Research Article

Peter Z. Schochet*

Design-based RCT estimators and central limit theorems for baseline subgroup and related analyses

https://doi.org/10.1515/jci-2023-0056
received August 25, 2023; accepted May 13, 2024

Abstract: There is a growing literature on design-based (DB) methods to estimate average treatment effects (ATEs) for randomized controlled trials (RCTs) for full sample analyses. This article extends these methods to estimate ATEs for discrete subgroups defined by pre-treatment variables, with an application to an RCT testing subgroup effects for a school voucher experiment in New York City. We consider ratio estimators for subgroup effects using regression methods, allowing for model covariates to improve precision, and prove a new finite population central limit theorem. We discuss extensions to blocked and clustered RCT designs, and to other common estimators with random treatment-control sample sizes or summed weights: post-stratification estimators, weighted estimators that adjust for data nonresponse, and estimators for Bernoulli trials. We also develop simple variance estimators that share features with robust estimators. Simulations show that the DB subgroup estimators yield confidence interval coverage near nominal levels, even for small subgroups.

Keywords: randomized controlled trials, subgroup analyses, design-based estimators, finite population central limit theorems

MSC 2020: 62K99, 62D99, 62E20, 62G20

1 Introduction

There is a growing literature on design-based (DB) methods to estimate overall average treatment effects (ATEs) for randomized controlled trials (RCTs). These nonparametric methods use the building blocks of experimental designs to generate consistent, asymptotically normal ATE estimators with minimal assumptions. The underpinnings of these methods were introduced by Neyman [1] and later developed in seminal works by Rubin [2,3] and Holland [4] using a potential outcomes framework.

To date, the DB literature has focused on ATE estimation for full sample analyses. In this article, we build on these methods to develop ATE estimators for discrete *subgroups* defined by pre-treatment (baseline) characteristics of study participants. Subgroup analyses for RCTs are common across fields as they can be used to assess treatment effect heterogeneity and inform decisions about how to best target and improve treatments [5,6]. Guidelines for the planning, analysis, and reporting of RCT subgroup analyses have been proposed in the literature to ensure statistical rigor, such as approaches to reduce the chances of finding spurious positive effects due to multiple testing [5,7,8].

As a motivating example, consider the evaluation of the New York City (NYC) School Choice Scholarships Program, an RCT where low-income public school students in grades K–4 could participate in a series of lotteries to receive a private school voucher for up to 3 years [9,10]. A subgroup analysis was pre-specified for

*** Corresponding author: Peter Z. Schochet,** Mathematica, P.O. Box 2393, Princeton, NJ 08543-2393, United States of America, e-mail: pschochet@mathematica-mpr.com, tel: +(609) 936-2783

the study to examine differences in voucher effects for African-American and Latino students. The hypothesis was that African Americans might benefit more from the vouchers as they tended to live in poorer communities and attend lower-performing public schools.

Several key aspects of this subgroup analysis motivate the theory underlying this article. First, the study sample was not randomly sampled from a broader population. Rather, the sample included only a very small percentage of NYC families who applied for a scholarship. Thus, the study results cannot be generalized to a broader voucher program that would involve all children in NYC or elsewhere. This setting suggests a finite population framework for estimating ATEs where the sample and their potential outcomes are considered fixed, and study results are assumed to pertain to the study sample only. This is a common RCT setting across disciplines that often include volunteer samples of individuals and sites.

Second, the estimation strategy should allow for the inclusion of model baseline covariates to improve precision as power is often a concern for subgroup analyses due to small sample sizes. Third, the voucher study conducted randomization within strata, suggesting the need for a theory for blocked RCTs. Fourth, the study randomized families rather than students, suggesting a further need to consider a theory for clustered RCTs that are becoming increasingly prevalent across fields [11,12]. Finally, the study constructed weights to adjust for missing outcome data, a common strategy for RCT analyses that should be covered in the theory.

This article addresses these issues by developing DB ATE ratio estimators for subgroup-related analyses using regression models that allow for baseline covariates. We focus on ratio estimators due to the randomness of subgroup sizes in the treatment and control groups. We prove a new finite population central limit theorem (CLT) by building on the methods reported by Pashley [13] and Schochet et al. [14]. We also discuss extensions to blocked and clustered RCTs, and to other common estimators with random sample sizes or summed weights: post-stratification estimators, weighted estimators that adjust for data nonresponse, and estimators for Bernoulli trials (BTs). We provide consistent variance estimators that are compared to commonly used robust standard errors (SEs). Our simulations show that the DB subgroup ATE estimators yield confidence interval coverage near nominal levels, even for small subgroups. Finally, we demonstrate the methods using data from our motivating NYC voucher experiment.

The rest of this article proceeds as follows. Section 2 discusses the related literature. Section 3 provides the theoretical framework, ATE estimators and CLT results for the non-clustered RCT, and extensions. Section 4 discusses blocked and clustered RCTs. Section 5 presents simulation results, and Section 6 presents empirical results using the NYC voucher study. Section 7 concludes.

2 Related work

Our work builds on the growing literature on DB methods to estimate ATEs for full sample analyses [14–23]. These methods also pertain to subgroup analyses conditional on subgroup sizes observed in the treatment and control groups [21], but not to unconditional analyses that average over subgroup allocations.

Our work draws most directly on two studies. First, we draw on methods in Schochet et al. [14] who provide finite population CLTs for ratio estimators for blocked, clustered RCTs with general weights (using previous results in Scott and Wu [24], Li and Ding [23], and Pashley [13]). Our innovation is to adapt these methods by treating subgroup indicators as “weights” in the analysis. Second, we draw on results from the study by Miratrix et al. [25] who considered DB post-stratification estimators for overall effects, which share properties with baseline subgroup estimators. Miratrix et al. [25], however, do not consider asymptotic distributions, blocked or clustered RCT designs, the inclusion of other model covariates, or weights considered here.

Finally, there is a large statistical literature on DB methods for analyzing survey data with complex sample designs, including for estimating subpopulation means or totals [26–28]. However, these works do not consider RCT settings for estimating treatment-control differences in subpopulation means.

In what follows, we focus on the non-clustered RCT design without blocking and extensions to related estimators. We then discuss blocked and clustered designs.

3 DB subgroup analysis for non-clustered RCTs

We assume an RCT of n individuals, with $n^1 = np$ assigned to the treatment group and $n^0 = n(1-p)$ assigned to the control group, where p is the treatment assignment rate ($0 < p < 1$). Let T_i equal 1 if person i is randomly assigned to the treatment condition and 0 otherwise.

Let $Y_i(1)$ be the outcome of person i if assigned to the treatment group and $Y_i(0)$ be the outcome in the control condition. These potential outcomes can be continuous, binary, or discrete. We assume a finite population model where potential outcomes are fixed for the study.

For the subgroup analysis, we assume each sample member is allocated to a discrete category within a subgroup class with $K \geq 1$ levels. The subgroup classes (such as age or race/ethnicity groups) can be formed from continuous, categorical, or discrete variables measured at baseline, so are unaffected by the treatment. We consider estimation for each subgroup class in isolation. For a specific class, let G_{ik} equal 1 for a member of subgroup (level) k and 0 otherwise, for $k \in \{1, 2, \ldots, K\}$. Let $n_k = \sum_{i=1}^{n} G_{ik}$ denote the number of persons in subgroup k, with $\sum_{k=1}^{K} n_k = n$. Finally, let $\pi_k = \bar{G}_k = n_k/n$ be the subgroup population share, with $\sum_{k=1}^{K} \pi_k = 1$.

We assume two conditions. The first is the stable unit treatment value assumption (SUTVA) [29]:

(C1): SUTVA: Let $Y_i(\mathbf{T})$ denote the potential outcome given the random vector of all treatment assignments, $\mathbf{T}$. Then, if $T_i = T_i'$ for person i, we have that $Y_i(\mathbf{T}) = Y_i(\mathbf{T}')$.

SUTVA allows us to express $Y_i(\mathbf{T})$ as $Y_i(T_i)$, so that a person's potential outcomes depend only on the person's treatment assignment and not on those of other persons in the sample. This condition is assumed to hold within and across subgroups. SUTVA also assumes that a particular treatment unit cannot receive different forms of the treatment.

Under SUTVA, the ATE parameter for subgroup k under the finite population model is,

$$\tau_k = \frac{\sum_{i=1}^{n} G_{ik}(Y_i(1) - Y_i(0))}{n_k} = \bar{Y}_k(1) - \bar{Y}_k(0), \tag{1}$$

which is the mean treatment effect for members of subgroup k in the study sample.

Our second condition is complete randomization [21], where extensions to BTs are discussed in Section 3.3:

(C2): Complete randomization: For fixed n^1, if $\mathbf{t} = (t_1, \ldots, t_n)$ is any vector of randomization realizations such that $\sum_{i=1}^{n} t_i = n^1$, then $Prob(\mathbf{T} = \mathbf{t}) = \binom{n}{n^1}^{-1}$.

This condition implies that potential outcomes are independent of treatment status, $Y_i(1), Y_i(0) \perp\!\!\!\perp T_i$, which also holds for any baseline subgroup (e.g., males or females).

