NCEE 2009-4040

U. S. DEPARTMENT OF EDUCATION

Technical Methods Report: Estimation and Identification of the Complier Average Causal Effect Parameter in Education RCTs

NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

Technical Methods Report: Estimation and Identification of the Complier Average Causal Effect Parameter in Education RCTs

April 2009

Peter Z. Schochet
Hanley Chiang
Mathematica Policy Research, Inc.

Abstract

*In randomized control trials (RCTs) in the education field, the complier average causal effect (*CACE*) parameter is often of policy interest, because it pertains to intervention effects for students who receive a meaningful dose of treatment services. This report uses a causal inference and instrumental variables framework to examine the identification and estimation of the* CACE *parameter for two-level clustered RCTs. The report also provides simple asymptotic variance formulas for* CACE *impact estimators measured in nominal and standard deviation units. In the empirical work, data from ten large RCTs are used to compare significance findings using correct* CACE *variance estimators and commonly-used approximations that ignore the estimation error in service receipt rates and outcome standard deviations. Our key finding is that the variance corrections have very little effect on the standard errors of standardized* CACE *impact estimators. Across the examined outcomes, the correction terms typically raise the standard errors by less than 1 percent, and change p-values at the fourth or higher decimal place.*

NCEE 2009-4040
U.S. DEPARTMENT OF EDUCATION

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

This report was prepared for the National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences under Contract ED-04-CO-0112/0006.

Disclaimer
The Institute of Education Sciences (IES) at the U.S. Department of Education contracted with Mathematica Policy Research, Inc. to develop methods for examining the complier average causal effect in education evaluations. The views expressed in this report are those of the authors and they do not necessarily represent the opinions and positions of the Institute of Education Sciences or the U.S. Department of Education.

April 2009

Schochet, Peter Z. and Hanley Chiang (2009). *Estimation and Identification of the Complier Average Causal Effect Parameter in Education RCTs* (NCEE 2009-4040). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

This report is available on the IES website at http://ncee.ed.gov.

Alternate Formats
Upon request, this report is available in alternate formats such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department's Alternate Format Center at 202-260-9895 or 202-205-8113.

Disclosure of Potential Conflicts of Interest

The authors for this report, Dr. Peter Schochet and Dr. Hanley Chiang are employees of Mathematica Policy Research, Inc. with whom IES contracted to develop the methods that are presented in this report. Dr. Schochet, Dr. Hanley Chiang and other MPR staff do not have financial interests that could be affected by the content in this report.

Contents

List of Tables

Chapter 1: Introduction

Randomized control trials (RCTs) in the education field typically examine the intention-to-treat (*ITT*) parameter, which is estimated by comparing the mean outcomes of treatment group members (who are offered intervention services) to those of control group members (who are not). RCTs also sometimes examine two policy-relevant variants of the *ITT* parameter. The first variant is the complier average causal effect (*CACE*) parameter, defined as the average impact of intervention services on those who comply with their treatment assignments (Bloom 1984; Angrist et al. 1996). Estimators for this parameter are obtained by adjusting the *ITT* impact estimators for those in the treatment group who do not receive intervention services and for crossovers in the control group who erroneously receive intervention services. Second, it is becoming increasingly popular to standardize *ITT* and *CACE* impact estimates into effect size (standard deviation) units. This metric is useful for comparing findings across outcomes that are measured on different scales, for interpreting impacts that are difficult to understand in nominal units, and for comparing study findings to those from previous evaluations (Cohen 1988; Lipsey and Wilson 1995; Hedges 1981 and 2007).

This report addresses two main issues. First, it systematically examines the identification of the *CACE* parameter under *clustered* RCT designs that are typically used in the education field, where units (such as schools or classrooms) rather than students are randomly assigned to a treatment or control condition. Using a causal inference and instrumental variables (IV) framework, we extend the identification conditions in Angrist et al. (1996) to two-level clustered designs, where treatment compliance decisions can be made by both school staff and students. Our emphasis differs from Jo et al. (2008) who focus on parametric and path modeling of treatment noncompliance under clustered designs using multilevel mixture models and maximum likelihood methods.

The second purpose of the report is to theoretically and empirically examine variance estimation under clustered designs for two types of IV estimators: (1) *CACE* estimators in nominal units, and (2) *ITT* and *CACE* estimators in effect size units—hereafter referred to as standardized estimators. These estimators are ratio estimators, whose variances must account for estimation errors in their numerators and denominators. In practice, however, analysts often ignore the estimation error in the denominator terms, which are assumed to be known. Thus, in study reports, the same *t*-statistics and *p*-values are sometimes reported for all estimators.

A potential problem with this approach, however, is that it could lead to significance findings that are biased if the variance correction terms for the denominators matter. Accordingly, we present simple asymptotic variance estimation formulas for commonly-used ratio estimators by combining variance results in Hedges (2007) for standardized *ITT* estimators with those in Little et al. (2006) and Heckman et al. (1994) for *CACE* estimators. We then use data from ten large-scale RCTs to compare significance findings using the correct variance formulas with those that are typically used in practice, an empirical issue that has not been systematically addressed in the literature. The empirical results can be used to help guide future decisions as to whether the correct, but more complex variance formulas are warranted for RCTs in the education area to obtain rigorous significance findings for the full range of impact estimators.

The remainder of this report is in six chapters. Chapter 2 discusses the causal inference framework underlying the *ITT* estimator for two-level clustered designs, which forms the foundation for the *CACE* analysis. Chapter 3 discusses impact and variance estimation of the *ITT* parameter, and Chapter 4 discusses identification and estimation of the *CACE* parameter. Chapter 5 discusses estimation of the impact parameters in effect size units. Chapter 6 discusses empirical findings, and the final chapter presents a summary and conclusions.

Chapter 2: The Theoretical Framework Underlying the *ITT* Parameter

We consider two-level clustered designs where students are nested within units (such as schools, classrooms, or districts) that are randomly assigned to a single treatment or control group—the most common designs used in large-scale RCTs in the education field. The results that are presented for two-level designs, however, can be collapsed to obtain results for nonclustered designs where students are the unit of random assignment. This is because nonclustered designs are a special case of clustered designs where every cluster has one student and there is no within-cluster variance.

We consider a "superpopulation" version of the Neyman-Rubin causal inference model (see Rubin 1974; Imbens and Rubin 2007; Schochet 2008). It is assumed that the sample contains n units (groups), with np treatment units and $n(1-p)$ control units, where p is the sampling rate to the treatment group ($0<p<1$). Let W_{Ti} be the "potential" unit-level outcome for unit i when assigned to the treatment condition and W_{Ci} be the potential outcome for unit i in the control condition. These potential outcomes are assumed to be random draws from potential treatment and control outcome distributions in the study population with means μ_T and μ_C, respectively, and common variance σ_B^2. It is assumed that the potential outcomes for each unit are independent of the treatment status of other units.

Suppose next that m_i students are sampled from the student superpopulation within study unit i. Let Y_{Tij} and Y_{Cij} be student-level potential outcomes (conditional on unit-level potential outcomes) that are random draws from potential outcome distributions with means W_{Ti} and W_{Ci}, respectively, and common variance $\sigma_W^2 > 0$.

Under this causal inference model, the difference between the two potential outcomes, $(W_{Ti} - W_{Ci})$, is the unit-level treatment effect for unit i, and the *ITT* (or average treatment effect) parameter is $E(W_{Ti} - W_{Ci}) = \mu_T - \mu_C$. The unit-level treatment effects, and hence, the *ITT* parameter, cannot be calculated directly because for each unit and student, the potential outcome is observed in either the treatment or control condition, but not in both. Formally, if T_i is a treatment status indicator variable that equals 1 for treatments and 0 for controls, then the *observed* outcome for a unit, w_i, can be expressed as follows:

$$(1) \quad w_i = T_i W_{Ti} + (1 - T_i) W_{Ci}.$$

Similarly, the observed outcome for a student y_{ij} is:

$$(2) \quad y_{ij} = T_i Y_{Tij} + (1 - T_i) Y_{Cij}.$$

The simple equations in (1) and (2) form the basis for the estimation models that are considered in this report.

The terms in (1) can be rearranged to create the following regression model:

$$(3) \quad y_{ij} = \alpha_0 + \alpha_{ITT} T_i + (u_i + e_{ij}), \text{ where}$$

1. $\alpha_0 = \mu_C$ and $\alpha_{ITT} = \mu_T - \mu_C$ (the *ITT* parameter) are coefficients to be estimated;

2. $u_i = T_i(W_{Ti} - \mu_T) + (1 - T_i)(W_{Ci} - \mu_C)$ is a unit-level error term with mean zero and between-unit variance σ_B^2 that is uncorrelated with T_i; and

3. $e_{ij} = T_i(Y_{Tij} - W_{Ti}) + (1 - T_i)(Y_{Cij} - W_{Ci})$ is a student-level error term with mean zero and within-unit variance σ_W^2 that is uncorrelated with u_i and T_i.

Importantly, (3) can also be derived using the following two-level hierarchical linear model (HLM) (Bryk and Raudenbush 1992):

$$\begin{aligned} &Level\,1: \quad y_{ij} = w_i + e_{ij} \\ &Level\,2: \quad w_i = \alpha_0 + \alpha_{ITT} T_i + u_i, \end{aligned}$$

where Level 1 corresponds to students and Level 2 to units. Inserting the Level 2 equation into the Level 1 equation yields (3). Thus, the HLM approach is consistent with the causal inference theory presented above.

Finally, baseline covariates can be included in (3) as "irrelevant" variables to improve the precision of the impact estimates, which yields the following estimation model:

$$(4) \quad y_{ij} = \alpha_0 + \alpha_{ITT} T_i + (X_{ij} - \bar{X}_i)'\gamma + Z_i'\delta + (u_i^* + e_{ij}^*)$$

where X_{ij} is a vector of student-level baseline covariates that is centered around the unit-level covariate mean $\bar{X}_i$; γ is a parameter vector that is associated with X_{ij}; Z_i is a vector of unit-level baseline covariates (that could include $\bar{X}_i$ and stratum indicators) with associated parameter vector δ; and u_i^* and e_{ij}^* are error terms that are now conditional on the covariates. We center the X_{ij} covariates around the unit-level means so that we can separately identify the effects of covariates on the within- and between-unit variance components.

Chapter 3: *ITT* Impact and Variance Estimators

In this chapter, we use the models in (3) and (4) to discuss *ITT* estimators in nominal units, because they form the foundation for the *CACE* and standardized estimators. We focus on commonly used differences-in-means and analysis of covariance (ANCOVA) estimators, which are used for the empirical analysis.

We make the simplifying assumption that $m_i = m$ for all units (that is, equal cluster sizes). Cluster sizes are often similar for RCTs in the education area (and for the RCTs examined in our empirical work), and variance formulas are much more complex with unequal cluster sizes. Furthermore, the formulas presented in this chapter apply approximately for unequal unit sizes that do not vary substantially across units if *m* is replaced in the formulas by the average unit size $\overline{m}$ (Kish 1965) or, preferably, by $[n / \sum (1/m_i)]$ (Hedges 2007).

