Title

Sensitivity analyses for clustered data: An illustration from a large-scale clustered randomized controlled trial in education

Permalink
https://escholarship.org/uc/item/1bd479tw

Journal
Evaluation and Program Planning, 47

ISSN
01497189

Authors
Abe, Yasuyo
Gee, Kevin A

Publication Date
2014-12-01

DOI
10.1016/j.evalprogplan.2014.07.001

Peer reviewed

Published in *Evaluation and Program Planning*

Suggested Citation

Yasuyo, A., Gee, K.A. (2014). Sensitivity analyses for clustered data: An illustration from a large-scale clustered randomized controlled trial in education. *Evaluation and Program Planning, 47: 26-34*

Sensitivity Analyses for Clustered Data: An Illustration from a Large-Scale Clustered Randomized Controlled Trial in Education

Kevin A. Gee, Ed.D.
Assistant Professor
University of California, Davis
School of Education
One Shields Ave.
Davis, CA 95616
(530) 752-9334

Abstract

In this paper, we demonstrate the importance of conducting well-thought-out sensitivity analyses for handling clustered data (data in which individuals are grouped into higher order units, such as students in schools) that arise from cluster randomized controlled trials (RCTs). This is particularly relevant given the rise in rigorous impact evaluations that use cluster randomized designs across various fields including education, public health and social welfare. Using data from a recently completed cluster RCT of a school-based teacher professional development program, we demonstrate our use of four commonly applied methods for analyzing clustered data. These methods include: (1) Hierarchical Linear Modeling (HLM); (2) Feasible Generalized Least Squares (FGLS); (3) Generalized Estimating Equations (GEE); and (4) Ordinary Least Squares (OLS) regression with cluster-robust (Huber-White) standard errors. We compare our findings across each method, showing how inconsistent results—in terms of both effect sizes and statistical significance—emerged across each method and our analytic approach to resolving such inconsistencies.

Keywords

Impact Evaluation; Sensitivity Analyses; Data Analysis for Clustered Randomized Trials; Hierarchical Linear Modeling; Generalized Estimating Equations

1. Introduction

Cluster randomized controlled trials[1] (RCTs) have become an increasingly popular way to evaluate the impact of interventions which are applicable to intact groups of individuals. Common examples include schools that are randomly assigned to offer its students an educational intervention. Similarly, there are studies in which clinics are randomized to offer a particular treatment to an intact group of patients it serves. One notable feature of such trials is that individuals (e.g., students or patients) are clustered together in higher level units (e.g., schools or clinics) with the higher level unit serving as the unit of randomization. [2] Evaluators who analyze data from clustered RCTs must select from a variety of methods that appropriately account for the correlation between study participants within the higher level units. Ignoring such correlation, especially when the correlation between individuals within clusters is relatively high (as captured by the intra-class correlation coefficient (ICC) may lead to erroneous inferences due to downward biased standard errors (Garson, 2012; Hox, 2010; Liang and Zeger (1993); Zyzanski, Flocke, & Dickinson, 2004).

For evaluation analysts, deciding upon which method to use when analyzing clustered data is not an exact science. Often, the choice depends upon a combination of factors including analysts' professional judgment and their prior quantitative training. Also, the choice of method is driven by the methodological conventions and traditions of the disciplinary field (e.g., public health, education, etc.) in which the evaluation is conducted. However, one overarching principal is that analysts are entrusted to choose the most appropriate approach among various data analytic methods, *prior to conducting analyses*, based on their prior assessment of the design and data limitations. This prevents researchers from selecting, or being suspected of selecting, a particular analytic method to influence the results.

[1] *Cluster* randomized trials are also commonly referred to as *place-based* or *group* randomized trials (Boruch, 2005, p. 14).

[2] Often, these clustered structures (e.g., students within schools) are also referred to as multi-level, hierarchical or nested structures.

Yet, when accounting for clustering, analysts often rely only upon one preferred methodological approach without considering how and if the results remain consistent across different methods. Carrying out analyses using different methods and checking for the consistency in results across such methods is one class of a broader set of *sensitivity analyses* (Thabane et al., 2013) which analysts often undertake. We believe that well-thought-out sensitivity analyses to handle clustered data and the transparent reporting of such analyses are important, particularly as different methods can and—as we show in our case—lead to discrepant findings. When conflicting findings emerge across different methodological approaches, we believe that evaluation analysts must then proceed to understand the conflicting results, plan alternate analyses to reconcile such findings, and carefully document those alternative approaches. Finally, analysts should be transparent in communicating their analytic decisions to their evaluation audience.

In this paper, we review our results from a recently completed cluster randomized trial of a teacher professional development program. We compare our results across four methods we used to account for clustering in our data: (1) Hierarchical Linear Modeling (HLM); (2) Feasible Generalized Least Squares (FGLS); (3) Generalized Estimating Equations (GEE); and (4) Ordinary Least Squares (OLS) regression with robust clustered (Huber-White) standard errors. Importantly, we show how inconsistent results emerged across these different methods and our approach to resolving inconsistencies. We present and discuss our work primarily from an *applied* point of view, forgoing technical descriptions of the methods we have employed (with the exception of the statistical model we present for our main analytic approach using HLM). We do assume, however, that readers have basic familiarity with statistical concepts and the analytic issues that arise due to clustered data.

We structure the rest of our paper in five sections. In Section 2, we briefly review clustered randomized controlled trials and introduce the concept of the *intra-cluster correlation coefficient* (ICC). The ICC is a key quantitative measure capturing the extent to which individuals are correlated within an intact group. We also discuss sensitivity analyses for clustered data, methods for handling clustered data and prior empirical studies that have compared methods for clustered data. Next, in Section 3, we describe our research design, providing background about our study intervention, the site and sample as well as our data and measures. In Section 4, we describe our primary analytic method along with our selected alternative methods. Then, in Section 5, we present results from the four analytic approaches we used to analyze our data, discussing the inconsistencies that emerged across the methods and ways in which we reconciled those inconsistent results. Finally, in Section 6, we close with several substantive "lessoned learned" of our work, providing advice to evaluation analysts who face the task of analyzing clustered data.