3.1 ATE estimators

Under the potential outcomes framework and SUTVA, the data generating process for the observed outcome measure, y_i, is a result of the random assignment process:

$$y_i = T_i Y_i(1) + (1 - T_i) Y_i(0). \tag{2}$$

This relation states that we can observe $y_i = Y_i(1)$ for those in the treatment group and $y_i = Y_i(0)$ for those in the control group, but not both.

Rearranging (2) generates the following nominal full sample regression model:

$$y_i = \alpha + \tau \tilde{T}_i + u_i, \tag{3}$$

where $\tilde{T}_i = (T_i - p)$ is the centered treatment indicator; $\tau = \bar{Y}(1) - \bar{Y}(0)$ is the full sample ATE estimand; $\bar{Y}(t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)$ is the mean potential outcome for $t \in \{1,0\}$; $\alpha = p\bar{Y}(1) + (1-p)\bar{Y}(0)$ is the intercept (expected outcome); and the "error" term, u_i, is

$$u_i = T_i(Y_i(1) - \bar{Y}(1)) + (1 - T_i)(Y_i(0) - \bar{Y}(0)).$$

We center the treatment indicator in (3) to facilitate the theory without changing the estimator.

In contrast to usual formulations of the regression model, the residual, u_i, is random solely due to T_i [16,17,20]. This framework allows individual-level treatment effects, $\tau_i = Y_i(1) - Y_i(0)$, to vary across the sample, and is nonparametric because it makes no assumptions about the potential outcome distributions. The model does not satisfy key assumptions of the usual regression model: over the randomization distribution (R), u_i is heteroscedastic, and $E_R(u_i)$, $Cov_R(u_i, u_{i'})$, and $E_R(\tilde{T}_i u_i)$ are nonzero if τ_i varies across the sample.

The model in (3) also applies to each subgroup due to randomization. Thus, if we combine each subgroup model using the G_{ik} indicators, we obtain the following pooled model:

$$y_i = \sum_{k=1}^{K} \tau_k G_{ik} \tilde{T}_i + \sum_{k=1}^{K} \alpha_k G_{ik} + \epsilon_i, \tag{4}$$

where $\tau_k = \bar{Y}_k(1) - \bar{Y}_k(0)$ is the subgroup ATE estimand; $\alpha_k = p\bar{Y}_k(1) + (1-p)\bar{Y}_k(0)$ is the subgroup intercept; and $\epsilon_i = \sum_{k=1}^{K} G_{ik} u_{ik}$ is the error term with $u_{ik} = T_i(Y_i(1) - \bar{Y}_k(1)) + (1 - T_i)(Y_i(0) - \bar{Y}_k(0))$. We include model terms for all subgroups and exclude the grand intercept.

Consider the ordinary least squares (OLS) differences-in-mean estimator for τ_k from (4) using data on the full sample:

$$\hat{\tau}_k = \bar{y}_k^1 - \bar{y}_k^0 = \frac{1}{n_k^1} \sum_{i=1}^{n} G_{ik} T_i Y_i(1) - \frac{1}{n_k^0} \sum_{i=1}^{n} G_{ik}(1 - T_i) Y_i(0), \tag{5}$$

where $n_k^1 = \sum_{i=1}^{n} T_i G_{ik}$ and $n_k^0 = \sum_{i=1}^{n} (1 - T_i) G_{ik}$ are subgroup sizes in the treatment and control groups with sample shares, $\pi_k^1 = n_k^1/n^1$ and $\pi_k^0 = n_k^0/n^0$. We see that $\hat{\tau}_k$ is a ratio estimator because n_k^1 and n_k^0 are random variables (with hypergeometric distributions).

The finite population CLT in Theorem 4 in the study by Li and Ding [23] applies to $\hat{\tau}_k$ conditional on n_k^1 and n_k^0, which are ancillary to (independent of) the potential outcomes. There is a long-standing debate on the merits of conditional inference in such settings [30]. In our DB RCT context, we view repeated sampling over the randomization distribution as applying to the *full sample*, which leads to random subgroup allocations and the need for unconditional inference to capture what "could" occur. In contrast, a conditional analysis measures what "did" occur and parallels a blocked subgroup design with fixed subgroup sizes. Accordingly, we focus on an unconditional CLT for $\hat{\tau}_k$, but compare variance estimators using both approaches in our theory and simulations. The key difference is that an unconditional analysis leads to nonlinear ratio estimators with random numerators and denominators (subgroup sizes), which complicates the asymptotic analysis.

Our CLT is provided in Section 3.2 for a more general covariate-adjusted estimator from a working model that includes in (4) a $1xV$ vector of fixed, baseline covariates other than the subgroup indicators, $\mathbf{x}_i$, with parameter vector, $\boldsymbol{\beta}$.

$$y_i = \sum_{k=1}^{K} \tau_k G_{ik} \tilde{T}_i + \sum_{k=1}^{K} \alpha_k G_{ik} + \tilde{\mathbf{x}}_i \boldsymbol{\beta} + e_i, \tag{6}$$

where $\tilde{\mathbf{x}}_i = (\mathbf{x}_i - \sum_{k=1}^{K} G_{ik} \bar{\mathbf{x}}_k)$ are centered covariates; $\bar{\mathbf{x}}_k = \frac{1}{n_k} \sum_{i=1}^{n} G_{ik} \mathbf{x}_i$ are subgroup covariate means; and e_i is the error term. While the covariates do not enter the true RCT model in (4) and the ATE estimands do not change, they will increase precision to the extent they are correlated with the potential outcomes. We do not need to assume that the true conditional distribution of y_i given $\mathbf{x}_i$ is linear in $\mathbf{x}_i$. We define $\boldsymbol{\beta}$ in Section 3.2.

We focus on the pooled covariate model in (6) because it is commonly used in practice. In Section 3.3, we discuss extensions to models that interact $\tilde{T}_i$ with $\tilde{\mathbf{x}}_i$ and G_{ik}.

Using OLS to estimate the working model in (6) yields the following covariate-adjusted estimator for τ_k that is produced by standard OLS statistical packages:

$$\hat{\tau}_k^x = (\bar{y}_k^1 - \bar{y}_k^0) - (\bar{\mathbf{x}}_k^1 - \bar{\mathbf{x}}_k^0)\hat{\boldsymbol{\beta}}, \tag{7}$$

where $\bar{\mathbf{x}}_k^1$ and $\bar{\mathbf{x}}_k^0$ are subgroup covariate means for treatments and controls, and $\hat{\boldsymbol{\beta}}$ is the OLS estimator for $\boldsymbol{\beta}$ (see [23] for a parallel result for full sample analyses).

Our DB theory is conditional on randomizations that yield $n_k^1 > 0$ and $n_k^0 > 0$ so that $\hat{\tau}_k$ and its variance can be defined [25]. These restrictions yield subgroup allocation distributions that are truncated, but these effects disappear as n and n_k increase (where we assume $p = n^1/n$ and $\pi_k = n_k^1/n^1$ have finite limits). To see this, consider the general case where $n_k \leq n^1$ and $n_k \leq n^0$ so that either n_k^1 or n_k^0 can equal 0, and define $a_k^1 = E(n_k^1 \mid 0 < n_k^1 < n_k)$ as the expected value of the associated *Truncated hypergeometric* (n, n^1, n_k) distribution. We can express a_k^1 in terms of the non-truncated expectation, $n_k p$, using $a_k^1 = (n_k pf)\left(\frac{a_k^1}{n_k pf}\right)$, where $(n_k pf)$ is the mean of the *Truncated binomial* (n_k, p) distribution with $f = \frac{1 - p^{(n_k - 1)}}{1 - (1-p)^{n_k} - p^{n_k}}$ (derivation not shown). As n and n_k increase, then a_k^1 converges to $n_k p$ because $\left(\frac{a_k^1}{n_k pf}\right)$ converges to 1 (as a hypergeometric mean converges to a binomial mean) and f also converges to 1. A similar argument holds for n_k^0 and when $n_k > n^1$ or $n_k > n^0$ (where there is no truncation if both hold). Relatedly, in finite samples, the restrictions are likely to have little effect on our results as they will hold with probability near 1 for typical subgroup analyses. For instance, even for a very small subgroup with $n_k = 12$, $n = 40$, and $p = 0.5$, the restrictions will hold with probability 0.9999. Thus, to simplify notation, we omit the conditioning on positive subgroup allocations to each research group and use the unconditional expectations, $n_k p$ and $n_k(1 - p)$, in the analysis.

3.2 Main CLT result

To consider the asymptotic properties of $\hat{\tau}_k^{\mathbf{x}}$ (which also apply to $\hat{\tau}_k$ without covariates), we consider a hypothetical increasing sequence of finite populations where $n \to \infty$. Parameters should be subscripted by n, but we omit this notation for simplicity. We assume that $n^1/n \to p^*$ as $n \to \infty$, so the numbers of treatments and controls both increase with n. In addition, we assume that $n_k/n \to \pi_k^*$ for all k, where $\pi_k^* > 0$ and $\sum_{k=1}^K \pi_k^* = 1$. This implies that each subgroup also grows with n, where the number of subgroups, K, is assumed fixed.