The Simple Differences-In-Means Estimator

The simple differences-in-means *ITT* estimator $\hat{\alpha}_{ITT1}$ can be obtained by applying standard regression methods to (3). The resulting estimator is as follows:

$$(5) \quad \hat{\alpha}_{ITT1} = \overline{\overline{y}}_T - \overline{\overline{y}}_C,$$

where $\overline{\overline{y}}_T = \frac{1}{np}\sum_{i:T_i=1}^{np} \overline{y}_i$; $\overline{\overline{y}}_C = \frac{1}{n(1-p)}\sum_{i:T_i=0}^{n(1-p)} \overline{y}_i$; and $\overline{y}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij}$. This estimator is the average difference between cluster means across the treatment and control groups.

Schochet (2008) shows that $\hat{\alpha}_{ITT1}$ is asymptotically normally distributed with mean α_{ITT} and the following asymptotic variance:

$$(6) \quad AsyVar(\hat{\alpha}_{ITT1}) = \frac{1}{p(1-p)}\left[\frac{\sigma_B^2}{n} + \frac{\sigma_W^2}{nm}\right].$$

The within-unit (second) variance term in (6) is the conventional variance expression for an impact estimator in a nonclustered design where random assignment is conducted within units. Design effects in a clustered design arise because of the first between-unit variance term, which represents the extent to which mean outcomes vary across units (Murray 1998; Donner and Klar 2000).

An asymptotically unbiased estimator for the within-unit variance σ_W^2 is as follows (Cochran 1963; Hedges 2007):

$$(7) \quad \hat{\sigma}_W^2 = S_W^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - \overline{y}_i)^2}{n(m-1)}.$$

Similarly, an asymptotically unbiased estimator for the between-unit variance σ_B^2 is:

(8) $\hat{\sigma}_B^2 = S_B^2 - \frac{S_W^2}{m}$, where

(9) $S_B^2 = \frac{\sum_{i:T_i=1}^{np} (\bar{y}_i - \bar{\bar{y}}_T)^2 + \sum_{i:T_i=0}^{n(1-p)} (\bar{y}_i - \bar{\bar{y}}_C)^2}{n-2}$.

Note that equation (9) can also be expressed in terms of regression residual sums of squares:

(10) $S_B^2 = \frac{\sum_{i=1}^{n} (\bar{y}_i - \hat{\bar{y}}_i)^2}{n-2}$,

where $\hat{\bar{y}}_i$ is the predicted value for unit *i* from the between-unit regression of $\bar{y}_i$ on T_i and an intercept.

Inserting (7) and (8) into (6) yields the following variance estimator for $\hat{\alpha}_{ITT1}$:

(11) $\hat{AsyVar}(\hat{\alpha}_{ITT1}) = \frac{S_B^2}{np(1-p)}$.

This estimator also applies to *nonclustered* designs where units are defined as students.

The Analysis of Covariance (ANCOVA) Estimator

The ANCOVA estimator $\hat{\alpha}_{ITT2}$ can be obtained by applying regression methods to (4) where baseline covariates (such as pretests) are included in the analytic models, primarily to improve the precision of the impact estimates. Schochet (2008) shows that $\hat{\alpha}_{ITT2}$ is asymptotically normally distributed with mean α_{ITT} and the following asymptotic variance:

(12) $AsyVar(\hat{\alpha}_{ITT2}) = \frac{1}{p(1-p)}\left[\frac{\sigma_{B1}^2}{n} + \frac{\sigma_{W1}^2}{nm}\right]$.

In this expression, σ_{B1}^2 and σ_{W1}^2 are between- and within-unit variances, respectively, that are conditional on the covariates, and reduce σ_B^2 and σ_W^2 depending on the size of the outcome-covariate correlations in the joint superpopulation distributions (these are R^2 adjustments).

Using methods that are parallel to the simple differences-in-means estimator presented above, a consistent variance estimator for $\hat{\alpha}_{ITT2}$ in (12) is as follows:

(13) $\hat{AsyVar}(\hat{\alpha}_{ITT2}) = \frac{S_{B1}^2}{np(1-p)}$,

where S_{B1}^2 is obtained using (10) with the following changes: (1) $\hat{\bar{y}}_i$ is now the predicted value for unit *i* from the between-unit regression of $\bar{y}_i$ on $Q_i = [1\ T_i\ Z_i]$; and (2) $(n-2)$ is replaced by $(n-k)$ where

k is the rank of the matrix Q whose rows contain the Q_i s. In practice, T_i and Z_i may be weakly correlated due to random sampling and missing data. Thus, (13) can be refined as follows:

(14) $Asy\hat{V}ar(\hat{\alpha}_{ITT2}) = (Q'Q)^{-1}_{2,2} S^2_{B1}$.

Finally, in our empirical work, we also used STATA to estimate more efficient generalized least squares models that allowed for unequal cluster sample sizes. Specifically, we used generalized estimating equation (GEE) methods with the sandwich variance estimator (Liang and Zeger 1986), and full and restricted maximum likelihood approaches to general linear mixed models (Littell et al. 1996; Bryk and Raudenbush 1992). The empirical results using these methods are very similar to those that are presented in this report, and thus, are not reported.

Chapter 4: The *CACE* Parameter

The *ITT* estimator provides information on treatment effects for those in the study population who were *offered* intervention services. The treatment group sample used to estimate this parameter, however, might include not only students who received services but also those who did not. Similarly, the control group sample may include crossovers who received embargoed intervention services for advertent or inadvertent reasons. In these cases, the *ITT* estimates may understate intervention effects for those who were eligible for and actually received services (assuming that the intervention improves outcomes). Thus, it is often of policy interest to estimate the *CACE* parameter that pertains to those who complied with their treatment assignments.

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

Sources of Noncompliance

Under clustered RCT designs in the education area, the extent to which students receive intervention services could depend on compliance decisions made by *both* school staff (such as superintendents, principals, and teachers) and students. The interplay between these sources will depend on the particular intervention and study design (Jo et al. 2008). Furthermore, the extent of compliance will depend on the approach for defining service dosage, which is a topic that is beyond the scope of this report. For context, however, in what follows, we briefly discuss general sources of noncompliance at the school and student levels.

Noncompliance by School Staff

School staff in treatment units may not offer intervention services for several reasons. First, school principals or district superintendents may change their minds about implementing the intervention, due to changes in school priorities or for other reasons. Second, even if schools agree to participate, some teachers may not, perhaps due to initial problems implementing the intervention or because they prefer their status quo teaching methods or curricula. In addition, noncompliance could occur if school personnel are not adequately trained in intervention procedures. Similarly, crossovers could occur if staff in control schools decide to offer the intervention (or a very similar one), perhaps because of a strong belief that the intervention is effective (from discussions with evaluators) and a strong desire to implement it immediately rather than after the embargo period.

Noncompliance by Students

Students may also play a role in noncompliance for several reasons. First, a student may not receive meaningful intervention services due to a lack of school attendance. This could occur, for example, if the

student is suspended, is chronically absent, or, if relevant, decides not to attend a voluntary program (for example, an after-school program).

Second, student mobility in and out of the study schools could lead to a low dosage of service receipt. In some designs, follow-up data are collected only for students who are present in the study schools at baseline (to ensure that the treatment and control group student samples will have similar baseline characteristics). In these designs, noncompliers may include those who left the treatment schools soon after the start of the school year. A more common "placed-based" design, however, is when follow-up data are collected for all students in the target grades who are in the study schools at data collection, including those who entered the schools after baseline. In these designs, noncompliers could include students who entered the study schools soon before follow-up data collection.[1] Under either design, crossovers could occur due to student mobility if control students in the follow-up sample transfer to treatment schools or classrooms.

Identification of the *CACE* Parameter

This section discusses the identification of the *CACE* parameter under two scenarios. First, to fix ideas, we assume that compliance is determined solely by school staff, and that all students who are offered services receive them. Second, we consider the more general case where compliance is determined by both schools and students, in which case some students may not receive services even if their schools offer them. For both scenarios, treatment status (T_i) is determined at random assignment and is *fixed* thereafter; T_i values are *not* affected by compliance decisions. We assume also that if the RCT uses a "placed-based" design as discussed above, there are no treatment effects on student mobility. Finally, because the literature has conceptualized compliance decisions as dichotomous (Angrist et al. 1996), we model the offer and receipt of services as binary decisions.

In what follows, we introduce some new notation. Let $R_i = R_i(T_i)$ denote an indicator variable that equals 1 if unit *i* would *offer* intervention services if assigned to a given treatment condition ($T_i = 0$ or $T_i = 1$), and let $W_i(T_i, R_i)$ denote the unit's potential outcome for a given value of (T_i, R_i); there are four such potential outcomes. Similarly, let $D_{ij} = D_{ij}(T_i, R_i)$ denote an indicator variable that equals 1 if the student *receives* intervention services from *any* study school, given one of the four possible combinations of (T_i, R_i). Finally, let $Y_{ij}(T_i, R_i, D_{ij})$ denote the student's potential outcome, given one of the possible combinations of (T_i, R_i, D_{ij}).

The *CACE* Parameter When Compliance Decisions Are Made by Units Only

To identify the *CACE* parameter when treatment compliance decisions are made by units only, we classify units into four mutually exclusive compliance categories: compliers, never-takers, always-takers, and defiers (Angrist et al. 1996). *Compliers (CL)* are those who would offer intervention services only if

[1] The *ITT* estimates under this design pertain to the *combined* effects of the intervention on student mobility and student outcomes, because of potential intervention effects on the fraction and types of students who enter and leave the study schools.

they were assigned to the treatment group [$R_i(1) = 1$ and $R_i(0) = 0$]. *Never-takers (N)* are those who would never offer treatment services [$R_i(1) = 0$ and $R_i(0) = 0$], and *always-takers (A)* are those would always offer treatment services [$R_i(1) = 1$ and $R_i(0) = 1$]. Finally, *defiers (D)* are those who would offer treatment services only in the control condition [$R_i(1) = 0$ and $R_i(0) = 1$]. Outcome data are assumed to be available for all sample members. Note that this scenario applies also to nonclustered designs where units are students.

The *ITT* parameter for the pooled sample α_{ITT} can be expressed as a weighted average of the *ITT* parameters for each of the four unobserved compliance groups:

$$(15) \quad \alpha_{ITT} = p_{CL}\alpha_{ITT\_CL} + p_N\alpha_{ITT\_N} + p_A\alpha_{ITT\_A} + p_D\alpha_{ITT\_D},$$

where p_g is the fraction of the study population in compliance group g ($\sum p_g = 1$), and α_{ITT_g} is the associated *ITT* impact parameter (as defined earlier).

Following Angrist et al. (1996), the α_{ITT_CL} parameter in (15) can then be identified under three key assumptions (U1-U3):

U1. ***The Unit-Level Stable Unit Treatment Value Assumption (SUTVA):*** Unit-level potential compliance decisions $[R_i(T_i)]$ and outcomes $[W_i(T_i, R_i)]$ are unrelated to the treatment status of other units. This allows us to express $R_i(T_i)$ and $W_i(T_i, R_i)$ in terms of T_i rather than the vector of treatment statuses of all units. This condition is likely to hold in clustered education RCTs where random assignment is conducted at the school level (the most common design), unless there is substantial interaction between students and staff across study schools.

U2. ***Unit-Level Monotonicity*****:** $R_i(1) \geq R_i(0)$. This means that units are at least as likely to offer intervention services in the treatment than control condition, and implies that there are no defiers (that is, $p_D = 0$). Under this assumption, $p_{CL} = P(R_i(1) = 1) - P(R_i(0) = 1)$, which is the difference between service offer rates in the treatment and control conditions.