2. Clustered Randomized Trials and Clustered Data

A *cluster* randomized controlled trial (RCT) refers to an experiment in which intact groups of individuals are randomly assigned to receive an offer to participate in a treatment or not[3]. The groups that do not receive an offer of the treatment serve as the control group. This is in contrast to a standard RCT in which *individuals* are randomly assigned into a treatment or control group. The level at which randomization occurs—whether it be at the group or the individual-level—is commonly referred to as *the unit of randomization*. The cluster is the unit of randomization in numerous experimental evaluations of programs in education, public health and criminology (Boruch, 2005). Randomizing clusters of individuals not only avoids potential cross contamination between control and treatment conditions, but the interventions themselves are often designed to be administered to intact groups rather than individuals (Raudenbush, 1997). Finally, there may be ethical issues that can be ameliorated by randomizing at the cluster level. For example, in a ground-breaking study of an incentive-based cash subsidy program in Mexico known as *Progresa* (now *Opportunidades*), intact communities rather than households were randomly assigned to receive an

[3] Here, we assume the most basic design of a randomized experiment with only one treatment and one control condition. There are, of course, various randomized designs that have multiple treatment and control conditions.

offer of a subsidy or not (Parker & Teruel, 2005). Randomizing households within these relatively small and close-knit communities could have created tension between treatment and control group households (Parker & Teruel, 2005). Also, randomization could have led to a "perception of discretionality" (Parker & Teruel, 2005, p. 208) with respect to which households—despite being equally eligible—were selected to receive subsidies or not.

When analyzing data from cluster RCTs, evaluation analysts often want to understand the impact of a program, on average, *across individuals'* outcomes even though these individuals are part of an existing intact group. For example, in the evaluation of the *Progresa* program, researchers wanted to understand whether children living in communities randomized to receive cash subsidies had improved health outcomes versus children in control communities (Gertler, 2004). To determine the impact of the program on individuals' outcomes in a RCT with individual-level random assignment, an analyst may apply standard t-tests to compare the means of outcome measures collected on individuals assigned into the control condition versus the treatment condition or to apply ordinal least squares (OLS) regression to test the estimate effects on the treatment condition. However, such a strategy, if applied to a cluster RCT, ignores the fact that individuals are members of existing groups and may not be completely independent of each other—a critical assumption of standard statistical techniques such as the t-test or OLS regression. Ignoring clustering can lead to erroneous inferences (Garson, 2012; Hox, 2010; Liang & Zeger, 1993; Zyzanski et al., 2004) due to standard errors that are biased downwards (Clarke, 2008; Steenbergen & Jones, 2002) leading to inflated Type I error rates (i.e., stating that there is an effect when there is not). Modeling the degree to which individuals are correlated within clusters requires different methods, such as the ones we illustrate in this paper.

The degree to which individuals are interdependent within a cluster can be quantitatively measured by the intra-class correlation coefficient (ICC), often denoted by the Greek symbol ρ (rho)

(Killip, Mahfoud, & Pearce, 2004). In the most basic case where individuals (e.g., students) are clustered into higher level units (e.g. schools), the ICC is calculated as the ratio of the between-cluster variance on a particular continuous outcome measure of interest (e.g. achievement) to the total variance (the between- plus within-cluster variance) of that outcome. The ICC can be expressed as:

$$\frac{\sigma^2_{between}}{\sigma^2_{between} + \sigma^2_{within}} \quad (1)$$

where $\sigma^2_{between}$ represents the between-cluster variance and σ^2_{within} is the within-cluster variance. The ICC ranges from 0 to 1, with values closer to 1 indicating a higher degree of correlation for a particular outcome of interest within an intact group.[4] If there is no variability between clusters, the ICC would equal 0 ($\frac{0}{0+\sigma^2_{within}} = 0$). This suggests that individuals' outcomes are independent of each other. In other words, all of the variation lies between individuals and there is no correlation between individuals within a cluster. On the other hand, in the instance where all individuals are homogenous on an outcome so that there is no within-cluster variance, the ICC would equal 1 ($\frac{\sigma^2_{between}}{\sigma^2_{between}+0}$ =1). The ICC can also be interpreted in percentage form. For example, for an ICC that equals .35, we can say that 35% of the total variation in a particular outcome lies between clusters, while the remaining 65% lies between individuals.

The clustering of individuals (e.g., clustering of students within schools) increases the standard errors of the effect estimates because of the correlation across observations within clusters. A non-zero ICC, which measures the degree of such correlations, thus indicates the presence of clustering effects and warns researchers against applying the conventionally estimated "incorrect"

[4] There are instances in which the ICC can be negative (see Lohr, 2010, p. 175); however, as Lohr (2010) notes, "The ICC is rarely negative in naturally occurring clusters (p. 175).

variance estimates in inference tests. Without addressing clustering effects, such tests could lead to potentially misleading conclusions (Hox, 2010; Schochet, 2008).

2.1 Sensitivity Analyses for Clustered Data from RCTs

Sensitivity analyses refer to a set of ancillary analyses that analysts undertake to assess the robustness of a study's findings (Thabane et al., 2013) lending both credibility and rigor to any empirical studies. Sensitivity analyses vary in their intent and purpose ranging from alternative ways analysts handle missing data to the types of analyses that are the focus of this paper—different analytic approaches for clustered (or correlated) data structures. As is typical in conducting large-scale cluster RCTs in the social sciences, analysts tend to specify sensitivity analyses as part of a comprehensive study protocol that includes a detailed discussion of proposed data analytic methods. We strongly advise analysts who design such protocols, particularly when it is known that data will be clustered (as in a clustered RCT), to develop and articulate *a priori* a sensible set of sensitivity analyses to handle clustered data. This is important not only for the sake of transparency but to preserve the scientific and ethical rigor of cluster RCTs.

2.2 Approaches for Analyzing Clustered Data

Below, we briefly describe the four different approaches for analyzing clustered data that we illustrate in this paper and address some practical tradeoffs between the methods. Then, we discuss studies that have examined how results compare across these approaches and address important considerations for analyzing clustered data.

(1) Hierarchical Linear Modeling (HLM): HLM (also known as random coefficient models, mixed-level models, or multilevel models) explicitly models clustered data structures by specifying how an outcome indicator is explained by explanatory variables at different levels that are nested within each other. For example, in the case where students are nested in schools, there are two levels (students and schools) and thus analysts would specify a

student-level model and a school level model. Importantly, these models allow analysts to account for the variances in the outcome that arise from different levels of the clustered data; in the school example, HLM allows analysts to appropriately model variation in the outcome that lies between schools (i.e., the between cluster variation). HLM can accommodate multiple levels of clustering (e.g., children in classrooms in schools). For a thorough overview of HLM, readers should consult Raudenbush and Bryk (2002) and Hox (2010). For an important discussion of the limitations of HLM, readers should consult Gelman (2006).