Our CLT builds on Schochet et al. [14] who provided CLTs for RCT ratio estimators with general weights for clustered, blocked designs. We adapt these methods to our setting by treating the subgroup indicators, G_{ik}, as "weights" when computing the subgroup sample means.

Before presenting our CLT, we need to define several terms. First, for $t \in \{1,0\}$, let $\varepsilon_{ik}(t) = (Y_i(t) - \bar{Y}_k(t) - \tilde{\mathbf{x}}_i\boldsymbol{\beta})$ denote model residuals for subgroup k, where $R_{ik}(t) = \frac{G_{ik}}{\pi_k}\varepsilon_{ik}(t)$ are scaled residuals using the normalized weights, $w_i/\bar{w} = G_{ik}/\pi_k$, that sum to n. Second, let $S_{R_k}^2(t) = \frac{1}{n-1}\sum_{i=1}^n R_{ik}^2(t)$ denote the variance of $R_{ik}(t)$, and let $S_{R_k}^2(1,0) = \frac{1}{n-1}\sum_{i=1}^n R_{ik}(1)R_{ik}(0)$ denote the treatment-control covariance. Third, we define $\bar{D}_k$ as the mean treatment-control difference in the $R_{ik}(t)$ residuals, with associated variance:

$$\mathrm{Var}(\bar{D}_k) = \frac{S_{R_k}^2(1)}{n^1} + \frac{S_{R_k}^2(0)}{n^0} - \frac{S^2(\tau_k)}{n}, \tag{8}$$

where $S^2(\tau_k) = \frac{1}{n-1}\sum_{i=1}^n (R_{ik}(1) - R_{ik}(0))^2$ is the heterogeneity of treatment effects. Fourth, we define the variance of G_{ik} as $S^2(G_k) = \frac{1}{n-1}\sum_{i=1}^n \frac{1}{\pi_k^2}(G_{ik} - \pi_k)^2 = \frac{n}{(n-1)}\frac{(1-\pi_k)}{\pi_k}$. Fifth, we require the variances of each covariate, $S_{x_k,v}^2 = \frac{1}{n-1}\sum_{i=1}^n \frac{G_{ik}}{\pi_k^2}([\tilde{\mathbf{x}}_i]_v)^2$ for $v \in \{1,\ldots,V\}$, and the full variance-covariance matrix for the covariates, $\mathbf{S}_{\mathbf{x},k}^2 = \frac{1}{n}\sum_{i=1}^n \frac{G_{ik}}{\pi_k}\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i$. Finally, we need two outcome-covariate variance-covariance matrices: $\mathbf{S}_{\mathbf{x},Y,k}^2(t) = \frac{1}{n}\sum_{i=1}^n \frac{G_{ik}}{\pi_k}\tilde{\mathbf{x}}_i'Y_i(t)$ and $\mathbf{S}_{\mathbf{x}Y,k}^2(t) = \frac{1}{n}\sum_{i=1}^n \frac{1}{\pi_k}(G_{ik}\tilde{\mathbf{x}}_i'Y_i(t) - \bar{\boldsymbol{\theta}}_k)^2$, where $\bar{\boldsymbol{\theta}}_k = \frac{1}{n}\sum_{i=1}^n G_{ik}\tilde{\mathbf{x}}_i'Y_i(t)$ is the mean covariance.

We now present our CLT theorem, proved in Supplementary Materials S1.

Theorem 1. *Assume* (*C1*), (*C2*), *and the following conditions for* $t \in \{1, 0\}$ *and* $k \in \{1,\ldots, K\}$, *with fixed* $K \geq 1$:

(*C3*) Letting $g_k(t) = \max_{1\leq i\leq n}\{R_{ik}^2(t)\}$, as $n \to \infty$,

$$\frac{1}{(n^t)^2}\frac{g_k(t)}{\text{Var}(\bar{D}_k)} \to 0.$$

(*C4*) $f^1 = n^1/n$ and $f^0 = n^0/n$ have limiting values, p^* and $(1 - p^*)$, for $0 < p^* < 1$.

(*C5*) The subgroup shares, n_k/n, converge to π_k^* for $0 < \pi_k^* < 1$ and $\sum_{k=1}^K \pi_k^* = 1$.

(*C6*) As $n \to \infty$,

$$(1 - f^t)\frac{S^2(G_k)}{n^t} \to 0.$$

(*C7*) Letting $h_v(t) = \max_{1\leq i\leq n}\left\{\frac{G_{ik}}{\pi_k}([\tilde{\mathbf{x}}_i]_v)\right\}^2$ for all $v \in \{1,\ldots, V\}$, as $n \to \infty$,

$$\frac{1}{\min(n^1, n^0)}\frac{h_v(t)}{S_{x,v}^2} \to 0.$$

(*C8*) $S_{R_k}^2(t)$, $S_{R_k}^2(1,0)$, $S_{x_k,v}^2$, $\mathbf{S}_{\mathbf{x},k}^2$, $\mathbf{S}_{\mathbf{x},Y,k}^2(t)$, and $\mathbf{S}_{\mathbf{x}Y,k}^2(t)$ have finite limiting values.

Then, as $n \to \infty$, $\hat{\tau}_k^{\text{x}}$ is a consistent estimator for τ_k, and

$$\frac{\hat{\tau}_k^{\text{x}} - (\bar{Y}_k(1) - \bar{Y}_k(0))}{\sqrt{\text{Var}(\bar{D}_k)}} \xrightarrow{d} N(0, 1),$$

where $\text{Var}(\bar{D}_k)$ is defined as in (8).

Remark 1. The $\text{Var}(\bar{D}_k)$ expression in (8) is difficult to interpret because $S_{R_k}^2(t)$ and $S^2(\tau_k)$ are scaled by $\pi_k^2(n - 1)$ to facilitate the theory. To address this, we apply the following relations in (8): $n^1\pi_k^2(n - 1) = n_k p(n_k - \pi_k)$ and $n^0\pi_k^2(n - 1) = n_k(1 - p)(n_k - \pi_k)$, which yields,

$$\text{Var}(\bar{D}_k) = \phi_k\left[\frac{\Omega_{R_k}^2(1)}{n_k p} + \frac{\Omega_{R_k}^2(0)}{n_k(1 - p)} - \frac{\Omega^2(\tau_k)}{n_k}\right], \tag{9}$$

where $\Omega_{R_k}^2(t) = \frac{1}{n_k - 1}\sum_{i=1}^n G_{ik}\varepsilon_{ik}^2(t)$; $\Omega^2(\tau_k) = \frac{1}{n_k - 1}\sum_{i=1}^n G_{ik}(\varepsilon_{ik}(1) - \varepsilon_{ik}(0))^2$; and $\phi_k = (n_k - 1)/(n_k - \pi_k) \leq 1$ is a correction term that reflects the single treatment indicator "shared" by each subgroup (and can be ignored as it converges to 1). The $\Omega_{R_k}^2(t)$ and $\Omega^2(\tau_k)$ terms are population variances for those in subgroup k, and $n_k p$ and $n_k(1 - p)$ are expected subgroup sizes in the two research groups. This variance expression is more intuitive as it parallels the full sample asymptotic results in Li and Ding [23], the key difference being that (9) is based on expected subgroup sizes rather than actual ones. Note that for $\phi_k = 1$, (9) is the same as for an RCT that stratifies on subgroup k to select fixed subgroup sample sizes, $n_k p$ and $n_k(1 - p)$.

Remark 2. The first two terms in (9) pertain to separate variances for the two research groups because we allow for heterogeneous treatment effects. The third term pertains to the treatment-control covariance, $\Omega_{R_k}^2(1,0)$, expressed in terms of the heterogeneity of treatment effects, $\Omega^2(\tau_k)$, which cannot be identified from the data but can be bounded [31].

Remark 3. (*C3*) and (*C7*) are Lindeberg-type conditions from Li and Ding [23] that control the tails of the potential outcome and covariate distributions. (*C6*) yields a weak law of large numbers for the observed subgroup shares so that $\pi_k^t/\pi_k \xrightarrow{p} 1$ (using Theorem B in Scott and Wu [24]). This condition is used to account for the randomness in n_k^t, because, for instance, it allows us to express the sample mean for the treatment group as $\left(\frac{np}{n_k^1}\right)\sum_{i=1}^n \frac{G_{ik}T_iY_i(1)}{np}$, where the bracketed term converges to 1 by (*C6*) and the denominator in the summation term is fixed, so we can apply the CLT results in Li and Ding [23] and Slutsky's theorem. While (*C6*) is implied by *(C4)* and (*C5*), it facilitates the addition of other weights (Section 3.3). (*C8*) specifies limiting values of the variances and variance-covariance matrices.