U3. ***The Unit-Level Exclusion Restriction:*** $W_i(1, r) = W_i(0, r)$ for $r = 0, 1$. This means that the outcome for a unit that offers services would be the same in the treatment or control condition, and similarly for a unit that does not offer services. Stated differently, this restriction implies that any effect of T_i on outcomes must occur only through an effect of T_i on service offer rates. This restriction implies that impacts on always-takers and never-takers are zero, that is, $\alpha_{ITT\_N} = \alpha_{ITT\_A} = 0$.

Under these assumptions, the final three terms on the right-hand-side of (15) cancel. Thus, the following *CACE* impact parameter can be identified:

$$(16) \quad \alpha_{CACE0} = \alpha_{ITT\_CL} = E[W_i(1,1) - W_i(0,0)] = [\alpha_{ITT} / p_{CL}].$$

This parameter represents the average causal effect of the treatment for compliers.

Importantly, follow-up data on *all* sample members are required to estimate the *CACE* parameter even though this parameter pertains to the complier subgroup only. Thus, noncompliance is different than data nonresponse.

The *CACE* Parameter When Compliance Decisions Are Made by Units and Students

In this section, we generalize the *CACE* parameter from above to the case where compliance decisions are made by both school staff and students. For this analysis, we require assumptions on both students and schools to identify the *CACE* parameter.

Table 4.1 displays and labels the 16 possible *student-level* complier groups that depend on treatment status (T_i), whether the school offers services (R_i), and whether the student receives services (D_{ij}). In this scenario, there are four groups each of compliers, never-takers, defiers, and always-takers. For example, Never-Taker Group 2 includes students who would never receive services even though their schools would always offer them. Note that students with $R_i = 0$ and $D_{ij} = 1$ are assumed to receive services from a different study school than their baseline school. The frequency of each of the 16 combinations will depend on the particular application, and some may be rare. However, all combinations are included for completeness.

To derive the *CACE* parameter under this scenario, we define $\alpha^i_{ITT} = E(Y_{Tij} - Y_{Cij} \mid W_{Ti}, W_{Ci}, p_{1i}, ..., p_{16i})$ as the *within-unit ITT* for the student population in unit *i*, where p_{gi} is the fraction of students in compliance group *g* ($\sum_{g=1}^{16} p_g = 1$). Note that $E(\alpha^i_{ITT}) = \alpha_{ITT}$, where the expectation is taken over the joint unit-level potential outcome and compliance distributions. Note next that α^i_{ITT} can be expressed as a weighted average of the within-unit *ITT* parameters for each of the 16 student-level compliance groups shown in Table 4.1:

$$(17) \quad \alpha^i_{ITT} = \sum_{g=1}^{16} p_{gi} \alpha^i_{ITT\_g},$$

where $\alpha^i_{ITT\_g}$ is the impact parameter for compliance group *g*.

Table 4.1: Possible Student-Level Compliance Groups

Compliance Group: Number and Label	Compliance Status in the Treatment Condition		Compliance Status in the Control Condition	
	Unit: $R_i(1)$	Student: $D_{ij}[1, R_i(1)]$	Unit: $R_i(0)$	Student: $D_{ij}[0, R_i(0)]$
1. Complier 1	1	1	0	0
2. Always-Taker 1	1	1	0	1
3. Complier 2	1	1	1	0
4. Always-Taker 2	1	1	1	1
5. Never-Taker 1	1	0	0	0
6. Defier 1	1	0	0	1
7. Never-Taker 2	1	0	1	0
8. Defier 2	1	0	1	1
9. Complier 3	0	1	0	0
10. Always-Taker 3	0	1	0	1
11. Complier 4	0	1	1	0
12. Always-Taker 4	0	1	1	1
13. Never-Taker 3	0	0	0	0
14. Defier 3	0	0	0	1
15. Never-Taker 4	0	0	1	0
16. Defier 4	0	0	1	1

Note: $R_i(1)=1$ and $D_{ij}(1,1)=1$ if the student's unit (for example, school) would offer intervention services in the treatment condition and the student would then agree to receive services, and similarly for other combinations.

Using (17) and Table 4.1, the *CACE* parameter for Complier Group 1 can be identified under the following assumptions (that are analogs to the unit-level assumptions from above):

S1. ***SUTVA:*** Potential student-level service receipt decisions $[D_{ij}(T_i, R_i)]$ and outcomes $[Y_{ij}(T_i, R_i, D_{ij})]$ are unrelated to the treatment status of other students and schools. In addition, we impose the unit-level SUTVA condition U3 from above that $R_i(T_i)$ is unrelated to the treatment status of other units.

S2. ***Monotonicity on Compliance*****:** $R_i(1) \geq R_i(0)$ or $D_{ij}(1, R_i(1)) \geq D_{ij}(0, R_i(0))$. This assumption will be satisfied if a unit is at least as likely to offer services in the treatment than control condition, or if students in that unit are at least as likely to receive services in the treatment than control condition. Using Table 4.1, this condition implies that $p_{16} = 0$.

S3. ***Student-Level Monotonicity on the Take-Up of Services:*** $D_{ij}(s,1) \geq D_{ij}(t,0)$ for $s,t \in \{0,1\}$. This assumption means that students are at least as likely to take up services if they are offered them than if they are not, which implies that $p_6 = p_{11} = 0$.

S4. ***The Student-Level Exclusion Restriction on Compliance:*** $D_{ij}(1,r) = D_{ij}(0,r)$ for $r = 0,1$. This means that for a given service offer status, the student's compliance decision would be the same in the treatment or control condition. Stated differently, this restriction implies that any effect of T_i on student compliance decisions must be a result of a treatment effect on service offer rates. Using Table 4.1, this restriction implies that $p_3 = p_8 = p_9 = p_{14} = 0$.

S5. ***The Student-Level Exclusion Restriction on Outcomes:*** $Y_{ij}(1, R_i(1), d) = Y_{ij}(0, R_i(0), d)$ for $d = 0,1$. This means that student outcomes are determined solely by whether or not the student receive services, and it does not matter where these services are received (or not received) or how many other students are receiving them. This restriction implies zero impacts for Groups 2, 4, 5, 7, 10, 12, 13, and 15.

These assumptions imply that the only term on the right-hand-side of (17) that does not cancel is the first term that pertains to Complier Group 1 (see Table 4.1). Thus, after taking expectations in (17), the following *CACE* parameter can be identified:

$$(18) \quad \alpha_{CACE} = \alpha_{ITT} / E(p_{1i}) = E(p_{1i}\alpha^i_{ITT\_1}) / E(p_{1i}).$$

This *CACE* parameter is a weighted average of within-unit impacts for Complier Group 1, with weights p_{1i} (and reduces to $E(\alpha^i_{ITT\_1})$ if $\alpha^i_{ITT\_1}$ and p_{1i} are independent). Denoting $p_g = E(p_{gi})$, assumptions S2 to S4 imply that $p_1 = (p_T - p_C)$, where $p_T = p_1 + p_2 + p_4 + p_{10} + p_{12}$ and $p_C = p_2 + p_4 + p_{10} + p_{12}$ are the fractions of students receiving services in the treatment and control conditions, respectively. Thus, the only difference between α_{CACE0} in (16) and α_{CACE} in (18) is that p_{CL} refers to service offer rates for *units* whereas p_1 refers to service receipt rates for *students*. Clearly, (18) is more general and reduces to

(16) if compliance decisions are made by school staff only. Thus, in what follows, we focus on estimation issues for the α_{CACE} parameter.[2]

Impact and Variance Estimation of the *CACE* Parameter

In this section, we discuss estimation of the *CACE* parameter. We use an IV approach, because simple closed-form variance formulas exist, the variance correction terms can easily be understood because they enter the formulas linearly, and the formulas can be readily generalized to the standardized *CACE* parameter. An alternative approach, without these properties, is to use (more efficient) maximum likelihood estimation methods and the EM algorithm (Jo et al. 2008).

Impact Estimation

A consistent estimator for α_{CACE} in (18) can be obtained by dividing consistent estimators for α_{ITT} and p_1:

$$(19) \quad \hat{\alpha}_{CACE} = \hat{\alpha}_{ITT} / \hat{p}_1 .$$

Estimators for $p_1 = p_T - p_C$ can be obtained by noting that this parameter represents an impact on the rate of service receipt. Thus, estimation methods similar to those discussed above for α_{ITT} can be used to estimate p_1. For example, analogous to (5), the simple differences-in-means estimator is $\hat{p}_1 = (\bar{\bar{d}}_T - \bar{\bar{d}}_C)$, where d_{ij} is an observed service receipt status indicator variable that equals 1 if student *i* in school *j* received intervention services, and zero otherwise.

Variance Estimation

The *CACE* estimator in (19) is a ratio estimator (Little et al. 2008 and Heckman et al. 1994). Both the numerator and denominator are measured with error, and thus, both sources of error should be taken into account in the variance calculations. A variance estimator for $\hat{\alpha}_{CACE}$ can be obtained using an asymptotic Taylor series expansion of $\hat{\alpha}_{CACE}$ around the true value α_{CACE}:

$$(20) \quad \left(\hat{\alpha}_{CACE} - \alpha_{CACE}\right) \approx \frac{(\hat{\alpha}_{ITT} - \alpha_{ITT})}{p_1} - \frac{\alpha_{ITT}(\hat{p}_1 - p_1)}{p_1^2} .$$

[2] A special case of our general framework is when always-takers (groups 2, 4, 10, and 12) are not present, possibly as a result of strict implementation rules ensuring that students from control schools cannot receive intervention services. In this case, recipients of intervention services belong only to complier group 1 in the treatment group, and the *CACE* parameter is equivalent to the average treatment effect on those who receive services (the "treatment-on-the-treated" parameter).

Taking squared expectations on both sides of (20) and inserting estimators for unknown parameters yields the following variance estimator for $\hat{\alpha}_{CACE}$:

$$(21) \quad \hat{AsyVar}(\hat{\alpha}_{CACE}) = \frac{\hat{AsyVar}(\hat{\alpha}_{ITT})}{\hat{p}_1^2} + \frac{\hat{\alpha}_{CACE}^2 \hat{AsyVar}(\hat{p}_1)}{\hat{p}_1^2} - \frac{2\hat{\alpha}_{CACE} \hat{AsyCov}(\hat{\alpha}_{ITT}, \hat{p}_1)}{\hat{p}_1^2} \ .$$

The first term in (21) is the variance of the *CACE* estimator assuming that estimated service receipt rates are measured without error. The second and third terms are therefore correction terms. The second term accounts for the estimation error in $\hat{p}_1$, and the third term accounts for the covariance between $\hat{\alpha}_{ITT}$ and $\hat{p}_1$.[3] Importantly, these correction terms depend on the size of α_{ITT}, and thus, become more important with larger impacts. Finally, because $\hat{\alpha}_{ITT}$ and $\hat{p}_1$ are asymptotically normal, the delta method (Greene 2000, p.118) implies that $\hat{\alpha}_{CACE}$ is also asymptotically normal.