(2) Feasible Generalized Least Squares (FGLS). FGLS is an estimation method that models the correlated nature of errors and leads to estimates that can be more efficient[5] than standard OLS estimates (Cameron & Miller, 2010). FGLS estimates are derived from a multistep process in which individual level residuals are obtained by fitting a first stage OLS regression model; these residuals are then used to estimate a variance-covariance matrix of the errors. This estimated matrix, which is specified so that the errors can be correlated, is then used to obtain feasible GLS estimation of model's regression coefficients. One important practical consideration, as Wooldridge (2010) notes, is that properties of the FGLS estimators can be poor for finite (i.e., small) samples. For a detailed discussion of FGLS in the context of clustered data see Cameron and Miller (2013).

(3) Generalized Estimating Equations (GEE). GEE (Liang & Zeger, 1993), an extension of general linearized models (Burton, Gurrin, & Sly, 1998), explicitly models the correlated nature of the data. Unlike HLM which accounts for clustering via random effects, GEE models rely on estimating how the model residuals are correlated within each cluster. These residual estimates and the correlation structure of these residuals (which analysts often

[5] This assumes that the structure of the errors is correctly specified.

specify *a priori*) are used in an iterative process to arrive at the coefficient estimates (Burton et al., 1998). As argued by Hubbard et al. (2010), GEE models versus HLM (or mixed models), provide a better "approximation of the truth" (p. 467); in other words, GEE models more accurately describe the underlying population parameters of interest. Also, as Gelman (2006) notes, though both approaches are similar, models based on HLM allow the estimates of parameters to vary across groups (e.g. schools) while GEE models focus on estimating parameters that do not vary, but account for clustering. Readers further interested in the technical differences between GEE and HLM as well as issues to consider when deciding between GEE and HLM should consult Hubbard et al. (2010).

(4) Ordinary Least Squares (OLS) regression with robust clustered (Huber-White) standard errors. To account for the non-independence of observations (and thus correlated errors) that can lead to downward biased estimates from OLS regression, analysts can calculate robust clustered standard errors for models fit using OLS regression (Cameron & Miller, 2010; White, 1980, 1982). Robust clustered standard errors, calculated using a sandwich variance estimator (Burton et al., 1998), tend to be adjusted upwards, leading to wider confidence intervals and decreasing the risk of making Type I errors. In our study, we estimated robust standard errors by relaxing the independence assumptions within clusters, while still assuming that observations are assumed independent across clusters.

There are numerous examples of studies, particularly from the medical literature (Thabane et al., 2013) as well as in political science (Steenbergen & Jones, 2002), comparing results across analytic methods for clustered data. For example, from the political science literature, Arceneaux and Nickerson (2009) compare results across clustered robust SEs, random effect models, HLM and aggregated OLS. Similarly, Zorn (2006) compares GEE with robust standard errors. From the medical field, Galbraith, Daniel, and Vissel (2010) use simulated data to illustrate how results can vary across a range of methods, including Linear Mixed Models (i.e., HLM) as well as GEE. Finally, Peters, Richards, Bankhead, Ades, and Sterne (2003) compare results across eight methods for analyzing data from a cluster RCT of a screening intervention, ranging from completely ignoring clustering to both GEE and random effects models.

Across these studies, there are several important lessons about analyzing clustered data and choosing a method for handling clustering.

(1) There is strong consensus that completely ignoring clustering will lead to erroneous inferences even if the degree to which observations are correlated is relatively small (Peters et al., 2003).

(2) Different analytic approaches can, but not necessarily always, lead to different results. Both Peters et al. (2003) and Galbraith et al. (2010) demonstrate that different point estimates and statistical significance of parameter estimates can arise due to different analytic methods, while both Arceneaux and Nickerson (2009) and Zorn (2006) demonstrate relatively consistent estimates across methods. Interestingly, though Zorn (2006) finds results under GEE and robust standard errors fairly consistent, the unit of clustering matters as well.

(3) Finally, as Peters et al. (2003) note, the number of clusters as well as the distributional assumptions of the data within clusters (i.e., the extent to which data are normally distributed or skewed) can influence the overall findings.

3. Research Design

3.1 Site and Intervention

We illustrate our results from analyzing clustered data in the context of a recently completed Institute of Education Science (IES) supported cluster RCT of a teacher professional development program known as the Pacific Communities with High Performance in Literacy Development (Pacific CHILD) (Authors, 2012). This evaluation was conducted across three sites in the Pacific region: Hawaii, American Samoa and the Commonwealth of the Northern Marian Islands (CNMI). Pacific CHILD provided fourth and fifth grade English language arts teachers with approximately 300 hours of professional development over two years, focusing on reading instruction. The year-round program consisted of off-campus training institutes, on-campus sessions that include one-on-one coaching and in-class demonstrations, and peer group meetings. This unusually intensive program aimed to improve teacher quality in reading instruction, a critical area of concern in the region, and to improve students' reading achievement. The impact analyses and subsequent sensitivity analyses we describe in this paper were aimed at evaluating the effect of Pacific CHILD on students' reading comprehension.[6]

3.2 Sample and Randomization Scheme

Our study sample was based on a convenience sample of 45 elementary schools across three sites (which, for the purposes of this study, we abbreviate as HI (Hawaii); AS (American Samoa) and the CNMI (Commonwealth of the Northern Mariana Islands)). Of the 45 schools, 23 were randomly assigned to the treatment condition, and 22 to the control condition. Random assignment occurred within strata that were based upon geographical location as well as size. This ensured that the resulting allocation of schools in the treatment and control groups was balanced, both in the number of schools and key school demographic characteristics.

The student impact sample consisted of all grade 5 students enrolled in the 45 study schools at the time of data collection in the spring of the second year of the intervention[7]. In total, 3,078 students and met the impact sample criteria. From this universe, students for whom outcome data were collected comprised the data analytic sample. Due to attrition and non-response our final

[6] The study also examined the causal impact of Pacific CHILD on teachers' knowledge and their classroom practices. However, for the purposes of this paper, we focus only on our student outcome measure.

[7] Unlike a randomized trial that follows a particular cohort of students over time, in our study, our sample includes all grade 5 students who were at the study schools near the *end* of the second year of the intervention, irrespective of whether or how much they had been potentially or actually exposed to the intervention. Thus, the impact study measured the intent-to-treat effects of Pacific CHILD as implemented in the field.

sample included 3,052 students. Of those students, 1,566 were in treatment schools and 1,486 were in control schools.

3.3 Measures

3.3.1. Outcome Variable. To assess the impact of the program on student achievement in reading, we utilized reading comprehension subtests of two national, norm-referenced tests administered annually in the study sites (SAT 10 and TerraNova). To create a common metric across these two different tests, we used published norming tables and equipercentile methods (Kolen & Brennan, 2004). The analysis sample for measuring impacts on achievement in reading consisted of 3,052 students.