Remark 4. Theorem 1 is proved in two stages by expressing the ATE estimator in (7) as,

$$\hat{\tau}_k^{\mathrm{x}} = \hat{\tau}_k^{\mathrm{x}\beta} - (\bar{\mathbf{x}}_k^1 - \bar{\mathbf{x}}_k^0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}), \tag{10}$$

where $\hat{\tau}_k^{\mathrm{x}\beta} = (\bar{y}_k^1 - \bar{y}_k^0) - (\bar{\mathbf{x}}_k^1 - \bar{\mathbf{x}}_k^0)\boldsymbol{\beta}$ and $\boldsymbol{\beta} = \left(\sum_{k=1}^K \pi_k \mathbf{S}^2_{\mathbf{x},k}\right)^{-1} [\sum_{k=1}^K p\pi_k \mathbf{S}^2_{\mathbf{x},Y,k}(1) + \sum_{k=1}^K (1-p)\pi_k \mathbf{S}^2_{\mathbf{x},Y,k}(0)]$ is assumed known. This $\boldsymbol{\beta}$ parameter is the (hypothetical) population OLS coefficient that would result from a regression of $[pY_i(1) + (1-p)Y_i(0)]$ on the covariates. In the first stage, we obtain a CLT for $\hat{\tau}_k^{\mathrm{x}\beta}$. In the second stage, we prove that $\hat{\tau}_k^{\mathrm{x}}$ has the same asymptotic distribution as $\hat{\tau}_k^{\mathrm{x}\beta}$ by showing that $(\bar{\mathbf{x}}_k^1 - \bar{\mathbf{x}}_k^0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = o_p(n^{-1/2})$, which holds under our conditions because $\bar{\mathbf{x}}_k^1$ and $\bar{\mathbf{x}}_k^0$ are both asymptotically normal and $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \xrightarrow{p} \mathbf{0}$.

Remark 5. Under (*C1*)–(*C6*), Theorem 1 also applies to $\hat{\tau}_k$ for the model without covariates by setting $\boldsymbol{\beta} = \mathbf{0}$ in (6) and in the residuals, $\varepsilon_{ik}(t)$ and $R_{ik}(t)$, which enter the variances in (8) and (9). In Supplement S2.1, we prove that $\hat{\tau}_k$ is unbiased using the approach reported by Miratrix et al. [25] and Schochet [18], where we condition on $n_k^1 > 0$ and $n_k^0 > 0$ and then average over possible subgroup allocations (A) to the two research groups to show that $E_R(\hat{\tau}_k) = E_A E_R(\hat{\tau}_k | n_k^1, n_k^0) = \tau_k$. Similarly, using the law of total variance, Supplement S2.1 shows that

$$\mathrm{Var}_R(\bar{D}_k) = E_A\left(\frac{1}{n_k^1}\right)\Omega^2_{R_k}(1) + E_A\left(\frac{1}{n_k^0}\right)\Omega^2_{R_k}(0) - \frac{\Omega^2(\tau_k)}{n_k}, \tag{11}$$

where $\Omega^2_{R_k}$ and Ω^2 are defined in (9) with $\boldsymbol{\beta} = \mathbf{0}$. Note that $\lim_{n\to\infty} E_A\left(\frac{1}{n_k^1}\right) = \frac{1}{E_A(n_k^1)} = \frac{1}{n_k p}$ and $\lim_{n\to\infty} E_A\left(\frac{1}{n_k^0}\right) = \frac{1}{n_k(1-p)}$ which aligns with (9) in large samples. In finite samples, however, the variance in (11) is at least as large as in (9) because $E_A\left(\frac{1}{n_k^1}\right) \geq \frac{1}{n_k p}$ and $E_A\left(\frac{1}{n_k^0}\right) \geq \frac{1}{n_k(1-p)}$ by Jensen's inequality. Our simulations include both sets of sample sizes (Section 5).

The following corollary to Theorem 1, proved in Supplement S1, provides the joint asymptotic distribution of the subgroup estimators, $(\hat{\tau}_1^{\mathrm{x}}, \ldots, \hat{\tau}_K^{\mathrm{x}})$.

Corollary 1. *Under the conditions of Theorem* 1, as $n \to \infty$, the ATE estimators, $\hat{\tau}_k^{\mathrm{x}}$ *and* $\hat{\tau}_{k'}^{\mathrm{x}}$ *for two subgroups* k *and* k', *are asymptotically independent, for* $(k, k') \in \{1, \ldots, K\}$. *Further, the joint asymptotic distribution of the* K *subgroup ATE estimators,* $(\hat{\tau}_1^{\mathrm{x}}, \ldots, \hat{\tau}_K^{\mathrm{x}})$, *is multivariate normal.*

This corollary is important for real-world applications because it supports the use of standard F-tests (or chi-square tests) to test the null hypothesis of equal subgroup effects.

3.3 Extensions to related estimators

This section outlines extensions of our CLT result to post-stratification estimators, models that interact $\tilde{T}_i$ with $\tilde{\mathbf{x}}_i$ and G_{ik}, BTs, and the use of nonresponse weights to adjust for missing outcome data.

Post-stratification estimators. Miratrix et al. [25] considered variance estimation for a DB post-stratification ATE estimator that obtains overall effects for the model without covariates by averaging $\hat{\tau}_k$ across subgroups. With model covariates, we can express this estimator as, $\hat{\tau}_{PS}^{\mathrm{x}} = \frac{1}{n}\sum_{k=1}^K n_k \hat{\tau}_k^{\mathrm{x}}$. Corollary 1 from above can then be applied to yield a new CLT for this estimator: $\frac{\hat{\tau}_{PS}^{\mathrm{x}} - (\bar{Y}(1) - \bar{Y}(0))}{\sqrt{\mathrm{Var}(\bar{D}_{PS})}} \xrightarrow{d} N(0,1)$, where $\mathrm{Var}(\bar{D}_{PS}) = \frac{1}{n^2}\sum_{k=1}^K n_k^2 \mathrm{Var}(\bar{D}_k)$.

Interacted models. Theorem 1 can be extended to a model that replaces $\tilde{\mathbf{x}}_i\boldsymbol{\beta}$ in (6) with the interaction terms, $\sum_{k=1}^K G_{ik}(1 - T_i)\tilde{\mathbf{x}}_i\boldsymbol{\beta}_k^0$ and $\sum_{k=1}^K G_{ik}T_i\tilde{\mathbf{x}}_i\boldsymbol{\beta}_k^1$, which allows covariate effects to differ by subgroup and treatment status. The ATE estimator for this model is, $\hat{\tau}_k^{\mathrm{x}GT} = [\bar{y}_k^1 - (\bar{\mathbf{x}}_k^1 - \bar{\mathbf{x}}_k)\hat{\boldsymbol{\beta}}_k^1] - [\bar{y}_k^0 - (\bar{\mathbf{x}}_k^0 - \bar{\mathbf{x}}_k)\hat{\boldsymbol{\beta}}_k^0]$. Theorem 1 can then be applied by redefining the residuals as, $R_{ik}(t) = \frac{G_{ik}}{\pi_k}(Y_i(t) - \bar{Y}_k(t) - (\mathbf{x}_i - \bar{\mathbf{x}}_k)\boldsymbol{\beta}_k^t)$, where $\boldsymbol{\beta}_k^t = (\mathbf{S}^2_{\mathbf{x},k})^{-1}\mathbf{S}^2_{\mathbf{x},Y,k}(t)$.

The proof (not shown) follows using the same arguments as in Remark 4 by replacing (10) with $\hat{\tau}_k^{xGT} = \hat{\tau}_k^{xGT\beta} - \sum_{t=0}^{1}(\bar{\mathbf{x}}_k^t - \bar{\mathbf{x}}_k)(\hat{\boldsymbol{\beta}}_k^t - \boldsymbol{\beta}_k^t)$, and noting that under our regularity conditions, $\hat{\boldsymbol{\beta}}_k^t - \boldsymbol{\beta}_k^t \xrightarrow{p} \mathbf{0}$ and $(\bar{\mathbf{x}}_k^t - \bar{\mathbf{x}}_k)(\hat{\boldsymbol{\beta}}_k^t - \boldsymbol{\beta}_k^t) = o_p(n^{-1/2})$ for $t \in \{1,0\}$, so that $\sum_{t=0}^{1}(\bar{\mathbf{x}}_k^t - \bar{\mathbf{x}}_k)(\hat{\boldsymbol{\beta}}_k^t - \boldsymbol{\beta}_k^t) = o_p(n^{-1/2})$.

BTs. Our results also extend to BTs where each sample member is independently randomized to the treatment group with probability p, leading to random treatment-control sizes. This design pertains, e.g., to an RCT with rolling study intake. First, we can show that our result in Remark 5 on the unbiasedness of $\hat{\tau}_k$ for the model without covariates also applies to BTs. To see this, consider the full sample analysis. Then, Bernoulli sampling has the same properties as a two-stage design that first randomly selects n^1 from a truncated binomial distribution, and then selects a simple random sample of size n^1 to the treatment group [32]. Thus, this setting parallels the one in Remark 5 that calculates sample moments by first conditioning on subgroup sizes. The only difference is that E_A is now taken over a truncated binomial rather than truncated hypergeometric distribution. The same argument applies to the subgroup analysis as for the full sample analysis.