An asymptotic variance estimator for $\hat{p}_1$ that adjusts for clustering can be obtained from (10) and (11) using simple differences-in-means methods or from (13) and (14) using linear probability models, where y_{ij} is replaced by d_{ij}. For our empirical work, we used a slightly different variance estimator that allows for different processes underlying service receipt decisions for treatments and controls:

$$(22) \quad \hat{AsyVar}(\hat{p}_1) = \frac{S_{BT}^2}{np} + \frac{S_{BC}^2}{n(1-p)},$$

where $S_{BT}^2 = \sum_{i:T_i=1}^{np} (\bar{d}_i - \hat{\bar{d}}_i)^2 / [np-k]$, $S_{BC}^2 = \sum_{i:T_i=0}^{n(1-p)} (\bar{d}_i - \hat{\bar{d}}_i)^2 / [n(1-p)-k]$, and k is the number of unit-level covariates (including the intercept) that are included in the model.

Similarly, an unbiased estimator for $\hat{AsyCov}(\hat{\alpha}_{ITT}, \hat{p}_1)$ is as follows:

$$(23) \quad \hat{AsyCov}(\hat{\alpha}_{ITT}, \hat{p}_1) = \frac{S_{\alpha p}^T}{np} + \frac{S_{\alpha p}^C}{n(1-p)} \ ,$$

where $S_{\alpha p}^T = \sum_{i:T_i=1}^{np} (\bar{y}_i - \hat{\bar{y}}_i)(\bar{d}_i - \hat{\bar{d}}_i) / [np-k]$ and $S_{\alpha p}^C = \sum_{i:T_i=0}^{n(1-p)} (\bar{y}_i - \hat{\bar{y}}_i)(\bar{d}_i - \hat{\bar{d}}_i) / [n(1-p)-k]$.

Finally, the *CACE* impact and variance estimators discussed above are IV estimators (Angrist et al. 1996). To see this, consider the following variant of the model in (3):

$$(24) \quad y_{ij} = \alpha_0 + \alpha_{CACE} d_{ij} + (u_i^{IV} + e_{ij}^{IV}) \ ,$$

[3] Little et al. (2008) and Heckman et al. (1994) ignore the covariance term.

where u_i^{IV} and e_{ij}^{IV} are random error terms. If T_i is used as an instrument for d_{ij} in (24), then the estimated IV regression coefficient $\hat{\alpha}_{CACE}$ is the simple differences-in-means *CACE* estimator in (19) with the variance estimator in (21). Treatment status T_i is likely to be a "strong" instrument if service receipt rates differ markedly for treatment and control students (see Murray 2006 and Stock et al. 2002 for a discussion of weak and strong instruments).[4]

[4] Due to the correction terms in (21), the correct *p*-values of the *ITT* and *CACE* estimates will generally differ, and the choice of the parameter on which inference is conducted should be determined by the population of interest. However, when the problem of weak instruments precludes valid inference on the *CACE* parameter, inference about the absence of an effect may have to be conducted solely on the *ITT* parameter.

Chapter 5: The Standardized *ITT* and *CACE* Estimators

It is becoming increasingly popular in educational research to standardize estimated impacts into standard deviation units (Hedges 1981 and 2007). This approach can be used to facilitate the comparison of impact findings across outcomes that are measured on different scales. It has also been used extensively in meta-analyses to contrast and collate impact findings across a broad range of disciplines (Cohen 1988; Lipsey and Wilson 1993). The use of effect sizes is especially important for helping to understand impact findings on outcomes that are difficult to interpret when measured in nominal units (for example, impacts on behavioral scales or test scores). In addition, this approach is useful for creating composite measures across multiple outcomes, and for scaling an outcome that is measured differently across students (such as state achievement test scores from different states). Finally, it has become standard practice in education evaluations to conduct power analyses using primary outcomes that are measured in effect size units, to ensure adequate study sample sizes for detecting impacts that are meaningful and attainable based on findings from previous studies.

Impact Estimation for the Standardized *ITT* Estimator

The *ITT* parameter in effect size units, α_{ITT_E}, can be expressed as follows:

$$(25) \quad \alpha_{ITT\_E} = \alpha_{ITT} / \sigma_y ,$$

where σ_y is the standard deviation of the outcome across all treatment and control students.[5]

An unbiased standard deviation estimator for $\sigma_y = \sqrt{\sigma_B^2 + \sigma_W^2}$ can be obtained as follows:

$$(26) \quad S_y = \sqrt{S_B^2 + \frac{(m-1)S_W^2}{m}} ,$$

where S_B^2 and S_W^2 are defined as in (9) and (7) above. Thus, a consistent estimator for α_{ITT_E} is:

$$(27) \quad \hat{\alpha}_{ITT\_E} = \hat{\alpha}_{ITT} / S_y .$$

Variance Estimation for the Standardized *ITT* Estimator

The effect size estimator in (27) is a ratio estimator where both the numerator and denominator are measured with error. We discern, however, two competing views on whether it is necessary, when reporting impact results, to adjust the variance of this estimator for the estimation error in S_y. One view, that opposes variance corrections, is that standardized impact estimators are descriptive statistics for

[5] In clustered designs, σ_y could also be defined as the within- or between-unit unit standard deviation, and could also be measured using the control group only or an outside sample (for example, a sample with pertinent data that is larger and more representative of the study "universe" than the sample for the current study) (Hedges 2007). All formulas below can be adapted using these alternative definitions for σ_y.

interpreting and benchmarking the impacts in nominal units. In this view, standardized outcomes are not measures per se, and thus, the nominal estimator is the relevant impact for assessing whether an education intervention had a significant impact on the outcome. The alternative view is that the standardized impact estimator is often the impact measure on which researchers and policymakers focus. Thus, standardized outcomes are effectively the outcome measures of interest, and standardized impacts should have proper standard errors attached to them.

Given these opposing views, we believe that it is appropriate that impact studies report correct standard errors for *ITT* impact estimates in *both* nominal and effect size units. Thus, in what follows, we discuss simple asymptotic variance formulas for the standardized estimators (see Hedges 2007 for similar results using finite populations and unequal cluster sizes).

A variance estimator for $\hat{\alpha}_{ITT\_E}$ can be obtained from the delta method using a Taylor series expansion of $\hat{\alpha}_{ITT\_E}$ around the true value α_{ITT_E}, which after inserting estimators for unknown parameters, yields the following expression:

$$(28) \quad \hat{AsyVar}(\hat{\alpha}_{ITT\_E}) = \frac{\hat{AsyVar}(\hat{\alpha}_{ITT})}{S_y^2} + \frac{\hat{\alpha}_{ITT\_E}^2 \hat{AsyVar}(S_y)}{S_y^2},$$

where the asymptotic covariance term between $\hat{\alpha}_{ITT}$ and S_y can be shown to be zero using results on the independence of linear functions and quadratic forms for normal distributions.

The first term in (28) is the variance expression for the effect size impact ignoring the estimation error in S_y (the usual approach found in the literature). The second term, therefore, is a correction term. This term increases as $\hat{\alpha}_{ITT\_E}$ increases, and is zero if and only if $\hat{\alpha}_{ITT\_E} = 0$.

Finally, (28) requires an estimator for $AsyVar(S_y)$, which can be obtained as follows (see Appendix A for a proof):

$$(29) \quad \hat{AsyVar}(S_y) = \frac{S_B^4}{2(n-2)S_y^2} + \frac{(m-1)S_W^4}{2nm^2 S_y^2}.$$

This expression also applies to *nonclustered* designs where units are defined as students. In this case, $S_W^2 = 0$ and $S_y^2 = S_B^2$ so that (29) reduces to $S_B^2/[2(n-2)]$.

Impact and Variance Estimation for the Standardized *CACE* Estimator

Using results from above, a *CACE* estimator in *effect size units* can be expressed as follows:

$$(30) \quad \hat{\alpha}_{CACE\_E} = \hat{\alpha}_{ITT}/(S_y \hat{p}_1),$$

where it is assumed that the standard deviation for compliers is the same as it is for the full sample. Using the delta method, a variance estimator for $\hat{\alpha}_{CACE\_E}$ is:

$$(31)\quad \hat{Asy Var}(\hat{\alpha}_{CACE\_E}) = \frac{\hat{Asy Var}(\hat{\alpha}_{ITT})}{S_y^2 \hat{p}_1^2} + \frac{\hat{\alpha}_{CACE\_E}^2 \hat{Asy Var}(S_y)}{S_y^2} + \frac{\hat{\alpha}_{CACE\_E}^2 \hat{Asy Var}(\hat{p}_1)}{\hat{p}_1^2} - \frac{2\hat{\alpha}_{CACE\_E} \hat{Asy Cov}(\hat{\alpha}_{ITT}, \hat{p}_1)}{S_y \hat{p}_1^2},$$

where we have ignored the covariance term between S_y and $\hat{p}_1$. The estimator $\hat{\alpha}_{CACE\_E}$ is asymptotically normal because each estimator component is asymptotically normal.

Chapter 6: Empirical Analysis

RCTs in the education field often report the same significance levels for each of the *ITT* and *CACE* estimators considered above. This chapter uses data from ten RCTs to assess this approach.

Data

Data for our analysis come from ten large published RCTs conducted by Mathematica Policy Research, Inc. (MPR). We selected these RCTs due to their significance for policy and their coverage of a wide range of interventions found in the education and social policy fields. Most of these evaluations were advised by national panels of evaluation and subject-area experts. Appendix Table B.1 lists the RCTs and summarizes the basic features of each one, including the 20 key outcome variables selected for our analysis, the covariates used in regression adjustment, and the unit of random assignment (that is, the level of clustering).

The RCTs include six evaluations of K-12 educational interventions. The remaining four RCTs include evaluations of interventions in welfare, labor, and early childhood education, which are included to help gauge the robustness of our findings beyond the K-12 setting. Overall, the ten studies span a wide range of outcomes, geographic areas, and target populations, and there is a mix of clustered and nonclustered designs. All ten studies were used for the standardized *ITT* analysis.

The *CACE* analysis was conducted using data from seven RCTs where noncompliers were identified using service receipt data. Appendix Table B.2 provides definitions of program "participation" used in our *CACE* analysis, and shows unadjusted service receipt rates in the treatment and control groups. For the 21st Century, New York City Voucher, Power4Kids, Early Head Start, and Job Corps evaluations, we defined program participation using the same rules as used by the studies. The Teacher Induction and Education Technologies evaluations did not conduct *CACE* analyses, so we developed illustrative rules for defining noncompliers using available service receipt data. The *CACE* analysis was not conducted for the remaining three RCTs (the Teach for America, San Diego Food Stamp Cash-Out, and Teenage Parent Demonstration evaluations) due to full compliance of study subjects.

For the *CACE* analysis, individuals were coded as service recipients if they received at least a *minimal* amount of services. It is appropriate to set the bar low for defining service receipt to ensure that impacts on never-takers are likely to be zero (see assumptions U3 and S5 above).

Methods

Data from each RCT were used to obtain (1) *uncorrected* variance estimators where the denominator terms of the impact estimators were assumed to be known, and (2) *corrected* variance estimators that accounted for all sources of estimation error.

Variance estimators for $\hat{\alpha}_{ITT\_E}$ were obtained using (28). To apply (28), we estimated between-unit ANCOVA models to obtain $\hat{\alpha}_{ITT}$ and then used (14) to obtain $AsyV\hat{a}r(\hat{\alpha}_{ITT})$. The ANCOVA models included covariates as similar as possible to those used in the published studies (see Table B.1).[6] The

[6] The impact estimates and uncorrected variance estimates that we report are slightly different than those reported in the published study reports due to the standardization of the estimation methods that we used across

estimation of σ_y and $AsyVar(S_y)$ involved a straightforward application of (26) and (29). Equations (31), (22), and (23) were used to obtain $\hat{AsyVar}(\hat{\alpha}_{CACE\_E})$. Similar impact and variance results were found using simple differences-in-means procedures and the other estimation methods discussed above (not shown).