3.3.2. Covariates. In Table 1, we list the covariates we included in our analysis. We included both school- and student-level covariates. We also controlled for baseline characteristics of schools and students, as well as indicator variables for sampling strata. Including these covariates increased the study's statistical power by improving the precision of the impact estimates and helped remove any chance differences in the baseline characteristics of treatment and control group schools.

3.3.3. Baseline comparison of treatment and control groups. To evaluate whether random assignment resulted in statistically equivalent groups at baseline, we compared selected school-level baseline characteristics of the treatment and control group schools in the impact sample. In Table 2, we provide means and standard deviations of these baseline characteristics disaggregated by treatment and control group status. We also report the standardized mean differences between the treatment and groups along with the *p*-value for the t-test on the mean differences. We collected school-level data from the U.S. Department of Education's Common Core of Data, enrollment records, and student test records to check baseline equivalence. School-level characteristics included school size, the student-teacher ratio, the percentage of students eligible for free or reduced-price lunch, student race/ethnicity, and student achievement in reading. In Hawai'i, we also compared the proportion of

English language learner students (reliable official data on English language learner status were not available for American Samoa or the CNMI). Student test records at baseline were available for grade 5 students in the CNMI and Hawai'i and for grade 4 students in American Samoa and Hawai'i. We compared each grade separately. Finally, we compared school level characteristics by averaging the characteristics within schools, which served as the unit of random assignment. As shown in column 5 of Table 2, the estimated standardized mean differences across all baseline characteristics range from -0.38 to 0.12. As shown in column 6 of Table 2, none of these differences were statistical significant at the $\alpha=.05$ level (*p*-values ranged from .26 to .98).

4. Data Analytic Plan

4.1 Main estimation method: Hierarchical Linear Modeling (HLM)

As is typical in conducting data analyses for cluster RCTs in education, we proposed *a priori* to fit a two-level hierarchical linear model (HLM) to our data, in which the first level (students) was nested in the second level (schools)[8]. We also decided to fit our model using data that pooled all individuals across all three sites in our study, thus yielding an unbiased estimate of program impact irrespective of potential cross-site heterogeneity.

More formally, we specified a two-level random-intercept model, in which the first level (student-level) was nested in the second level (school-level). In particular, for individual student *i* in school *j*, our hierarchical linear model was specified as the following system of equations (Raudenbush & Bryk, 2002):

[8] In our study, students are also clustered within classrooms. However, we did not account for this level of clustering in our analysis because our data did not allow us to match students to their specific classrooms and because the intervention is considered a school-based program For a discussion of the consequences of omitting a level of clustering in multilevel models, see Moerbeek (2004) and Van Landeghem, De Fraine, and Van Damme (2005).

$$Y_{ij} = \alpha_j + \sum_{q=1}^{Q} \beta_q X_{qij} + \varepsilon_{ij} \qquad \text{Level 1 (individual level)} \qquad (2)$$

$$\alpha_j = \gamma_0 + \gamma_1 (STATUS)_j + \sum_{s=2}^{S} \gamma_s W_{sj} + u_j \qquad \text{Level 2 (school level)} \qquad (3)$$

Substituting equation 2 into equation 1 allows the system of equations to be rewritten in reduced-form:

$$Y_{ij} = \gamma_0 + \gamma_1 (STATUS)_j + \sum_{q=1}^{Q} \beta_q X_{qij} + \sum_{s=2}^{S} \gamma_s W_{sj} + u_j + \varepsilon_{ij} \qquad (4)$$

In this model, Y_{ij} is our reading comprehension measure for student *i* in school *j*; X_{qij} is the *q*th individual-level covariate for observed baseline characteristics, for $q = 1\ldots$ Q; *STATUS* is a dummy variable indicating whether school *j* received a randomized offer to participate in Pacific CHILD (*STATUS* = 1) or not (*STATUS* = 0); and W_{sj} is the *s*th school-level covariate, for $s = 2\ldots S$.

In addition, γ_0 represents the adjusted mean outcome across control group schools (when *STATUS* = 0). Importantly, γ_1 is the impact estimator of interest, representing the regression-adjusted mean difference in reading comprehension scores between treatment and control group schools. β_q and γ_s are estimators for marginal effects of individual- and school-level covariates. Note that we constrained the effects of the individual-level covariates, β_q for $q = 1\ldots$ Q, as fixed across the school level (Level 2).

Finally, our model includes two error terms: ε_{ij} is the residual term specific to student *i* in school *j*; u_j is the residual specific to the *j*th school. We assume that ε_{ij} and u_j are independently and normally distributed, each with mean 0 and constant variance (σ_ε^2 and σ_u^2), such that $\varepsilon_{ij} \mid u_j \sim N(0, \sigma_\varepsilon^2)$ and $u_j \sim N(0, \sigma_u^2)$. In this random intercept HLM model, the effects of clustering of students within schools were explicitly specified, in the form of between-school heterogeneity represented by school-specific random intercepts (u_j).

In our study, we fit an expanded version of model (4) that incorporated site fixed effects in level 2 to account for the site-to-site variation in the outcome.[9] The reduced form model with the entity-specific intercepts can be expressed as follows:[10]

$$Y_{ij} = \lambda_{HI} HI + \lambda_{AS} AS + \lambda_{NM} CNMI + \gamma_1 (STATUS)_j + \sum_{q=1}^{Q} \beta_q X_{qij} + \sum_{s=2}^{S} \gamma_s W_{sj} + u_j + \varepsilon_{ij} \qquad (5)$$

where HI, AS and CNMI are dummy variables for Hawai'i, American Samoa and Commonwealth of the Northern Mariana Islands respectively; and λ_{HI}, λ_{AS}, and λ_{NM} are the parameters representing the fixed site-specific effects for Hawai'i, American Samoa, and CNMI respectively. We used the restricted maximum likelihood estimation (REML) method to estimate the coefficients and covariance parameters. [11]

Alternative estimation methods

To verify the results from our main estimation of the HLM model, we re-analyzed our data using alternate estimation methods to handle clustered data as discussed in Section 2.2. First, we re-fit our reduced-form model using the maximum likelihood estimation (MLE) method, instead of REML: MLE is another commonly used estimation method but provides downward-biased

[9] We treated site as a fixed effect rather than a random effect given that our study sites, which we purposively selected, did not necessarily represent a sample from a larger population of sites. Also, since we purposively selected these sites, we did not intend to draw inferences that were generalizable beyond these specific sites.