Second, we can adapt the CLT in Theorem 1 to BTs by using expected rather than actual sample sizes in the theorem conditions (i.e., by replacing n^1 and n^0 with np and $n(1-p)$). To see this, consider again the full sample analysis, omitting the $k = 1$ subscript for simplicity. Let p^1 be the observed treatment share, and express the observed mean outcomes as, $\bar{y}^1 = \bar{y}_{BT}^1 g^1$ and $\bar{y}^0 = \bar{y}_{BT}^0 g^0$, where $\bar{y}_{BT}^1 = \frac{1}{np}\sum_{i=1}^{n} T_i y_i$ and $\bar{y}_{BT}^0 = \frac{1}{n(1-p)}\sum_{i=1}^{n}(1-T_i)y_i$ are divided by expected sample sizes rather than actual ones, and $g^1 = \frac{p}{p^1}$ and $g^0 = \frac{(1-p)}{(1-p^1)}$. Note that $g^t \xrightarrow{p} 1$ for $t \in \{1,0\}$, so $\bar{y}^t$ and $\bar{y}_{BT}^t$ have the same asymptotic distributions by Slutsky's theorem. Then, under our conditions, Theorem 4 in the study by Li and Ding [23] provides a CLT for $(\bar{y}_{BT}^1 - \bar{y}_{BT}^0)$, and the same proof as in Supplement S1.2 for Theorem 1 extends this CLT to $(\bar{y}^1 - \bar{y}^0)$. A similar approach yields a CLT for the covariate-adjusted ATE estimator using the variance in (9) by applying $\bar{\mathbf{x}}^t = \bar{\mathbf{x}}_{BT}^t g^t$ and noting that the asymptotic properties of $\hat{\boldsymbol{\beta}}$ do not change from Theorem 1. The same argument applies to the subgroup analysis.

Data nonresponse weights. Theorem 1 also applies, with additional assumptions, to a "subgroup" analysis that adjusts for missing outcome data using respondents only with nonresponse weights, w_i^r. To show this, let $R_i(T_i)$ denote an indicator of potential data response in the treatment or control condition, where $R_i(T_i) = 1$ for a respondent and 0 for a nonrespondent. Further, let $r_i = R_i(T_i)$ denote the observed response indicator. If baseline covariates are available for the full sample, a common approach is to set $w_i^{rt} = 1/e^t(\mathbf{x}_i)$ as the weight for a respondent in research group $t \in \{1,0\}$, where $e^t(\mathbf{x}_i) = \Pr(r_i = 1 | \mathbf{x}_i, T_i = t)$ is the propensity score [33].

We invoke two missing data assumptions for each subgroup: (i) data are missing at random for each research group conditional on covariates [34]: $Y_i(1), Y_i(0) \perp\!\!\!\perp r_i \mid T_i = t, G_{ik} = 1, \mathbf{x}_i$; and (ii) data response and treatment status are independent conditional on the covariates: $r_i \perp\!\!\!\perp T_i \mid G_{ik} = 1, \mathbf{x}_i$. The first ignorability (selection-on-observables) condition – which is commonly invoked for observational studies using inverse probability weighting methods [35,36] – ensures that weighted ATE estimators using the respondent sample will consistently estimate τ_k. The second condition implies that the response, $r_i = R_i(T_i) = R_i$, will be *identical* in the treatment and control conditions, so that $e^t(\mathbf{x}_i) = e(\mathbf{x}_i)$ and $w_i^{rt} = w_i^r$ are independent of t. Stated differently, this condition implies that respondents and their weights are randomly allocated to the two research groups. Thus, the summed weights, $\sum_{i:T_i=t} w_i^r$, which enter the denominators of the weighted differences-in-means estimators become random, which parallels the subgroup analysis from above with random subgroup sizes in the two research groups.

Accordingly, we can apply Theorem 1, assuming known weights, where the respondent sample and weighted least squares are used to obtain the ATE estimator, $\hat{\tau}_k^{rx}$, using (6). We assume $e(\mathbf{x}_i)$ is known and converges to $e^*(\mathbf{x}_i)$ as $n \to \infty$, where $0 < e^*(\mathbf{x}_i) \leq 1$ for all $\mathbf{x}_i$ in its finite population support so that (*C6*) holds. For the proof, we replace the weights, $w_i = G_{ik}$, in the theorem with $w_i = G_{ik} r_i w_i^r$ to define the variables and regularity conditions. The resulting variance for the CLT has the same form as (9) but is based on expected subgroup respondent sizes (Supplement S2.2). Developing a finite population CLT that allows for estimated nonresponse weights rather than known weights is a topic for future research.

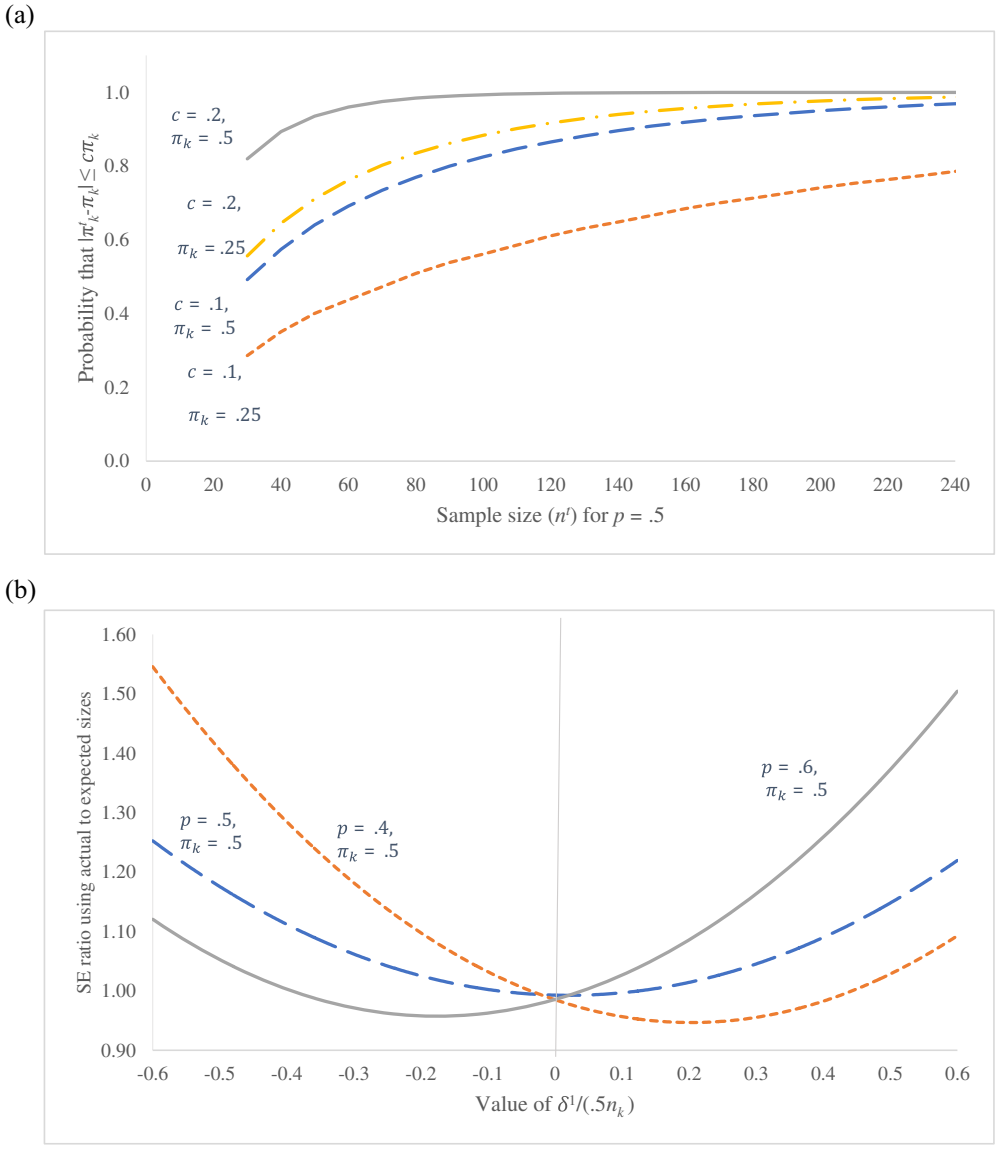

Figure 1: (a) Probabilities for the differences, $(\pi_k^t - \pi_k)$, relative to π_k, and (b) SE ratios using actual to expected subgroup sizes, as a function of $\delta^1/5n_k$ for $\delta^1 = (n_k^1 - n_k p)$. Note: See text for definitions and formulas. (b) assumes n = 100, φ = 1.1, and ϑ = 0.05. SE = Standard error.

3.4 Variance estimation

To obtain consistent variance estimators for (9) (and model variants), we can either use expected subgroup sizes, $n_k p$ and $n_k(1 - p)$, or actual ones, n_k^1 and n_k^0, as for the conditional analysis. Our simulations find very similar results using either approach (Section 5). This occurs because the difference between the hypergeometric random variable, π_k^t, and its expected value, π_k, decreases exponentially with n^t [37,38]. For instance, Figure 1a shows that for modest n^t, there is a high probability that $|\pi_k^t - \pi_k|/\pi_k \leq c$ for small c (defined as 10 or 20%).