The *CACE* analysis required the estimation of the fraction p_1 of individuals who were compliers. To do this, we defined a binary variable d_{ij} that was set to 1 if the individual received services and zero otherwise. We then estimated p_1 as the coefficient on T_i from a between-unit regression of $\overline{d}_i$ on T_i and the same covariates that were used to estimate the *ITT* parameters. Similar results were found using simple differences-in-means procedures and logit models.

For each outcome, we quantified the importance of the variance corrections in two ways. First, we calculated the difference between the corrected standard error (the square root of the sum of the two terms in (28) or four terms in (31)) and the uncorrected standard error (the square root of the first terms in (28) or (31)) as a percentage of the uncorrected standard error. Second, we used *t*-statistics to assess the effect of the variance corrections on the statistical significance of $\hat{\alpha}_{ITT\_E}$ and $\hat{\alpha}_{CACE\_E}$ by calculating the absolute difference between the corrected and uncorrected *p*-values.[7]

The importance of the variance corrections will depend on the size of the impact estimates. Thus, to assess the sensitivity of our main findings to larger impacts than were typically found in the considered RCTs, we conducted simulations assuming that impacts were 0.25 standard deviations, which is a value that education RCTs are often powered to detect (Schochet 2007). For these simulations, variances of nominal *ITT* impact estimators were assumed to be the same as those observed in the data. Finally, for each outcome, we conducted a related analysis by identifying the smallest positive impact values for which the variance corrections would raise the standard errors of the impact estimators by 5 percent from the uncorrected values. Such an increment to the standard errors would cause an impact estimate with an uncorrected *p*-value of 0.04 to become, as a result of the correction, barely insignificant at the 0.05 level.

Because our variance formulas are based on asymptotic normality of the impact estimators and assume equal cluster sizes, they are only approximations. Thus, to evaluate whether our formulas apply well to sample and cluster sizes that arise in practice, we compared *p*-values based on our variance formulas with those based on a nonparametric bootstrap. The two methods yield very similar *p*-values (Appendix Table B.3).

(continued)
studies, and small differences in covariate sets, the treatment of strata, and weighting schemes. However, the two sets of findings are very similar (see Appendix Table B.1).

[7] We focus on absolute, rather than percentage changes in the *p*-values because a large percentage change in a *p*-value may have only a trivial effect on statistical significance if the original, uncorrected *p*-value is already small.

Results

The nominal *ITT* estimates are statistically significant at the five percent level for half of the 20 outcomes included in the analysis (Table 6.1; Column 4). These significance levels also apply to the standardized *ITT* and *CACE* estimates (discussed later) using the *uncorrected* variance estimates. Estimates of α_{ITT_E} are less than 0.15 standard deviations for 16 of the 20 outcomes (Table 6.1; Column 5). The Power4Kids study had the largest intervention effects (0.38 and 0.22 standard deviations for the two reading outcomes, respectively).

Compliance rates varied somewhat across the 7 studies included in the *CACE* analysis (Table 6.2; Column 2). The compliance rate was at least 88 percent in four RCTs, and ranged from 72 to 77 percent in the three other RCTs. By construction, $\hat{\alpha}_{CACE\_E}$ becomes closer to $\hat{\alpha}_{ITT\_E}$ as estimated compliance rates increase (Table 6.2).

Is Accounting for the Estimation Error in the Denominator of $\hat{\alpha}_{ITT\_E}$ Important?

The answer to this question is "no." We find strong evidence that accounting for the estimation error in S_y has a negligible effect on the standard error of $\hat{\alpha}_{ITT\_E}$ (Table 6.3). In our data, the correction term raises the standard error of $\hat{\alpha}_{ITT\_E}$ by less than one-quarter of 1 percent for 18 out of 20 outcome variables, and the correction never increases the standard error by more than 2 percent (Table 6.3; Column 5). Similarly, the correction has a trivial effect on the statistical significance of $\hat{\alpha}_{ITT\_E}$. As shown in the final column of Table 6.3, the correction changes the *p*-value of $\hat{\alpha}_{ITT\_E}$ only at the fourth or higher decimal place.

The correction for the estimation error in S_y would remain ignorable even if the *ITT* estimates were 0.25 standard deviations, which is larger than most of our observed $\hat{\alpha}_{ITT\_E}$ values (Table 6.4; Column 2). As expected, when $\hat{\alpha}_{ITT\_E}$ is set to 0.25, the correction becomes more important than before, but the correction still has a very small effect on the standard errors (less than a 2 percent increase in all but one instance). Similarly, the *p*-value of $\hat{\alpha}_{ITT\_E}$ is hardly affected; the absolute increase in the *p*-value due to the correction never exceeds 0.001 (Table 6.4; Column 3). In fact, if $\hat{\alpha}_{ITT\_E}$ were 0.25, the *t*-statistic of the estimate would typically be so far out in the right tail of the distribution that the slight decrease in the *t*-statistic from the correction would leave the *p*-value virtually unchanged.

Similarly, we find that on average across the considered RCTs, $\hat{\alpha}_{ITT\_E}$ would need to be about 0.8 standard deviations for the corrections to increase the standard error of $\hat{\alpha}_{ITT\_E}$ by 5 percent (last column in Table 6.4). This is a large effect size in social policy evaluations, and is more than double the largest *ITT* impact found in our studies.

Table 6.1: Dependent Variable Information and *ITT* Impact Estimates in Nominal and Effect Size Units, by Study

Study and Dependent Variable	Dependent Variable Information		*ITT* Impact Estimate[a]	
	Measurement Units	Standard Deviation	Nominal Units	Standard Deviation Units
21st Century				
Reading score	Percentiles	25.57	-0.70	-0.027
Math course grade	Percentage points	9.98	-0.62	-0.062
Teach for America				
Reading score	NCE[b]	21.99	0.34	0.016
Math score	NCE[b]	18.56	2.35*	0.127*
Education Technologies				
Grade 1 reading score	NCE[b]	20.62	0.28	0.014
Grade 4 reading score	NCE[b]	18.81	0.31	0.017
NYC Vouchers				
Reading score	Percentiles	23.08	0.74	0.032
Math score	Percentiles	23.50	0.82	0.035
Power4Kids				
Word attack score	Standard points	10.77	4.06*	0.377*
GRADE score	Standard points	14.33	3.16*	0.221*
Teacher Induction				
Whether stay in district	Binary outcome	0.38	-0.01	-0.030
Lesson implement score	Scale points	0.92	-0.01	-0.015
Early Head Start				
Bayley MDI score	Scale points	12.63	1.42*	0.113*
HOME score	Scale points	4.79	0.39*	0.081*
Job Corps				
Earnings	Dollars per week	195.02	12.04*	0.062*
Arrests	Number	1.44	-0.14*	-0.095*
Cash-Out				
Value of purchased food	Dollars per week	42.36	-6.38*	-0.151*
Energy as percent of RDA	Percentage points	62.21	-7.41*	-0.119*
Teenage Parent				
Percent of months active	Percentage points	32.94	5.96*	0.181*
Earnings	Dollars per month	268.53	18.79	0.070

Source: Data from studies listed in Appendix Table B.1.

Note: Impact estimates are regression-adjusted using the covariates indicated in Appendix Table B.1.

[a] *ITT* impacts are estimated by the authors. See Appendix Table B.1 for nominal *ITT* impact estimates from the published study reports.

[b] Denotes normal curve equivalents.

* The *ITT* impact estimate is significantly different from zero at the 0.05 level, two-tailed test using the uncorrected standard error.

Table 6.2: Standardized *ITT* and *CACE* Impact Estimates, by Study

Study and Dependent Variable	Estimated Fraction of Individuals in the Study Who Are Compliers	Standardized Impact Estimates	
		ITT	*CACE*
21st Century			
Reading score	0.753	-0.027	-0.036
Math course grade	0.765	-0.062	-0.081
Education Technologies			
Grade 1 reading score	0.881	0.014	0.016
Grade 4 reading score	0.890	0.017	0.019
NYC Vouchers			
Reading score	0.745	0.032	0.043
Math score	0.745	0.035	0.047
Power4Kids			
Word attack score	0.996	0.377*	0.379*
GRADE score	0.996	0.221*	0.222*
Teacher Induction			
Whether stay in district	0.947	-0.030	-0.031
Lesson implement score	0.960	-0.015	-0.016
Early Head Start			
Bayley MDI score	0.919	0.113*	0.123*
HOME score	0.919	0.081*	0.088*
Job Corps			
Earnings	0.717	0.062*	0.086*
Arrests	0.718	-0.095*	-0.133*

Source: Data from studies listed in Appendix Table B.1.

Note: Impact estimates and estimated fractions of individuals who are compliers are regression-adjusted using the covariates indicated in Appendix Table B.1.

* The impact estimate is significantly different from zero at the 0.05 level, two-tailed test using the uncorrected standard error.

Table 6.3: Uncorrected and Corrected Standard Errors of $\hat{\alpha}_{ITT\_E}$, by Study

Study and Dependent Variable	$\hat{\alpha}_{ITT\_E}$ (x 10^2)	Standard Error of $\hat{\alpha}_{ITT\_E}$			Absolute Change in the *p*-value of $\hat{\alpha}_{ITT\_E}$ due to the Correction (x 10^4)
		Uncorrected Value (x 10^2)	Corrected Value (x 10^2)	Percentage Change due to the Correction	
21st Century					
Reading score	-2.7	4.142	4.142	0.01	0.280
Math course grade	-6.2	4.873	4.874	0.03	1.179
Teach for America					
Reading score	1.6	3.436	3.437	0.02	0.636
Math score	12.7	5.523*	5.534*	0.21	2.736
Education Technologies					
Grade 1 reading score	1.4	4.481	4.481	0.00	0.043
Grade 4 reading score	1.7	3.797	3.798	0.01	0.200
NYC Vouchers					
Reading score	3.2	5.138	5.138	0.01	0.363
Math score	3.5	5.328	5.329	0.01	0.404
Power4Kids					
Word attack score	37.7	8.020*	8.151*	1.63	0.011
GRADE score	22.1	9.571*	9.609*	0.40	5.134
Teacher Induction					
Whether stay in district	-3.0	7.100	7.100	0.01	0.154
Lesson implement score	-1.5	8.547	8.547	0.00	0.018
Early Head Start					
Bayley MDI score	11.3	4.417*	4.421*	0.10	0.772
HOME score	8.1	4.009*	4.011*	0.06	1.187
Job Corps					
Earnings	6.2	1.965*	1.965*	0.02	0.041
Arrests	-9.5	1.935*	1.936*	0.05	0.000
Cash-Out					
Value of purchased food	-15.1	6.066*	6.074*	0.13	1.146
Energy as percent of RDA	-11.9	6.066*	6.072*	0.09	2.024
Teenage Parent					
Percent of months active	18.1	4.750*	4.761*	0.22	0.048
Earnings	7.0	4.815	4.817	0.03	1.332

Source: Data from studies listed in Appendix Table B.1.

Note: Impact estimates and standard errors are regression-adjusted using the covariates indicated in Appendix Table B.1. The percentage change in the standard error due to the correction is equal to the difference between the corrected and uncorrected standard error, divided by the uncorrected standard error, and multiplied by 100.