[10] We also conducted likelihood ratio tests based on statistics we obtained using MLE estimation to test the interaction of treatment condition and site effects (i.e., site-specific treatment effects), assuming these effects were fixed. These specification tests rejected the site-specific treatment effect assumptions, and we concluded that the fixed site-specific slopes (i.e., the interaction terms between the entity indicators and the treatment indicator) did not contribute to the estimation of the student outcome once we included fixed site-specific intercepts.

[11] We used the *xtmixed* command in Stata to estimate these models, specifying either the *reml* option to fit the REML models or the *mle* option to fit the MLE models. In addition to Stata, there are other software packages that can fit hierarchical linear models including SAS, SPSS, HLM and R. For a comprehensive overview of how to use these different software packages to fit hierarchical linear models and to understand how results compare across software packages, readers should consult West, Welch, and Galecki (2007).

estimates for σ_u^2.[12] Second, we used Feasible Generalized Least Squares (FGLS) based on the Swamy and Arora ANOVA method which provides an unbiased but inefficient estimator for σ_u^2. We then estimated our models using two methods that account for the interdependence between students clustered within schools, but did not explicitly estimate the within-school covariance structure: Generalized Estimating Equations (GEE) [13] and Ordinary Least Squares (OLS) with robust cluster (Huber-White) standard errors (SE)[14]. As the primary goal of our study was to estimate the coefficient (γ_1) on the treatment indicator variable (*STATUS*), not necessarily the random effects variance component, both GEE and OLS with Huber-White SE methods were regarded as reasonable alternative approaches for checking the robustness of our main impact estimation results.[15, 16]

4.2 *Weighted Average Approach*

One additional set of analyses involved estimating our models separately for each of the three sites and then taking a *weighted average* of separately estimated site-specific effects. More specially, weights (*w*) were defined as follows:

$$w_k = \frac{1}{se_i^2} \tag{4}$$

[12] Unlike the MLE, REML takes into account the loss of degrees of freedom that results from estimating the fixed-effects parameters in estimating the variance components (and provides an unbiased estimate for the variance components for balanced data). Consequently, the restricted maximum likelihood method yields more conservative (larger) estimates for standard errors for coefficient estimates than the maximum likelihood method.

[13] Generalized estimating equation parameters are estimated by an iterative optimization process, with the working covariance as a function of the working correlation matrix (of the dependent variable). The form of this working correlation matrix was assumed to be exchangeable. The covariance parameters are treated as nuisance variables in the iterative process. Estimates for the covariance based on a generalized estimating equation model are consistent, assuming the correlation matrix is correctly specified (Hanley, Negassa, & Forrester, 2003).

[14] As noted earlier, we did not use OLS regression without robust clustered standard errors given the correlation of students within schools, a violation of the independence assumption of OLS regression.

[15] See Schochet (2009) for a discussion of various estimation methods used in analyzing data from clustered randomized controlled trials. Also, see Cameron and Miller (2010) for a technical discussion of methods to handle clustered data, including FGLS and robust-clustered standard errors.

[16] We estimated all models in Stata using the following commands*:* *xtreg* with the *sa* option for the variance estimator (for FGLS); *xtgee* specifying the normal distribution for the dependent variable (for GEE) ; and *reg* specifying the cluster-robust option (OLS with robust cluster SEs).

where se_i is the standard error of the student effect estimate γ_{1k} in entity k. The weighted-average estimate, γ_{1k}, and its variance, $v(\gamma_1)$, were calculated as:

$$\gamma_1 = \frac{\sum_k w_k \gamma_{1k}}{\sum_k w_k} \tag{5}$$

and

$$v(\gamma_1) = \frac{1}{\sum_k w_k} \tag{6}$$

This weighted-average approach gives more weight to more precise estimates and less weight to less precise estimates.[17] The overall effect estimate for the student outcome thus reflected the effectiveness of the program measured across the three entities; the averaged effect took into account the variation in the impact estimate across entities.[18]

5. Results

In this section, we first present our ICC estimates, showing the extent to which students are correlated within schools in our sample. Then we present and discuss our results across our main and alternative estimation methods.

5.1 *Intra-cluster Correlation Coefficients*

Given potential heterogeneity across sites, we calculated unconditional ICCs based on an HLM model (estimated via REML) without any predictors (i.e., a null model (Hox, 2010)) for each of the three entities in our study. The ICCs were .04 for AS, .07 for HI and .06 in the CNMI. Thus, approximately 4% to 7% of the variation in student reading comprehension was due to between school differences; the rest was due to variation between students within schools. Though certainly

[17] This approach is frequently used in meta-analysis to compute weights for combining effects across independent samples (Cooper, Hedges, & Valentine, 2009). Our calculations of the weights and standard errors were based on fixed effects models, assuming a common effect across entities. We report the results based on the fixed effects models because we did not plan to generalize the results beyond the three entities.

[18] We fitted our models by site and calculated the weighted average using Stata.

not large (recall that ICCs closer to 1 indicate a higher degree of correlation within clusters), these ICCs do suggest that students within schools exhibit some correlation with each other and thus warranted the use of data analytic methods to account for clustering.

5.2 *Estimation Results*

In Table 3, we display our results across each of our selected estimation methods, beginning with HLM models in the rows (1) and (2) (using REML and MLE, respectively). Rows (3)-(5) summarize results for models based on our alternative methods: FGLS, GEE and OLS with cluster-robust Huber-White SEs, respectively. We report our results both in terms of the estimated difference in scale score points as well as effect sizes.[19] We first address our initial HLM findings using both REML and MLE.

As shown in row 1 in Table 3, our initial analysis of the impact of the intervention on our student outcome (as captured by the estimate of the parameter on *STATUS*, γ_1), based on our 2-level HLM using REML across all three entities, showed a 2.35 point difference between control and treatment group students (effect size=0.064). However, this estimate was not statistically significant at the standard .05 level (*p*=.258). On the other hand, the results of the HLM model using MLE, did yield a positive and statistically significant estimate (*p*=.037, effect size=0.083).

As shown in row (3) the coefficient estimate on *STATUS* based on FGLS was consistent in terms of its magnitude and statistical significance with the estimate from the model we fit using HLM-REML (2.21 versus 2.34 points; *p*=.33 versus *p*=.258; effect size=0.060 versus 0.064). However, this estimate was inconsistent with the HLM-MLE results, particularly in terms of the estimator's statistical significance (*p*=.33 versus *p*=.037). On the other hand, our results using GEE and OLS with Huber-White SEs were consistent with the HLM-MLE results, but conflicted with the findings under FGLS and HLM-REML (*p*=.015, effect size=0.088).

[19] For Table 3, we computed effect sizes by dividing the estimated difference in the treatment and control group means in the student outcome by the standard deviation of the student outcome for the control group.