Further, Figure 1b shows that for small c, the ratios of SEs using actual to expected subgroup sizes in (9) are close to 1, leading to similar confidence interval coverage. For example, for n = 100, p = 0.5, and π_k = 0.5, the ratios range only from 1 to 1.026 as c ranges from 0 to 0.2 (assuming $\Omega^2_{R_k}(1) = \varphi\Omega^2_{R_k}(0)$ and $\Omega^2(\tau_k) = \vartheta\Omega^2_{R_k}(0)$ with plausible values, φ = 1.1 and ϑ = 0.05). In *expectation*, the SE ratios are *greater* than 1 for all values of φ and ϑ (Remark 5 above), but the differences are small for typical subgroup sizes used in practice.

To further examine the pattern of the SE ratios in Figure 1b, suppose first that $\varphi = 1$. Then, all ratios are at least 1 when $p = 0.5$. However, if $p < 0.5$, the ratios are greater than 1 if $\delta^1 = n_k^1 - n_k p < 0$ or $\delta^1 > n_k(1-2p)$, but are less than 1 otherwise, and vice versa for $p > 0.5$. As a function of δ^1, the ratios are convex and symmetric around their minimum value at $\delta^1 = 0.5n_k(1-2p)$ when $n_k^1 = n_k^0$. This symmetry is lost when $\varphi \neq 1$, but the same overall patterns apply (Figure 1b).

Using expected sizes, a consistent (upper bound) plug-in variance estimator for (9) based on estimated subgroup regression residuals is as follows:

$$\hat{\text{Var}}(\bar{D}_k) = \frac{s^2_{R_k}(1)}{n_k p} + \frac{s^2_{R_k}(0)}{n_k(1-p)}, \tag{12}$$

where

$$s^2_{R_k}(1) = \frac{1}{(n_k^1 - Vp\pi_k^1 - 1)} \sum_{i=1}^{n} T_i G_{ik}(y_i - \hat{\alpha}_k^{\mathrm{x}} - (1-p)\hat{\tau}_k^{\mathrm{x}} - \tilde{\mathbf{x}}_i\hat{\boldsymbol{\beta}})^2$$

and

$$s^2_{R_k}(0) = \frac{1}{(n_k^0 - V(1-p)\pi_k^0 - 1)} \sum_{i=1}^{n} (1-T_i) G_{ik}(y_i - \hat{\alpha}_k^{\mathrm{x}} + p\hat{\tau}_k^{\mathrm{x}} - \tilde{\mathbf{x}}_i\hat{\boldsymbol{\beta}})^2.$$

Here we set $\phi_k = 1$, which can be relaxed by subtracting π_k^t in the denominators of $s^2_{R_k}(t)$ rather than 1. In (12), the losses in degrees of freedom (*df*) due to the V covariates are split proportionately across the K subgroups and two research conditions. Note that the same estimator results using a non-centered model in (6) that replaces the $G_{ik}\tilde{T}_i$ and $\tilde{\mathbf{x}}_i$ terms with $G_{ik}T_i$ and $\mathbf{x}_i$. Hypothesis testing can be conducted using t-tests with $df = (n_k - V\pi_k - 2)$ or z-tests. Using actual sizes, we can instead use n_k^1 and n_k^0 in (12) rather than $n_k p$ and $n_k(1-p)$.

As shown in Supplement S3, (12) is asymptotically equivalent to the robust Huber–White (HW) variance estimator [39,40], as has been shown for full sample estimators [16,20,41]. In finite samples, however, the DB variances will typically be larger for the model without covariates due to larger *df* corrections. We compare the two estimators in our simulations, along with other SE variants.

4 Blocked and clustered designs

The above CLT results extend directly to blocked RCTs where randomization is performed separately within strata (e.g., sites, demographic groups, or time cohorts), and to clustered RCTs where groups (e.g., schools, hospitals, or communities) are randomized rather than individuals.

4.1 Blocked RCTs

In blocked designs, the sample is first divided into subpopulations, and a mini-experiment is conducted in each one. Note that we do not consider blocks formed by subgroups slated for ATE estimation as the theory for the full sample analysis applies in this case (as n_k^1 and n_k^0 are fixed).

For the blocked design, we use similar notation as above with the addition of the subscript $b = (1, 2, \ldots, B)$ to indicate blocks. For instance, T_{ib} is the treatment indicator, p_b is the block treatment assignment rate, n_b is the number of persons in block b, $G_{ibk}$is the subgroup indicator, n_{bk} is the size of subgroup k, $\pi_{bk} = n_{bk}/n_b$ is the subgroup share, and $Y_{ib}(t)$ is the potential outcome. Further, we define S_{ib} as a 1/0 indicator of block membership and $q_b = n_b/n$ as the block population share. We assume SUTVA and complete randomization within each block, where vectors of possible treatment assignments are mutually independent across blocks.

With this notation, we can now define the ATE estimand for blocks containing members of subgroup k as $\tau_{bk} = \bar{Y}_{bk}(1) - \bar{Y}_{bk}(0)$, and the pooled ATE estimand across such blocks as,

$$\tau_k = \frac{\sum_{b:\pi_{bk}>0}^{B} w_{bk}\tau_{bk}}{\sum_{b:\pi_{bk}>0}^{B} w_{bk}}, \tag{13}$$

where w_{bk} is the block weight that can differ across subgroups. We set $w_{bk} = n_{bk}$, but other options exist [42]. We allow $n_{bk} = n_b$ and $n_{bk} = 0$.

Consider OLS estimation of the following extension of (6) to blocked RCTs:

$$y_{ib} = \sum_{k=1}^{K} \tau_{bk} S_{ib} G_{ibk} \tilde{T}_{ib} + \sum_{k=1}^{K} \alpha_{bk} S_{ib} G_{ibk} + \tilde{\mathbf{x}}_{ib}\boldsymbol{\beta} + \eta_{ib}, \tag{14}$$

where $\tilde{T}_{ib} = T_{ib} - p_b$ and $\tilde{\mathbf{x}}_{ib} = \mathbf{x}_{ib} - \sum_{k=1}^{K} S_{ib} G_{ibk} \bar{\mathbf{x}}_{bk}$ are block-centered variables; $\bar{\mathbf{x}}_{bk} = \frac{1}{n_{bk}} \sum_{i=1}^{n_b} G_{ibk} \mathbf{x}_{ib}$ are covariate means; and η_{ib} is the error term. The OLS ATE estimator for τ_{bk} in (14) is,

$$\hat{\tau}^{\mathrm{x}}_{bk} = (\bar{y}^1_{bk} - \bar{y}^0_{bk}) - (\bar{\mathbf{x}}^1_{bk} - \bar{\mathbf{x}}^0_{bk})\hat{\boldsymbol{\beta}}, \tag{15}$$

where $\bar{y}^t_{bk}$ and $\bar{\mathbf{x}}^t_{bk}$ are observed treatment and control group means.

Because a mini-experiment is conducted in each block, we can apply Theorem 1 to $\hat{\tau}^{\mathrm{x}}_{bk}$ as $n \to \infty$ for fixed B. This yields the following finite population CLT for the blocked design.

Theorem 2. *Assume* (*C1*)–(*C4*) *and* (*C6*)–(*C8*) *for each included block, and the following conditions for* $b \in \{1, \ldots, B\}$ *and* $k \in \{1, \ldots, K\}$, *for fixed* $B \geq 1$ *and* $K \geq 1$:

(*C4a*) The block shares, $n_b/n \to q^*_b$ as $n \to \infty$, where $q^*_b > 0$ and $\sum_{b=1}^{B} q^*_b = 1$.

(*C5a*) The subgroup shares, $n_{bk}/n_b \to \pi^*_{bk}$ as $n \to \infty$, with $0 \leq \pi^*_{bk} \leq 1$ and $\sum_{k=1}^{K} \pi^*_{bk} = 1$.

Then, as $n \to \infty$ for fixed B and K, $\hat{\tau}^{\mathrm{x}}_{bk}$ is a consistent estimator for τ_{bk}, and

$$\frac{\hat{\tau}^{\mathrm{x}}_{bk} - (\bar{Y}_{bk}(1) - \bar{Y}_{bk}(0))}{\sqrt{\mathrm{Var}(\bar{D}_{bk})}} \xrightarrow{d} N(0,1),$$

where $\mathrm{Var}(\bar{D}_{bk})$ is defined as in (8) or (9) at the block level.

The proof (not shown) parallels the one for Theorem 1, applied to each block, by redefining the residual as, $R_{ibk}(t) = \frac{G_{ibk}}{\pi_{bk}}(Y_{ib}(t) - \bar{Y}_{bk}(t) - (\mathbf{x}_{ib} - \bar{\mathbf{x}}_{bk})\boldsymbol{\beta})$, while invoking (*C4a*) that allows n_b to grow with n, and (*C5a*) that amends (*C5*) so that π_{bk} can equal 0 or 1. Note that Liu and Yang [43] considered asymptotics for full sample estimators as $B \to \infty$.