* The standardized *ITT* estimate is significant at the 0.05 level, two-tailed test using the indicated standard error.

Table 6.4: Simulated Effects of Variance Corrections on the Standard Error of $\hat{\alpha}_{ITT\_E}$, for an Assumed *ITT* Impact Value of 0.25 and by Study

Study and Dependent Variable	Percentage Change in the Standard Error of $\hat{\alpha}_{ITT\_E}$ due to the Correction if $\hat{\alpha}_{ITT\_E}$ Were Equal to 0.25	Absolute Change in the *p*-value (x 10^4) of $\hat{\alpha}_{ITT\_E}$ due to the Correction if $\hat{\alpha}_{ITT\_E}$ Were Equal to 0.25	Value of $\hat{\alpha}_{ITT\_E}$ so that the Correction Will Increase the Standard Error by 5 Percent
21st Century			
Reading score	0.55	0.000	0.762
Math course grade	0.42	0.000	0.872
Teach for America			
Reading score	4.87	0.000	0.253
Math score	0.80	0.011	0.631
Education Technologies			
Grade 1 reading score	0.61	0.000	0.723
Grade 4 reading score	1.41	0.000	0.475
NYC Vouchers			
Reading score	0.54	0.002	0.770
Math score	0.50	0.003	0.802
Power4Kids			
Word attack score	0.72	1.430	0.666
GRADE score	0.51	3.536	0.793
Teacher Induction			
Whether stay in district	0.36	0.209	0.944
Lesson implement score	0.36	1.182	0.941
Early Head Start			
Bayley MDI score	0.48	0.000	0.814
HOME score	0.54	0.000	0.771
Job Corps			
Earnings	0.37	0.000	0.925
Arrests	0.37	0.000	0.924
Cash-Out			
Value of purchased food	0.35	0.024	0.961
Energy as percent of RDA	0.39	0.027	0.905
Teenage Parent			
Percent of months active	0.43	0.000	0.867
Earnings	0.42	0.000	0.872

Source: Data from studies listed in Appendix Table B.1.

Note: Standard errors are regression-adjusted using the covariates indicated in Appendix Table B.1. The percentage change in the standard error due to the correction is equal to the difference between the corrected and uncorrected standard error, divided by the uncorrected standard error, and multiplied by 100.

Is Accounting for the Estimation Error in the Denominator of $\hat{\alpha}_{CACE\_E}$ Important?

The answer to this question is also "no." We find that the variance corrections exert a bit more influence on the variance estimates for $\hat{\alpha}_{CACE\_E}$ than $\hat{\alpha}_{ITT\_E}$, but the influence is still generally very small; only in rare instances do these corrections change the variance estimates by more than 1 percent.

Our key finding is that the standard error of $\hat{\alpha}_{CACE\_E}$ does not rise noticeably when correction terms involving S_y and $\hat{p}_1$ are included in the variance calculations (Table 6.5). The corrections increase the uncorrected standard errors by less than 0.5 percent for all studies except for the Word Attack Score in the Power4Kids study where the increase is 1.6 percent (Table 6.5; Column 5). The effect of the corrections on *p*-values is negligible; the corrections never raise or lower the *p*-value by more than 0.001 (Table 6.5; last column).

We find also that none of the individual correction terms in equation (31) is consistently important (Table 6.6). For 12 out of 14 outcome variables, every correction term is less than 0.5 percent of the uncorrected variance value for $\hat{\alpha}_{CACE\_E}$. Interestingly, $Asy\hat{C}ov(\hat{\alpha}_{ITT}, \hat{p}_1)$ has no consistent sign. In some instances, the variance reduction due to a negative covariance term offsets the positive variance contributions of the other correction terms. This explains why in some cases the corrections *reduce* the standard errors shown in Table 6.5 (as indicated by negative values in the fifth column of Table 6.5).

Simulations suggest that the results remain unchanged if $\hat{\alpha}_{CACE\_E}$ is set to 0.25 (Table 6.7; Columns 2 and 3). For this scenario, for all but one outcome, the correction terms raise the standard error of $\hat{\alpha}_{CACE\_E}$ by less than 2 percent; the corresponding rise in the *p*-value never exceeds 0.001. Furthermore, on average across the considered RCTs, the standardized *CACE* impact would need to be 0.7 standard deviations for the corrections to raise the standard error of $\hat{\alpha}_{CACE\_E}$ by 5 percent (Table 6.7; Column 4).

Table 6.5: Uncorrected and Corrected Standard Errors of $\hat{\alpha}_{CACE\_E}$, by Study

Study and Dependent Variable	$\hat{\alpha}_{CACE\_E}$ (x 10^2)	Standard Error of $\hat{\alpha}_{CACE\_E}$: Uncorrected Value (x 10^2)	Standard Error of $\hat{\alpha}_{CACE\_E}$: Corrected Value (x 10^2)	Percentage Change due to the Correction	Absolute Change in the *p*-value of $\hat{\alpha}_{CACE\_E}$ due to the Correction (x 10^4)
21st Century					
Reading score	-3.6	5.498	5.499	0.01	0.247
Math course grade	-8.1	6.368	6.378	0.16	7.173
Education Technologies					
Grade 1 reading score	1.6	5.084	5.081	-0.07	-1.647
Grade 4 reading score	1.9	4.265	4.262	-0.07	-2.363
NYC Vouchers					
Reading score	4.3	6.901	6.902	0.02	0.992
Math score	4.7	7.157	7.164	0.10	4.369
Power4Kids					
Word attack score	37.9	8.054*	8.186*	1.64	0.012
GRADE score	22.2	9.612*	9.653*	0.42	5.412
Teacher Induction					
Whether stay in district	-3.1	7.500	7.503	0.04	1.117
Lesson implement score	-1.6	8.900	8.902	0.02	0.330
Early Head Start					
Bayley MDI score	12.3	4.806*	4.811*	0.10	0.808
HOME score	8.8	4.362*	4.359*	-0.06	-1.324
Job Corps					
Earnings	8.6	2.740*	2.741*	0.03	0.055
Arrests	-13.3	2.694*	2.695*	0.04	0.000

Source: Data from studies listed in Appendix Table B.1.

Note: Impact estimates and standard errors are regression-adjusted using the covariates indicated in Appendix Table B.1. The percentage change in the standard error due to the correction is equal to the difference between the corrected and uncorrected standard error, divided by the uncorrected standard error, and multiplied by 100.

* The standardized *CACE* impact estimate is significantly different from zero at the 0.05 level, two-tailed test using the indicated standard error.

Table 6.6: Components of Variance Corrections for $\hat{\alpha}_{CACE\_E}$, by Study

Study and Dependent Variable	Percentage Change in the Estimated Variance of $\hat{\alpha}_{CACE\_E}$ due to Correction Terms Involving: All Corrections	Variance of the Sample Standard Deviation of the Dependent Variable	Variance of the Estimated Fraction of Individuals who Are Compliers	Covariance Between the Nominal *ITT* Estimator and the Fraction who Are Compliers
21st Century				
Reading score	0.01	0.01	0.02	-0.02
Math course grade	0.32	0.05	0.07	0.19
Education Technologies				
Grade 1 reading score	-0.14	0.00	0.02	-0.16
Grade 4 reading score	-0.15	0.01	0.04	-0.20
NYC Vouchers				
Reading score	0.05	0.02	0.04	0.00
Math score	0.21	0.02	0.04	0.15
Power4Kids				
Word attack score	3.31	3.29	0.09	-0.07
GRADE score	0.84	0.79	0.02	0.02
Teacher Induction				
Whether stay in district	0.07	0.01	0.00	0.06
Lesson implement score	0.05	0.00	0.00	0.04
Early Head Start				
Bayley MDI score	0.21	0.20	0.07	-0.06
HOME score	-0.13	0.11	0.04	-0.28
Job Corps				
Earnings	0.06	0.05	0.06	-0.05
Arrests	0.08	0.11	0.15	-0.18

Source: Data from studies listed in Appendix Table B.1.

Note: The indicated correction terms in the final three columns denote the second, third, and fourth terms, respectively, on the right hand side of equation (31). The percentage change in the estimated variance of the standardized *CACE* impact estimator due to the indicated correction term is equal to the indicated correction term divided by the first term on the right hand side of equation (31), and multiplied by 100.

Table 6.7: Simulated Effects of Variance Corrections on the Standard Error of $\hat{\alpha}_{CACE\_E}$, for an Assumed *CACE* Impact Value of 0.25 and by Study

Study and Dependent Variable	Percentage Change in the Standard Error of $\hat{\alpha}_{CACE\_E}$ due to the Correction if $\hat{\alpha}_{CACE\_E}$ Were Equal to 0.25	Absolute Change in the *p*-value (x 10^4) of $\hat{\alpha}_{CACE\_E}$ due to the Correction if $\hat{\alpha}_{CACE\_E}$ Were Equal to 0.25	Value of $\hat{\alpha}_{CACE\_E}$ so that the Correction Will Increase the Standard Error by 5 Percent
21st Century			
Reading score	0.86	0.011	0.626
Math course grade	0.30	0.043	0.800
Education Technologies			
Grade 1 reading score	1.46	0.004	0.404
Grade 4 reading score	3.33	0.000	0.298
NYC Vouchers			
Reading score	0.88	0.376	0.601
Math score	1.22	0.812	0.564
Power4Kids			
Word attack score	0.71	1.463	0.663
GRADE score	0.53	3.778	0.783
Teacher Induction			
Whether stay in district	0.21	0.221	0.913
Lesson implement score	0.07	0.285	0.980
Early Head Start			
Bayley MDI score	0.49	0.000	0.779
HOME score	0.22	0.000	0.807
Job Corps			
Earnings	0.39	0.000	0.854
Arrests	0.63	0.000	0.787

Source: Data from studies listed in Appendix Table B.1.

Note: Standard errors are regression-adjusted using the covariates indicated in Appendix Table B.1. The percentage change in the standard error due to the correction is equal to the difference between the corrected and uncorrected standard error, divided by the uncorrected standard error, and multiplied by 100.

Chapter 7: Summary and Conclusions

This report has examined the identification and estimation of the *CACE* parameter for two-level clustered RCTs that are commonly used in education research, where groups (such as schools or classrooms) rather than students are the unit of random assignment. We generalized the causal inference and IV framework developed by Angrist et al. (1996) to develop conditions for identifying the *CACE* parameter under clustered designs where multi-level treatment compliance decisions can be made by both school staff and students.

This report also provides simple asymptotic variance estimation formulas for *CACE* impact estimators measured in both nominal and standard deviation units. Because these IV impact estimators are ratio estimators, the variance formulas account for both the estimation error in the numerators (which pertain to the nominal *ITT* impact estimates) and the denominators (which pertain to the estimated service receipt rates and the estimated standard deviations of the outcomes).

Researchers sometimes assume that the denominator terms in these ratio estimators are known, and thus, present the *same p*-values from significance tests for all *ITT* and *CACE* impact estimates. This approach, however, could yield incorrect significance findings if the variance components due to the denominator terms matter. Accordingly, we used data from 10 large-scale RCTs in education and other social policy areas to compare significance findings for the considered impact estimates using uncorrected and corrected variance estimators.