These inconsistencies prompted us to further examine the data and potential factors contributing to these results. One additional set of analyses involved estimating our models separately for each of the three sites. In doing so, we discovered that the impact estimates varied considerably by site as shown in Table 4. Due to this variation across sites, we then pursued an alternative approach in consultation with external reviewers: estimating the program impact as a *weighted average* of separately estimated site-specific effects as described in Section 4.3.[20]

In Table 5, we report the estimated impacts based on the weighted average approach in terms of both scale score points and effect sizes.[21] Overall, using this weighted-average approach, we refit our HLM-REML model and found a statistically significant difference between the treatment and control groups (effect size = 0.244, p = .017). As shown in Table 5, the weighted average approach also yielded consistent results across different methodological approaches, both in terms of effect size (ranging from 0.208 under FGLS to 0.244 under HLM-REML) and statistical significance (p<.001 under HLM-MLE, GEE and OLS with Huber-White SE; p=.017 for HLM-REML; and p=.002 under FGLS). While achieving consistent results across methods was not necessarily a criterion for selecting this approach, using a weighted average approach did support the robustness of our findings. In the end, we were able to confidently conclude that there was a statistically significant impact of the intervention on student reading comprehension, which countered our initial results we obtained by fitting our models using an unweighted, pooled sample under HLM-REML.

However, modifying our original analytic approach (HLM-REML using a pooled sample) concerned us because it diverged from our original analysis plan and could have been seen as an

[20] Although the likelihood ratio test indicated that fixed entity-specific slopes did not make additional contribution to the estimation of the student outcome measure once entity-specific intercepts were included in the model, the fixed entity-specific slope estimate for American Samoa was statistically significant. The fixed entity-specific slopes for the two other entities were not statistically significant.

[21] We computed the overall effect size as the weighted mean of the entity-specific effect sizes. For each entity, we computed the effect size by dividing the regression-adjusted mean difference in reading comprehension scores between the control and treatment groups by the standard deviation of the control group mean. We then computed the overall effect size from the mean of the entity effect sizes, applying the same weights used to compute the overall impact estimates in the scale score unit.

attempt to manipulate the estimation results. To counter such suspicion, we decided to carefully document and report our analytic decision process which helped to maintain the face validity of the study, despite the divergence from our original data analytic plan.

6. Lessons Learned and Conclusion

Our work highlights the importance of verifying the results of cluster RCTs particularly when data are clustered and are subject to heterogeneity across sites. Though well-known data analytic methods are commonly used when analyzing data from cluster RCTs in education (e.g., HLM); however, more often than not, there is not a single "correct" estimation method, and analytic decisions depend primarily on the judgment of researchers. It is plausible that these decisions are made based solely upon the preferences of the researchers. In this case, thorough sensitivity analyses are particularly critical in order to verify the results.

In addition to the main lesson we learned about the value of verifying our results across alternative methods for handling clustered data, we learned several additional lessons:

1. *Methodological Bridging.* Often times conducting sensitivity analyses that incorporate different methodological approaches requires what we call *methodological bridging*. Particular methods are discipline-specific, so it is important to look broadly at other disciplinary areas to understand how they handle similar methodological issues. Not only can methods to handle similar issues—such as clustering—differ, but the methodological terminology can vary as well, so it is important to build bridges with analysts who are trained in evaluation but are grounded in different disciplines ranging from economics, statistics and, more broadly, the social sciences (e.g., public health, education and public policy).
2. *Selection of Analytic Methods for Clustering.* Based on our review of studies that compare ways to handle clustered data as well as our own empirical findings, we conclude that there is no one "right" way to handle clustering. Beyond our basic recommendation that analysts should

account for clustering when analyzing data from clustered RCTs, we also advise analysts to carefully consider the tradeoffs in analyzing clustered data. There are important practical considerations (e.g., the number of levels in the data) and distributional assumptions of the data.

3. *Weighted Average Approaches.* If there are multiple sites in a study, analysts may want to consider estimating program effects separately for each site as an exploratory step. In our case, this was helpful because it revealed cross-site heterogeneity. If there are differences in effects across sites, as we had discovered in our study, one option analysts may want to consider is using a weighted average approach as we have described above.
4. *Transparent Reporting of Methods and Results.* Finally, and most importantly, we highly recommend that analysts clearly develop *a priori* a plan for analyzing clustered data as part of their study protocols, including a description of the alternative approaches they will undertake. Also, analysts should clearly document and report their analytic methods and findings across methods. Doing so ensures that analysts will carry out their analytic work both thoughtfully and responsibly, preserving the overall integrity and rigor of the study. As we mentioned, in our study, we were concerned that deviating from our a priori specified data analytic plan would impact the face validity of the study; as such, we clearly documented the procedures and results of our analytic decisions in order to be fully transparent about how we obtained our final impact findings.

Our paper's contribution to the extant evaluation literature is to raise the awareness of the different methodological approaches to handle clustered data, the need to verify results across methods and—importantly—documenting and supplying information on the data analytic decision process if results of those sensitivity analyses are inconsistent. In sum, we believe that analysts should strive to become much more transparent and rigorous in their use, discussion and reporting of sensitivity analyses for clustered data arising from cluster RCTs. Doing so can greatly enhance the credibility and robustness of findings from impact evaluations that rely on cluster RCTs.

References

Arceneaux, K., & Nickerson, D. W. (2009). Modeling certainty with clustered data: A comparison of methods. *Political Analysis, 17*(2), 177-190.

Boruch, R. F. (2005). *Place randomized trials: Experimental tests of public policy*. Thousand Oaks: SAGE Publications.

Burton, P., Gurrin, L., & Sly, P. (1998). Tutorial in biostatistics. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modeling. *Stat Med, 17*, 1261-1291.

Cameron, A. C., & Miller, D. L. (2010). Robust inference with clustered data. *Handbook of empirical economics and finance*, 1-28.

Cameron, A. C., & Miller, D. L. (2013). A Practitioner's Guide to Cluster-Robust Inference: unpublished.

Clarke, P. (2008). When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. *Journal of Epidemiology and Community Health, 62*(8), 752-758.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis*. New York, NY: Russell Sage Foundation.

Galbraith, S., Daniel, J. A., & Vissel, B. (2010). A study of clustered data and approaches to its analysis. *The journal of Neuroscience, 30*(32), 10601-10608.

Garson, G. D. (2012). *Hierarchical linear modeling: Guide and applications*. Thousand Oaks: Sage.

Gelman, A. (2006). Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics, 48*(3).

Gertler, P. (2004). Do conditional cash transfers improve child health? Evidence from PROGRESA's control randomized experiment. *The American Economic Review, 94*(2), 336-341.