A variance estimator for $\hat{\tau}^{\mathrm{x}}_{bk}$ in (15), $\hat{\mathrm{Var}}(\bar{D}_{bk})$, can be obtained using (12), where $s^2_{R_{bk}}(t)$ is now calculated using residuals from the fitted model in (14). The *df* adjustments are $(n^1_{bk} - Vq_b p_b \pi^1_{bk} - 1)$ for $s^2_{R_{bk}}(1)$ and $(n^0_{bk} - Vq_b(1-p_b)\pi^0_{bk} - 1)$ for $s^2_{R_{bk}}(0)$.

Next we provide a corollary to Theorem 2 on the pooled subgroup estimator across blocks, $\hat{\tau}^{\mathrm{x}}_{k,Pooled} = \frac{1}{n_k} \sum_{b:n_{bk}>0}^{B} n_{bk} \hat{\tau}^{\mathrm{x}}_{bk}$, where each block is weighted by its subgroup size.

Corollary 2. *Under the conditions of Theorem 2, as* $n \to \infty$ *for fixed B and K,* $\hat{\tau}^{\mathrm{x}}_{k,\mathrm{Pooled}}$ *is a consistent estimator* for $\tau_{k,\mathrm{Pooled}} = \frac{1}{n_k} \sum_{b:n_{bk}>0}^{B} n_{bk}\tau_{bk}$, *and*

$$\frac{1}{\sqrt{\mathrm{Var}(\bar{D}_k)}}(\hat{\tau}^{\mathrm{x}}_{k,\mathrm{Pooled}} - \tau_{k,\mathrm{Pooled}}) \xrightarrow{d} N(0,1),$$

where $\mathrm{Var}(\bar{D}_k) = \frac{1}{(n_k)^2} \sum_{b:n_{bk}>0}^{B} n^2_{bk} \mathrm{Var}(\bar{D}_{bk})$.

This result follows because the $\hat{\tau}^{\mathrm{x}}_{bk}$ estimators are asymptotically independent across blocks, which can be shown using the same arguments as in the proof of Corollary 1 in Supplement S1. We can estimate $\mathrm{Var}(\bar{D}_k)$

using $\hat{\text{Var}}(\bar{D}_{bk})$ for each included block. Hypothesis testing for $\hat{\tau}_k^{\text{x}}$ can be conducted using t-tests with $df = (n_k - V\pi_k - 2B)$ or z-tests.

Finally, a future research topic is to develop a CLT for a restricted model that controls for block main effects but excludes block-by-treatment interactions. An example of such a model is to replace the first set of interactions in (14) with $\sum_{k=1}^{K}\tau_{k,R}G_{ibk}\tilde{T}_{ib}$. In this case, the OLS ATE estimator for subgroup k is $\hat{\tau}_{k,R} = \frac{1}{\sum_b w_{bk,R}}\sum_b w_{bk,R}\hat{\tau}_{bk}^{\text{x}}$, where $w_{bk,R} = n_{bk}\,p_{bk}^1(1 - p_{bk}^1)$ and $p_{bk}^1 = n_{bk}^1/n_{bk}$. Thus, $\hat{\tau}_{k,R}$ uses a form of precision weighting to weight the block-specific estimators. It is inconsistent but uses fewer parameters. Full sample CLTs for this estimator are considered in [14,21], which become more complex in the subgroup context.

4.2 Clustered RCTs

In clustered RCTs, groups rather than individuals are the unit of randomization. Consider a clustered, non-blocked RCT with m total clusters, where $m^1 = mp$ is assigned to the treatment group and $m^0 = m(1 - p)$ is assigned to the control group. All persons in the same cluster have the same treatment assignment. Let m_k denote the number of clusters in subgroup k, where m_k^1 and m_k^0 are observed counts. We assume that individual-level data are available for analysis, although our results also pertain to data averaged to the cluster level.

We index clusters by j. Thus, we have that $T_j = 1$ for treatment clusters and 0 for control clusters, n_{jk} is the number of subgroup k members in cluster j, $Y_{ij}(t)$ is the potential outcome for person i in cluster j, and so on. We also assume SUTVA and complete randomization as generalized to clustered RCTs [14].

Consider an individual-level subgroup ($G_{ijk} = 1$) where $\pi_{jk} > 0$ for all j. In this case, $m_k^1 = m^1$ and $m_k^0 = m^0$ are fixed, and the ATE estimand for subgroup k under the clustered RCT is

$$\tau_k = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n_j}G_{ijk}(Y_{ij}(1) - Y_{ij}(0))}{n_k} = \frac{\sum_{j=1}^{m}n_{jk}(\bar{Y}_{jk}(1) - \bar{Y}_{jk}(0))}{n_k} = \bar{\bar{Y}}_k(1) - \bar{\bar{Y}}_k(0), \tag{16}$$

where $\bar{Y}_{jk}(t) = \frac{1}{n_{jk}}\sum_{i=1}^{n_j}G_{ijk}Y_{ij}(t)$ is the mean cluster-level outcome and $\bar{\bar{Y}}_k(t)$ is the grand mean. Here clusters are weighted by their subgroup sizes, $w_{jk} = n_{jk}$, but other options exist, such as weighting clusters equally to estimate subgroup effects per cluster rather than per person.

Applying OLS to (6) with clustered data yields the following subgroup ATE estimator:

$$\hat{\tau}_{k,\text{clus}}^{\text{x}} = (\bar{\bar{y}}_k^1 - \bar{\bar{y}}_k^0) - (\bar{\bar{\mathbf{x}}}_k^1 - \bar{\bar{\mathbf{x}}}_k^0)\hat{\boldsymbol{\beta}}, \tag{17}$$

where $\bar{\bar{y}}_k^t = \frac{1}{n_k^t}\sum_{j:T_j=t}\sum_{i=1}^{n_j}G_{ijk}Y_{ij}(t)$ is the mean observed outcome, and similarly for $\bar{\bar{\mathbf{x}}}_k^t$.

We see that (17) is a ratio estimator because n_k^1 and n_k^0 are random under the clustered design (if clusters are weighted unequally). However, this is also the case for the full sample estimator, because n^1 and n^0 (i.e., the summed weights) are also random. Thus, as $m \to \infty$, the full sample CLT results reported by Schochet et al. [14] for the clustered (and blocked) RCT can be applied to $\hat{\tau}_{k,\text{clus}}^{\text{x}}$. This approach is outlined in Supplement S3.2 along with a consistent variance estimator using a version of (12) based on estimated cluster-level residuals. Parallel to the HW analysis, Supplement S3.2 also shows that this variance estimator is asymptotically equivalent to the cluster-robust SE estimator developed by Liang and Zeger [44].

Finally, Supplement S3.3 outlines cluster RCT results for a subgroup analysis defined by a cluster-level characteristic ($G_{jk} = 1$), such as a school, hospital, or community feature, rather than an individual-level characteristic. In this setting, m_k^1 and m_k^0 become random, which parallels the subgroup analysis for the non-clustered RCT. Note that a similar formulation also applies for the individual-level subgroup analysis when $\pi_{jk} = 0$ for some clusters.

5 Simulation analysis

We conducted simulations to examine the finite sample statistical properties of our DB subgroup ATE estimators. The focus is on the non-clustered RCT because prior full sample simulation results for the clustered RCT also pertain to individual-level subgroup analyses [14,45], as discussed above. For the simulations, we applied the variance estimator in (12) using expected and actual subgroup sizes, for models with and without covariates. We set $\phi_k = 1$ for most specifications, but also adjusted for ϕ_k for some runs. We also ran simulations using the HW estimator and several variants of (12) (Supplement S4).

5.1 Simulation setup

The following model was used to generate potential outcomes for $K = 2$ subgroups and $V = 2$ pre-treatment covariates:

$$Y_i(0) = G_{i1} + 2G_{i2} + 0.4G_{i1}x_{i1} + 0.8G_{i1}x_{i2} + 0.7G_{i2}x_{i1} + 0.5G_{i2}x_{i2} + e_i$$
$$Y_i(1) = Y_i(0) + G_{i1}\theta_{i1} + G_{i2}\theta_{i2}, \tag{18}$$

where e_i is *iid* $N(0,1)$ random error; x_{i1} and x_{i2} are *iid* $N(0,1)$ covariates; and θ_{i1} and θ_{i2} are *iid* $N(0,0.5)$ and $N(0,0.4)$ random errors that capture treatment effect heterogeneity.

We generated five draws of potential outcomes using (18) to help guard against unusual draws and report average results. For each draw, we conducted 10,000 replications, randomly assigning units to either the treatment or control group using $p = 0.5$ (or $p = 0.4$ or 0.6 for some runs), and only kept randomizations that met our minimum subgroup size criteria for variance estimation. For each replication, we estimated the model in (6) and stored the results. We ran simulations for total sample sizes of $n = 40, 100$, and 200 and Subgroup 1 shares of $\pi_1 = 0.25,\ 0.50$, and 0.75. To allow for skewed distributions, we also generated model errors and covariates for selected runs using a chi-squared distribution with the same means and variances as above.

In Supplement S4, we discuss variants of (12) used in our simulations. These include applying the *df* correction for hypothesis testing in Bell and McCaffrey [46]; subtracting a lower bound on the $\frac{1}{n_k}\Omega^2(\tau_k)$ heterogeneity term; multiplying by $(1 - R^2_{TXk})^{-1}$, where R^2_{TXk} is the R^2 from a regression of $G_{ik}\widetilde{T}_i$ on $\widetilde{\mathbf{x}}_i$ and the other terms in (6); and using the finite sample variance in (11).