Our key empirical finding is that the variance correction terms have very little effect on the standard errors of the standardized *ITT* and *CACE* impact estimators. Across the examined outcomes, the correction terms typically raise the standard errors by less than 1 percent, and change *p*-values at the fourth or higher decimal place. Furthermore, simulations indicate that, on average, the impact estimates would need to be 0.7 to 0.8 standard deviations, representing effect sizes that are rarely found in practice, before the variance corrections would raise the standard errors by 5 percent. These results occur because, by far, the most important source of variance in the considered ratio estimators is the variance of the nominal *ITT* impact estimators.

Despite these results, we advocate, for rigor, that education researchers use the correct standard error formulas for standardized *ITT* and *CACE* impact estimates. The formulas laid out in this report are relatively straightforward to apply, and their use will protect against the risk of finding incorrect significance findings, even if this risk is likely to be low based on our empirical findings.

Appendix A: Proof of Equation (29)

Following Hedges (2007), note that $n(m-1)S_W^2 / \sigma_W^2$ has an approximate chi-squared distribution with $n(m\text{-}1)$ degrees of freedom. Thus $AsyVar(S_W^2) = 2\sigma_W^4 / n(m-1)$. Similarly, $(n-2)S_B^2 / E(S_B^2)$ has an approximate chi-squared distribution with $n(m\text{-}1)$ degrees of freedom. Thus,

$$AsyVar(S_B^2) = 2(\sigma_B^2 + \{\sigma_W^2 / m\})^2 / (n-2).$$

Using (26), we find then that:

$$(A.1) \quad AsyVar(S_y^2) = \frac{2(\sigma_B^2 + \{\sigma_W^2 / m\})^2}{(n-2)} + \frac{[(m-1)/m]^2 2\sigma_W^4}{n(m-1)}.$$

To obtain a variance expression for S_y in terms of the variance expression for S_y^2 in (A.1), we apply a Taylor series expansion of S_y around σ_y:

$$(A.2) \quad S_y - \sigma_y \approx \frac{(S_y^2 - \sigma_y^2)}{2\sigma_y}.$$

Because S_y^2 is asymptotically normal, the delta method implies that S_y is asymptotically normal with the following asymptotic variance:

$$(A.3) \quad AsyVar(S_y) \approx AsyVar(S_y^2) / 4\sigma_y^2.$$

After some algebra, (29) follows after inserting (A.1) into (A.3) and replacing σ_B^2, σ_W^2, and σ_y^2 by their estimators.

Appendix B:

Table B.1: Summary of Data Sources

Study(Authors; Sponsor)[a]	Description of Program or Intervention and Study Design	Original Study Population and Number of Treatment and Control Observations Used in Current Analyses[b]	Level of Clustering (Intraclass Correlation Coefficient)[c]	Outcome Measures (Corresponding Estimate of Nominal *ITT* Impact from Published Study Report)	Baseline Covariates
Evaluation of the 21st Century Community Learning Centers Program (James-Burdumy et al. 2005; IES)	Study examined the effects of participation in after-school programs on academic and behavioral outcomes of elementary school students in 12 school districts and 26 centers. Students interested in attending after-school programs were randomly assigned to the treatment or control groups within each after-school program center.	Students in kindergarten to 6th grade in the 2000-2001 school year within 12 unspecified school districts 857; 796	None	Stanford-9 reading score in second year of study (0.3); math course grade in second year (-0.6)	Baseline test scores in reading and math; grade level; whether student is overage for grade; race/ethnicity; number of absences, tardies, and suspensions in year prior to study; whether student has been retained in any prior year; site indicators
Teach for America Evaluation (Decker et al. 2004; SRF; HF, CC)	Study examined the impact of teachers from Teach for America, a highly selective alternative certification program, on the academic achievement of elementary school students. Students were randomly assigned to classrooms taught by Teach for America teachers or traditional teachers in the same grade and school.	1st to 5th graders in the 2001-2002 school year; 17 schools in Baltimore, Chicago, Los Angeles, Mississippi Delta, and New Orleans 742; 911	Teacher (0.561)	Iowa Test of Basic Skills (ITBS) reading score (0.56); ITBS math score (2.43)	Baseline test scores in reading and math; grade level; school indicators

Table B.1: Summary of Data Sources

Study(Authors; Sponsor)[a]	Description of Program or Intervention and Study Design	Original Study Population and Number of Treatment and Control Observations Used in Current Analyses[b]	Level of Clustering (Intraclass Correlation Coefficient)[c]	Outcome Measures (Corresponding Estimate of Nominal *ITT* Impact from Published Study Report)	Baseline Covariates
Evaluation of Reading and Mathematics Education Technologies (Dynarski et al. 2007; IES)	Study examined the effects of 16 software products on students' academic achievement in 1st grade reading, 4th grade reading, 6th grade math, and algebra in 33 school districts. Within each participating school, teachers were randomly assigned to use a study product or not. For the purposes of our report, outcomes in 1st and 4th grades are used.	Students in 1st grade, 4th grade, 6th grade, and algebra classes in the 2004-05 school year in 33 districts 1,160; 777	Teacher (0.197)	1st grade Stanford-9 reading score (0.73); 4th grade Stanford-10 reading score (0.41)	Baseline test scores; student's age and gender; teacher's gender, experience, and highest degree; school's racial/ethnic composition; percent of school's students eligible for special education and subsidized lunch
New York City School Voucher Experiment (Mayer et al. 2002; SCSF)	Study examined the effects of three-year private school scholarship offers on the academic outcomes of children from low-income families. Eligible families who applied for scholarships for their children were randomly selected for scholarships in a series of lotteries.	Low-income children enrolled in kindergarten to fourth grade in 1997 in New York City public schools 672; 471	Family (0.436)	Iowa Test of Basic Skills (ITBS) reading score in third year of study (0.27); ITBS math score in third year (1.59)	Baseline test scores in reading and math; randomization strata indicators

Table B.1: Summary of Data Sources

Study(Authors; Sponsor)[a]	Description of Program or Intervention and Study Design	Original Study Population and Number of Treatment and Control Observations Used in Current Analyses[b]	Level of Clustering (Intraclass Correlation Coefficient)[c]	Outcome Measures (Corresponding Estimate of Nominal *ITT* Impact from Published Study Report)	Baseline Covariates
Power4Kids Study (Torgesen et al. 2006; IES)	Study examined the impact of four widely used remedial reading instructional programs on students' reading skills. 50 schools from 27 districts were randomly assigned to one of the interventions, and within each school eligible children who were identified as struggling readers were randomly assigned to receive the intervention or not. For the purposes of our report, the treatment group consists of students assigned to receive any of the four interventions, and third graders' outcomes are used.	3rd and 5th grade students in the 2003-04 school year within 27 school districts near Pittsburgh, PA who are identified as struggling readers 211; 127	None	Woodcock Reading Mastery Test-Revised Word Attack subtest score (5.0); Group Reading Assessment and Diagnostic Evaluation (GRADE) Passage Comprehension subtest score (4.6)	School indicators; baseline test scores
Evaluation of Comprehensive Teacher Induction Programs (Glazerman et al. 2008; IES)	Study examined the effects of comprehensive teacher induction programs on teacher retention, teachers' classroom practices, and student outcomes. The comprehensive programs provide beginning teachers with an orientation, mentoring sessions, and professional development.	Beginning teachers in elementary schools within 17 low-income school districts across 13 states in the 2005-06 school year 457; 425	School (0.125)	Binary variable indicating that the teacher stayed in the same district from the first year of the study to the start of the second year (0.002); teacher's score, as assigned by trained observer using the Vermont Classroom Observation Tool, for implementation of literacy lessons (0.0)	Grade level taught; teacher's age, gender, race/ethnicity, marital status, household structure, teaching experience, non-teaching experience, certification status, preparation type, educational

Table B.1: Summary of Data Sources

Study(Authors; Sponsor)[a]	Description of Program or Intervention and Study Design	Original Study Population and Number of Treatment and Control Observations Used in Current Analyses[b]	Level of Clustering (Intraclass Correlation Coefficient)[c]	Outcome Measures (Corresponding Estimate of Nominal *ITT* Impact from Published Study Report)	Baseline Covariates
Teacher Induction (continued)	Within 17 participating districts, elementary schools were randomly assigned to participate in comprehensive induction programs or to use their district's existing induction program.				attainment, college quality, residential location, and homeownership status; school's racial/ethnic and socioeconomic composition; district indicators
Evaluation of Early Head Start (Love et al. 2002; ACF)	Study examined the impacts of Early Head Start, which provides center-based or home-based services to families with children aged 0 to 3, on child development and parenting outcomes. Within each of 17 participating programs, eligible applicants were randomly assigned to receive Early Head Start services or not.	Low-income families with infants and toddlers aged 0 to 3 or pregnant women who applied to a study site in 1996 879; 780	None	Bayley Mental Development Index (MDI) score (1.456); Home Observation for Measurement of the Environment (HOME) total score (0.455)[d]	Mother's age, race/ethnicity, English ability, education, employment status, living arrangements, number of children, household income, welfare receipt, resource adequacy, mobility, and random assignment date; child's age, birth weight status, premature birth status, gender, and previous Head Start enrollment; site indicators

Table B.1: Summary of Data Sources

Study(Authors; Sponsor)[a]	Description of Program or Intervention and Study Design	Original Study Population and Number of Treatment and Control Observations Used in Current Analyses[b]	Level of Clustering (Intraclass Correlation Coefficient)[c]	Outcome Measures (Corresponding Estimate of Nominal *ITT* Impact from Published Study Report)	Baseline Covariates
National Job Corps Study (Schochet et al. 2001; DOL)	Study examined the impacts of Job Corps, a large federal program providing educational and vocational training services to disadvantaged youth aged 16-24 in a residential setting, on employment and related outcomes. Among youths applying to Job Corps in a thirteen-month period, a subset of applicants was randomly offered enrollment in Job Corps.	Disadvantaged youth between the ages of 16 and 24 who applied to Job Corps in 1995 and were determined eligible 6,518; 4,298	None	Weekly earnings in the fourth year of the study (15.9); total number of arrests during the four years of the study (-0.09)	None
San Diego Food Stamp Cash-Out Experiment (Ohls et al. 1992; FNS)	Study examined the effects of cashing-out food stamps on food-purchasing and food-use patterns of Food Stamp Program (FSP) participants in San Diego County. FSP households were randomly assigned to the cash-out status or the regular coupon status.	FSP recipients in 1989 in San Diego County 613; 613	None	Money value of purchased food used at home by the household in the last seven days (-5.17); average availability of food energy per equivalent nutrition unit as a percentage of the recommended daily allowance (RDA) (-6.42)	None

Table B.1: Summary of Data Sources

Study(Authors; Sponsor)[a]	Description of Program or Intervention and Study Design	Original Study Population and Number of Treatment and Control Observations Used in Current Analyses[b]	Level of Clustering (Intraclass Correlation Coefficient)[c]	Outcome Measures (Corresponding Estimate of Nominal *ITT* Impact from Published Study Report)	Baseline Covariates
Teenage Parent Demonstration (Maynard et al. 1993; ACF)	Study examined the impacts of a demonstration program in 3 cities for teenage welfare mothers. The program required that welfare mothers participate in employment, job training, or education activities in order to receive full welfare benefits and also provided child care and transportation assistance. All first-time teenage welfare mothers were randomly assigned to be subject to the enhanced set of requirements and services or not. For the purposes of our report, outcomes from Chicago are used.	Teenage mothers who applied for AFDC for the first time in 1987 in Camden NJ, Newark NJ, and Chicago, IL 805; 822	None	Percentage of months active in employment, job training, or education activities (6.1); average monthly earnings (24.4)	Teenage parent's race/ethnicity, living arrangements, health barriers to work, English proficiency, contact with father of child, diploma status, school enrollment status, math skills, prior work experience, and date of study entry; teen's residency with own father and residency in welfare household as child; education of teen's mother; age of teen's child

[a] Acronyms are defined as follows: IES = Institute of Education Sciences at the U.S. Department of Education; SRF = Smith Richardson Foundation; HF= Hewlett Foundation; CC=Carnegie Corporation; SCSF = School Choice Scholarships Foundation; ACF = Administration for Children and Families at the U.S. Department of Health and Human Services; DOL = U.S. Department of Labor; FNS = Food and Nutrition Service of the U.S. Department of Agriculture.