Hanley, J. A., Negassa, A., & Forrester, J. E. (2003). Statistical analysis of correlated data using generalized estimating equations: An orientation. *American Journal of Epidemiology, 157*(4), 364-375.

Hox, J. (2010). *Multilevel analysis: Techniques and applications*. New York: Routledge.

Hubbard, A. E., Ahern, J., Fleischer, N. L., Van der Laan, M., Lippman, S. A., Jewell, N., . . . Satariano, W. A. (2010). To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology, 21*(4), 467-474.

Killip, S., Mahfoud, Z., & Pearce, K. (2004). What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. *The Annals of Family Medicine, 2*(3), 204-208.

Kolen, M., & Brennan, R. (2004). *Test equating, linking, and scaling: Methods and practices*. New York: Springer.

Liang, K.-Y., & Zeger, S. L. (1993). Regression analysis for correlated data. *Annual review of public health, 14*(1), 43-68.

Lohr, S. (2010). *Sampling: design and analysis* (Second ed.). Boston, MA: Brooks/Cole, Cengage Learning.

Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research, 39*(1), 129-149.

Parker, S. W., & Teruel, G. M. (2005). Randomization and social program evaluation: The case of Progresa. *The Annals of the American Academy of Political and Social Science, 599*(1), 199-219.

Peters, T., Richards, S., Bankhead, C., Ades, A., & Sterne, J. (2003). Comparison of methods for analysing cluster randomized trials: an example involving a factorial design. *International journal of epidemiology, 32*(5), 840-846.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*(2), 173.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks: Sage Publications.

Schochet, P. Z. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics, 33*(1), 62-87.

Schochet, P. Z. (2009). Technical methods report: The estimation of average treatment effects for clustered RCTs of education interventions (NCEE 2009–0061rev). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education.

Steenbergen, M. R., & Jones, B. S. (2002). Modeling multilevel data structures. *american Journal of political Science*, 218-237.

Thabane, L., Mbuagbaw, L., Zhang, S., Samaan, Z., Marcucci, M., Ye, C., . . . Kosa, D. (2013). A tutorial on sensitivity analyses in clinical trials: The what, why, when and how. *BMC Medical Research Methodology, 13*(1), 92.

Van Landeghem, G., De Fraine, B., & Van Damme, J. (2005). The consequence of ignoring a level of nesting in multilevel analysis: A comment. *Multivariate Behavioral Research, 40*(4), 423-434.

West, B., Welch, K. B., & Galecki, A. T. (2007). *Linear mixed models: a practical guide using statistical software*. Boca Raton, FL: Chapman & Hall/CRC Press.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, 817-838.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, 1-25.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*: MIT press.

Zorn, C. (2006). Comparing GEE and robust standard errors for conditionally dependent data. *Political Research Quarterly, 59*(3), 329-341.

Zyzanski, S. J., Flocke, S. A., & Dickinson, L. M. (2004). On the nature and analysis of clustered data. *The Annals of Family Medicine, 2*(3), 199-200.

Tables

Table 1. Covariates used in an analysis of a clustered randomized controlled trial (RCT) of a teacher professional development program. ($n_{students}$=3,052; $n_{schools}$=45).

Level/type of covariate	**Variables**
School (Level 2)	
School performance	▪ Baseline year average scores on reading comprehension subtest of Stanford 10 Achievement Test (SAT 10)[a]
School characteristics	▪ School size (number of students in school) at baseline[b] ▪ Student-to-teacher ratio at baseline ▪ Percent free or reduced-price lunch-certified students at baseline
Student (Level 1)	
Demographics	▪ Gender (binary indicator for female) ▪ Special education status (binary indicator)

a. The baseline year school average was computed based on grade 5 scores for schools in Hawai'i and the CNMI and grade 4 scores for schools in American Samoa (where grade 5 scores were not available). For Hawai'i, TerraNova scores were used to estimate equivalent SAT 10 scores, using equipercentile methods.

b. The number of grades varied by school. The study examined the number of fourth and fifth graders per school, which was collected during the recruitment period, as a secondary source of the baseline school size. Estimations based on the fourth and fifth grade size yielded the same results as those based on the total school size.

Table 2. Selected univariate descriptive statistics (mean and standard deviations) of baseline characteristics for schools participating in a clustered randomized controlled trial (RCT) of a teacher professional development program. ($n_{students}$=3,052; $n_{schools}$=45).

	Mean (standard deviation)			**Standardized mean difference**[e]	**Test of difference**
Baseline characteristic	**Overall**	**Treatment schools**	**Control schools**		***p*-value**
Number of grade 4 students	65.0 (40.0)	64.7 (42.5)	65.3 (38.2)	–0.01	.962
Number of grade 5 students	64.1 (41.4)	64.2 (42.5)	64.0 (38.2)	0.00	.983
Number of grade 4 and grade 5 teachers	6.0 (3.1)	6.0 (3.4)	6.0 (2.9)	0.00	.962
Student-teacher ratio in grades 4 and 5	20.2 (4.8)	20.1 (4.9)	20.3 (4.9)	-0.04	.935
Proportion of students eligible for free or reduced-price meals, all grades	69.0 (28.7)	69.8 (28.5)	68.1 (29.6)	0.06	.842
Mean proportion of students of races/ethnicities other than White, all grades	86.9 (15.2)	87.7 (13.9)	86.1 (16.8)	0.11	.722
Mean proportion of English language learner students (Hawai'i only), all grades[a]	12.8 (9.7)	13.5 (8.2)	12.3 (11.4)	0.12	.764
Mean reading comprehension score (SAT 10 scale score),[b, c] grade 4	609.6 (20.5)	608.9 (21.1)	610.4 (20.6)	-0.07	.829
Mean reading comprehension score (SAT 10 scale score),[b, d] grade 5	636.5 (11.0)	634.4 (11.9)	638.6 (9.9)	-0.38	.260
Number of schools	45	23	22		
Number of schools in Hawai'i	26	13	13		

Note: Significance tests are based on two-tailed *t*-tests, accounting for clustering at the school level.

a. Data on English language learner status were available only for Hawai'i students.

b. TerraNova reading comprehension scores from Hawai'i were converted to estimated Stanford 10 reading comprehension equivalents using published norming tables and concordancing method.

c. At baseline, grade 4 students in American Samoa and Hawai'i completed standardized assessments. Thirty-five schools (18 treatment group schools and 17 control group schools) had grade 4 scores.

d. At baseline, grade 5 students in the CNMI and Hawai'i completed standardized assessment. Thirty six schools (18 treatment group schools and 18 control group schools) had grade 5 scores.

e. Calculated as the treatment mean minus the control mean divided by the pooled standard deviation.

Source: Authors' analysis based on data from U.S. Department of Education 2010a for American Samoa and the Commonwealth of the Northern Mariana Islands, U.S. Department of Education 2010b for Hawai'i. Figures for students and teachers per school in grades 4 and 5 are based on enrollment estimates from the American Samoa Department of Education, the Commonwealth of the Northern Mariana Islands Public School System, and the Hawai'i Department of Education.

Table 3. Estimated parameters, standard errors and *p*-values from models fitted using different methods to account for clustered data (students nested in schools). ($n_{students}$=3,052; $n_{schools}$=45).

Model	**Parameter estimate (standard error)**	***p*-value**	**95% confidence interval**	***Effect Size***
Hierarchical linear models (HLM)				
HLM using restricted maximum likelihood (REML)	**2.348** **(2.076)**	**0.258**	-1.72 - 6.42	**0.064**
HLM using maximum likelihood (MLE)	3.059* (1.465)	0.037	0.19 - 5.93	0.083
Alternative estimation methods				
Feasible generalized least squares (FGLS) with Swamy-Arora method	2.218 (2.277)	0.330	-2.25 - 6.68	0.060
Generalized estimating equations (GEE) with model-based standard errors	3.099* (1.445)	0.032	0.27 - 5.93	0.084
Ordinary least squares (OLS) with cluster-robust Huber-White standard errors	3.232* (1.275)	0.015	0.66 -5.80	0.088

*Significant at the .05 level (two-tailed test), **significant at the .01 level (two-tailed test).

Note: The bolded results (REML estimates for the combined sample) represent the initial benchmark results. The models in this table included the following covariates: blocking variables, school-level baseline reading comprehension scale score, school size, student-to-teacher ratio, percentage of students eligible for free or reduced-price lunch, student gender, student special education status, and student race (being white). The covariates included in the models in this table slightly differ from those included in the models in tables 4 and 5 (see footnotes of tables 4 and 5). The impact estimates based on the combined sample using the same covariates as tables 4 and 5 yield the equivalent results with very similar parameter estimates and standard errors (the results are available from authors).

Table 4. Estimated parameters and standard errors from models fitted using different methods to account for clustered data (students nested in schools) by study site (n_{students}=3,052; n_{schools}=45).

	Site					
	American Samoa (AS) **(n_{students}=185)[a]**		**The Commonwealth of the North Mariana Islands (CNMI)** **(n_{students}=692)[a]**		**Hawai'i** **(n_{students}=2,175;n_{schools}=26)**	
Model	**Parameter estimate (standard error)**	***p*-value**	**Parameter estimate (standard error)**	***p*-value**	**Parameter estimate (standard error)**	***p*-value**
Hierarchical linear models (HLM)						
HLM using restricted maximum likelihood (REML)	-3.019 (6.241)	0.629	11.529* (5.227)	0.027	5.134 (2.617)	0.050
HLM using maximum likelihood (MLE)	-2.712 (4.359)	0.534	10.970** (3.003)	<0.001	5.711* (1.931)	0.003
Alternative estimation methods						
Feasible generalized least squares (FGLS) with Swamy-Arora method	-2.712 (4.469)	0.544	10.970** (3.025)	<0.001	5.010 (2.987)	0.093
Generalized estimating equations (GEE) with model-based standard errors	-6.709 (4.834)	0.165	10.172** (3.346)	0.002	6.489** (1.593)	<0.001
Ordinary least squares (OLS) with cluster-robust Huber-White standard errors	-2.712 (2.953)	0.358	10.970** (2.905)	<0.001	5.711** (1.339)	<0.001

Notes: For all models (except for the model using GEE), covariates include: blocking variables, school-level baseline reading comprehension scale score, school size, student-to-teacher ratio, percentage of students eligible for free and reduced-price lunch, student gender, and student special education status. For the GEE model only, we excluded student-to-teacher ratio and percent eligible for free and reduced price lunch.

*Significant at the .05 level (two-tailed test), **significant at the .01 level (two-tailed test).

[a] There are a total of 19 schools within American Samoa and the CNMI. Following Institute of Education Sciences (IES) guidelines, the number of schools are combined to prevent disclosure risk.

Table 5. Weight-average regression adjusted results from models fitted using different methods to account for clustered data (students nested in schools). ($n_{students}$=3,052; $n_{schools}$=45).

	Weighted-average regression-adjusted means of the SAT10 scale scores						
Model	**Treatment schools**	**Control schools**	**Difference**	**Standard error**	***p*-value**	**95% confidence interval**	**Weighted average effect size**
Benchmark models							
Hierarchical linear model (HLM); Restricted maximum likelihood (REML)	634.3	629.0	5.3*	2.19	.017	0.96 - 9.55	0.244
Hierarchical linear model (HLM); Maximum likelihood (MLE)	634.1	628.1	6.0**	1.52	<.001	3.05 - 9.02	0.234
Alternative estimation methods							
Feasible generalized least squares (FGLS) with Swamy-Arora method	630.5	624.5	6.0**	1.92	.002	2.22 - 9.75	0.208
Generalized estimating equations (GEE) with model-based standard error	636.5	630.4	6.0**	1.38	<.001	3.34 - 8.74	0.216
Ordinary least squares (OLS) with cluster-robust Huber-White standard errors	633.6	628.3	5.3**	1.12	<.001	3.07 - 7.48	0.224

*Significant at the .05 level (two-tailed test), **significant at the .01 level (two-tailed test).

Note: The number of observations = 3 entities (3,052 students for the three entities combined). Scores are based on reading comprehension assessment data from the Stanford 10 Achievement Test (SAT 10) for American Samoa and the CNMI and the TerraNova for Hawai'i. TerraNova scores were converted to SAT 10–equivalent scores using equipercentile methods (Kolen & Brennan, 2004). For each entity, regression-adjusted means were computed at the means of the covariates; effect sizes were calculated by dividing the impact estimate by the standard deviation of the control group. Unless otherwise noted, models fit under each estimation method included the following covariates: blocking variables, school-level baseline reading comprehension scale score, school size, student-to-teacher ratio, percentage of students eligible for free or reduced-price lunch, student gender, and student special education status. For generalized estimating equations, the student-to-teacher ratio and the percentage of students eligible for free or reduced-price lunch were excluded, because the model with the full set of covariates failed to converge. The overall impacts in scale score and effect size were computed as weighted means of the three single-entity impacts and the three corresponding effect sizes, with weights defined as the inverse of the variance of each scale score impact estimates.