[b] For each study, the size of the treatment and control group pertains to the sample used for estimating impacts on the first listed outcome variable.

[c] For each study, the intraclass correlation coefficient pertains to the first listed outcome variable and is estimated by the authors.

[d] The published report for the evaluation of Early Head Start reported nominal *CACE* impacts. Nominal *ITT* impacts are calculated as the published nominal *CACE* impacts multiplied by 0.91, the reported fraction of individuals in the study who are compliers.

Table B.2: Information on the Receipt of Intervention Services, by Study

Study	Definition of Service Receipt Used in Current Analysis	Same as Definition Used in Published Study?	Estimated Fraction of Individuals in the Treatment Group who Received Intervention Services	Estimated Fraction of Individuals in the Control Group who Received Intervention Services
21st Century	Student attended a program center for at least 1 day in the 2-year study period	Yes	0.926	0.177
Education Technologies	Average student time using product in teacher's classroom was above 25% of the treatment group mean	No *CACE* analyses in published study	0.885	0.000
NYC Vouchers	Child attended private school in any year of study	Yes	0.854	0.118
Power4Kids	Student received positive hours of intervention	Yes	0.991	0.000
Teacher Induction	Teacher was assigned a mentor in a comprehensive induction program	No *CACE* analyses in published study	0.950	0.003
Early Head Start	Family received at least a minimal set of Early Head Start services	Yes	0.916	0.000
Job Corps	Individual was ever enrolled in a Job Corps center in first three years of study	Yes	0.729	0.012

Source: Data from studies listed in Appendix Table B.1.

Note: Estimated fractions of individuals who receive intervention services are averages of unit-level rates of service receipt in the relevant treatment status group and are not adjusted for covariates. For each indicated study, the estimation sample is the same as that used for estimating the *CACE* impact on the first outcome variable listed in Appendix Table B.1.

Table B.3: Analytical and Bootstrap *P*-Values of $\hat{\alpha}_{ITT\_E}$ and $\hat{\alpha}_{CACE\_E}$, by Study

Study and Dependent Variable	*P*-value of $\hat{\alpha}_{ITT\_E}$			*P*-value of $\hat{\alpha}_{CACE\_E}$		
	Analytical	Bootstrap	Difference	Analytical	Bootstrap	Difference
21st Century						
Reading score	0.509	0.513	-0.004	0.509	0.513	-0.004
Math course grade	0.201	0.198	0.003	0.202	0.196	0.006
Teach for America						
Reading score	0.650	0.661	-0.012			
Math score	0.022	0.032	-0.010			
Education Technologies						
Grade 1 reading score	0.759	0.777	-0.018	0.759	0.779	-0.020
Grade 4 reading score	0.661	0.651	0.010	0.661	0.649	0.012
NYC Vouchers						
Reading score	0.533	0.532	0.002	0.533	0.530	0.003
Math score	0.515	0.511	0.004	0.515	0.512	0.003
Power4Kids						
Word attack score	0.000	0.000	0.000	0.000	0.000	0.000
GRADE score	0.022	0.024	-0.003	0.022	0.025	-0.003
Teacher Induction						
Whether stay in district	0.676	0.685	-0.009	0.676	0.684	-0.008
Lesson implement score	0.861	0.862	-0.001	0.861	0.861	0.000
Early Head Start						
Bayley MDI score	0.011	0.009	0.001	0.011	0.009	0.001
HOME score	0.044	0.039	0.005	0.044	0.039	0.004
Job Corps						
Earnings	0.002	0.004	-0.002	0.002	0.004	-0.002
Arrests	0.000	0.000	0.000	0.000	0.000	0.000
Cash-Out						
Value of purchased food	0.013	0.013	0.000			
Energy as percent of RDA	0.050	0.047	0.003			
Teenage Parent						
Percent of months active	0.000	0.000	0.000			
Earnings	0.146	0.141	0.005			

Source: Data from studies listed in Appendix Table B.1.

Note: The analytical *p*-values of the standardized *ITT* [or *CACE*] impact estimate were obtained using (28) [or (31)]. The bootstrap *p*-values were obtained using the following steps: (1) Obtain the analytical *t*-statistic of the standardized impact estimate, as described previously; (2) Draw a stratified simulated sample of np treatment units and $n(1-p)$ control units by sampling with replacement; (3) Use the simulated sample to calculate the simulated *t*-statistic, which is equal to the difference between the simulated and analytical standardized impact estimate, divided by the simulated value of the square root of (28) [or (31)]; (4) Repeat steps (2) and (3) an additional 4,999 times to obtain a total of 5,000 simulated *t*-statistics; (5) Calculate the proportion of simulated samples for which the absolute value of the simulated *t*-statistic exceeds the absolute value of the analytical *t*-statistic.

References

Angrist, J., G. Imbens, and D. Rubin (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association* 91, 444-472.

Bloom, H. (1984). Accounting for No-Shows in Experimental Evaluation Designs. *Evaluation Review* 8(2), 225-246.

Bryk, A. and S. Raudenbush (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods.* Newbury Park, CA: Sage.

Cochran, W. (1963). *Sampling Techniques*. New York: John Wiley and Sons.

Cohen, J. (1988). *Statistical Power Analysis for Behavioral Sciences.* Hillside, NJ: Lawrence Erlbaum.

Decker, P., D. Mayer, and S. Glazerman (2004). The Effects of Teach For America on Students: Findings from a National Evaluation. Princeton, NJ: Mathematica Policy Research, Inc.

Donner, A. and N. Klar (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. London: Arnold.

Dynarski, M., R. Agodini, S. Heaviside, T. Novak, N. Carey, L. Campuzano, B. Means, R. Murphy, W. Penuel, H. Javitz, D. Emery, and W. Sussex (2007). Effectiveness of Reading and Mathematics Software Products: Findings from the First Student Cohort. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Freedman, D. (2008). On Regression Adjustments to Experimental Data. *Advances in Applied Mathematics*, 40, 180-193.

Glazerman, S., S. Dolfin, M. Bleeker, A. Johnson, E. Isenberg, J. Lugo-Gil, M. Grider, and E. Britton (2008). Impacts of Comprehensive Teacher Induction: Results from the First Year of a Randomized Controlled Study. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Greene, W. (2000). *Econometric Analysis.* Fourth Edition. Upper Saddle River, NJ: Prentice Hall.

Heckman, J., J. Smith, and C. Taber (1994). Accounting for Dropouts in Evaluations of Social Experiments. NBER Working Paper NO. 166.

Hedges, L. (1981). Distribution Theory for Glass's Estimator of Effect Size and Related Estimators. *Journal of Educational Statistics, 6*, 107-128.

Hedges, L. (2007). Effect Sizes in Cluster-Randomized Designs. *Journal of Educational and Behavioral Statistics, 32*(4), 341-370.

Holland, P. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396), 945-960.

Imbens, G. and D. Rubin (2007). *Causal Inference: Statistical Methods for Estimating Causal Effects in Biomedical, Social, and Behavioral Sciences,* Cambridge University Press.

James-Burdumy, S., M. Dynarski, M. Moore, J. Deke, W. Mansfield, and C. Pistorino (2005). When Schools Stay Open Late: The National Evaluation of the 21st Century Community Learning Centers Program: Final Report. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Jo, B., T. Asparouhov, B. Muthen, and N. Ialongo (2008). Cluster Randomized Trials with Treatment Noncompliance. *Psychological Methods* 13(1), 1-18.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.

Liang, K. and S. Zeger (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika* 73, 13-22.

Lipsey, M.W. and D.B. Wilson (1993). The Efficacy of Psychological, Educational, and Behavioral Treatment. *American Psychologist*, 48(12), 1181-1209.

Littell, R., G. Milliken, W. Stroup, and R. Wolfinger (1996). *SAS System for Mixed Models.* Cary, NC: SAS Institute Inc.

Little, R., Q. Long, and X. Lin (2008). A Comparison of Methods for Estimating the Causal Effect of a Treatment in Randomized Clinical Trials Subject to Noncompliance. *Biometrics*, forthcoming.

Love, J., E. Kisker, C. Ross, P. Schochet, J. Brooks-Gunn, D. Paulsell, K. Boller, J. Constantine, C. Vogel, A. Fuligni, and C. Brady-Smith (2002). Making a Difference in the Lives of Infants and Toddlers and Their Families: The Impacts of Early Head Start: Volume I: Final Technical Report. Princeton, NJ: Mathematica Policy Research, Inc.

Mayer, D., P. Peterson, D. Myers, C. Tuttle, and W. Howell (2002). School Choice in New York City After Three Years: An Evaluation of the School Choice Scholarships Program: Final Report. Washington, DC: Mathematica Policy Research, Inc.

Maynard, R., W. Nicholson, and A. Rangarajan (1993). Breaking the Cycle of Poverty: The Effectiveness of Mandatory Services for Welfare-Dependent Teenage Parents. Princeton, NJ: Mathematica Policy Research, Inc.

Murray, D. (1998). *Design and Analysis of Group-Randomized Trials*. Oxford: Oxford University Press.

Murray, M. (2006). Avoiding Invalid Instruments and Coping with Weak Instruments. *Journal of Economic Perspectives*, 20(4), 111-132.

Ohls, J., T. Fraker, A. Martini, and M. Ponza (1992). The Effects of Cash-Out on Food Use by Food Stamp Program Participants in San Diego. Princeton, NJ: Mathematica Policy Research, Inc.

Rubin, D. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Education Psychology*, 66, 688-701.

Rubin, D. (1977). Assignment to Treatment Group on the Basis of a Covariate. *Journal of Education Statistics*, 2(1), 1-26.

Schochet, P. (2007). Is Regression Adjustment Supported by the Neyman Model for Causal Inference? Working Paper: Mathematica Policy Research, Inc.: Princeton NJ.

Schochet, P. (2008). Statistical Power for Random Assignment Evaluations of Education Programs. *Journal of Educational and Behavioral Statistics, 33*(1), 62-87.

Schochet, P., J. Burghardt, and S. Glazerman (2001). National Job Corps Study: The Impacts of Job Corps on Participants' Employment and Related Outcomes. Princeton, NJ: Mathematica Policy Research, Inc.

Stock, J., J. Wright, and M. Yogo (2002). A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments. *Journal of Business and Economic Statistics*, 20(4), 518-529.

Torgesen, J., F. Stancavage, D. Myers, A. Schirm, E. Stuart, S. Vartivarian, W. Mansfield, D. Durno, R. Javorsky, and C. Haan (2006). Closing the Reading Gap: First Year Findings from a Randomized Trial of Four Reading Interventions for Striving Readers: Final Report. Washington, DC: U.S. Department of Education, Institute of Education Sciences.