5.2 Simulation results

Table 1 and Supplement Tables S1–S4 present the simulation results. Of the 300,000 draws of n_k^1 and n_k^0 used in Table 1, all yielded values of $n_k^1 > 0$ and $n_k^0 > 0$, so these restrictions have little effect on our theory. Focusing on Subgroup 1, we find negligible biases for all specifications with and without baseline covariates. Confidence interval coverage is close to 95% using t-distribution cutoff values, even with relatively small subgroup samples, but with slight over-coverage across specifications. Accordingly, Type 1 errors tend to be slightly below the nominal 5% level (Tables 1 and D.1). It is interesting that these results differ from those found for the clustered RCT where Type 1 errors tend to be inflated [14,45].

Estimated SEs are close to "true" values, as measured by the standard deviation of the ATE estimates across replications. Consistent with the theory on SE ratios in Section 3.4, the SEs are slightly larger using actual subgroup sizes than expected ones, leading to narrower confidence interval coverage using the expected sizes. Also consistent with the theory, the SEs are slightly smaller for the HW estimator for the model without covariates, and for specifications that adjust for $\phi_k < 1$. Type 1 errors for F-tests to gauge differences in Subgroup 1 and 2 effects are close to 5% but tend to be somewhat liberal (Tables S1 and S2). We find similar results using data generated from a chi-squared distribution (Table S3) and using $p = 0.4$ or 0.6 (Table S4). Finally, applying variants of the variance formula in (12) as detailed in Supplement S4 does not change the overall findings or improve performance (Table S3).

Table 1: Simulation results for the subgroup ATE estimators

Model specification	Bias of ATE estimator[a]	Confidence interval coverage	True SE[a,b]	Mean estimated SE
Model without covariates				
Sample size: n = 40, π_1 = 0.50				
Design-based (DB), actual subgroup sizes, ϕ_k = 1	−0.002	0.954	0.646	0.640
DB, expected sizes, ϕ_k = 1	−0.002	0.952	0.646	0.633
DB, actual sizes, adjust for ϕ_k	−0.002	0.948	0.646	0.621
HW	−0.002	0.953	0.646	0.639
Sample size: n = 100, π_1 = 0.50				
DB, actual sizes, ϕ_k = 1	0.000	0.958	0.376	0.385
DB, expected sizes, ϕ_k = 1	0.000	0.957	0.376	0.383
DB, actual sizes, adjust for ϕ_k	0.000	0.956	0.376	0.381
HW	0.000	0.958	0.376	0.385
Sample size: n = 100, π_1 = 0.25				
DB, actual sizes, ϕ_k = 1	0.000	0.953	0.626	0.628
DB, expected sizes, ϕ_k = 1	0.000	0.951	0.626	0.618
DB, actual sizes, adjust for ϕ_k	0.000	0.948	0.626	0.613
HW	0.000	0.948	0.626	0.613
Model with two covariates				
Sample size: n = 40, π_1 = 0.50				
DB, actual sizes, ϕ_k = 1	0.002	0.950	0.501	0.482
DB, expected sizes, ϕ_k = 1	0.002	0.948	0.501	0.476
DB, actual sizes, adjust for ϕ_k	0.002	0.944	0.501	0.467
HW	0.002	0.953	0.501	0.494
Sample size: n = 100, π_1 = 0.50				
DB, actual sizes, ϕ_k = 1	0.000	0.961	0.298	0.305
DB, expected sizes, ϕ_k = 1	0.000	0.960	0.298	0.303
DB, actual sizes, adjust for ϕ_k	0.000	0.959	0.298	0.302
HW	0.000	0.962	0.298	0.308
Sample size: n = 100, π_1 = 0.25				
DB, actual sizes, ϕ_k = 1	0.000	0.956	0.486	0.482
DB, expected sizes, ϕ_k = 1	0.000	0.954	0.486	0.475
DB, actual sizes, adjust for ϕ_k	0.000	0.951	0.486	0.470
HW	0.000	0.952	0.486	0.470

Note: See text for simulation details. The calculations assume two subgroups with a focus on results for Subgroup 1, a treatment assignment rate of p = 0.50, and normally distributed covariates and errors. For each specification, the figures are based on 10,000 simulations for each of 5 potential outcome draws, and the findings average across the 5 draws. Ordinary least square (OLS) methods are used for ATE estimation using the model in (6), and design-based SEs are obtained using (12). Huber-White estimates are obtained using the lm_robust procedure in R.

ATE = Average treatment effect; DB = Design-based; HW = Hubert–White.

[a]Biases and true SEs are the same for all specifications within each sample size category because they use the same data and OLS model for ATE estimation.

6 Empirical application using the motivating NYC voucher experiment

To demonstrate our DB subgroup ATE estimators, we used baseline and outcome data from the NYC School Choice Scholarships Foundation Program (SCSF) [9]. SCSF was funded by philanthropists to provide scholarships to public school students in grades K–4 from low-income families to attend any participating NYC private

school. In spring 1997, more than 20,000 students applied to receive a voucher. SCSF then used random lotteries to offer 3-year vouchers of up to $1,400 annually to 1,000 eligible families in the treatment group. Of the remaining families not offered the voucher, 960 were randomly selected to the control group.

SCSF assisted the treatment group in finding private-school placements. More than 78% of treatment families used a voucher, for 2.6 years on average, where 98% of users attended parochial schools. Here we focus on estimating ATEs (i.e., intention-to-treat effects on the voucher offer) for two race/ethnicity subgroups as defined in the original study [9]: African Americans and Latinos each of who comprise about 47% of the sample. The study authors hypothesized that African Americans might benefit more from the vouchers as they tended to live in more disadvantaged communities with lower-performing public schools.

Following the original study [9], the primary outcomes for our analysis are composite national percentile rankings in math and reading from the study-administered Iowa Test of Basic Skills (ITBS). We focus on first follow-up year test scores, where the response rate was 78% for treatments and 71% for controls. Our goal is not to replicate study results but to illustrate our subgroup ATE estimators.

The voucher study was a blocked RCT. Applicants from schools with average test scores below the city median were assigned a higher probability of winning a scholarship, and blocks were also formed by lottery date and family size (with 30 blocks in total). The design is also partly clustered because families were randomized, where all eligible children within a family could receive a scholarship; 30% of families had at least two children in the evaluation.

We used (14) for ATE estimation and (12) for variance estimation for each block, where blocks were weighted by their subgroup sizes to obtain the overall subgroup effects. To adjust for clustering, we averaged data to the family level. Following [9], we used weights to adjust for missing follow-up test scores. We ran models without covariates and those that included baseline ITBS scores to increase precision, though they were not collected for the entire kindergarten cohort. Following the original study, other demographic covariates were not included in the models due to the large number of blocks.

Table 2 presents the subgroup findings that mirror those from the original study. We find that the offer of a voucher had no effect on test scores overall or for Latinos across specifications. The effects on African Americans are also not statistically significant at the 5% level for the model without baseline test scores. However, the effects on African Americans become positive and statistically significant for the model with baseline scores, that excludes the kindergarteners but nonetheless yields SEs that are reduced by about 12%. These effects are 4.7 percentile ranking points, which translates into a 0.26 standard deviation increase, with a significant F-test for the subgroup interaction effect (p-value = 0.028). The effects for African Americans remain significant using the sample with baseline test scores without controlling for them in the model.

Table 2: Estimated ATEs on composite test scores for the NYC voucher experiment

Model specification	Overall sample	African American	Latino
Model excludes baseline test scores			
DB, actual subgroup sizes	0.25 (1.06)	2.54 (1.45)	−0.86 (1.58)
DB, expected subgroup sizes	0.25 (1.06)	2.54 (1.45)	−0.86 (1.58)
HW	0.25 (1.03)	2.54 (1.42)	−0.86 (1.49)
DB: actual sizes using sample with baseline test scores	0.88 (1.30)	4.47* (1.73)	−1.11 (1.76)
Model includes baseline test scores			
DB, actual subgroup sizes	1.70 (1.01)	4.70* (1.27)	0.50 (1.44)
DB, expected subgroup sizes	1.70 (1.01)	4.70* (1.27)	0.50 (1.44)
HW	1.70 (0.98)	4.70* (1.24)	0.50 (1.39)
Student sample size (without/with baseline test scores)	2,012/1,434	902/643	964/682

Note: SEs are in parentheses. See text for ATE and SE formulas. All estimates are weighted to adjust for follow-up test score nonresponse.

ATE = Average treatment effect.

*Statistically significant at the 5% level, two-tailed test.

We find across specifications that the DB SEs are nearly identical using actual and expected sample sizes. Further, consistent with theory, the DB SEs are slightly larger than the HW SEs, but both yield the same study conclusions: the vouchers did not improve test scores overall, but there is evidence they had a positive effect on African American students in grades 1–4. A detailed reanalysis of the original study data, however, cautions that the results for African Americans are sensitive to alternative race/ethnicity definitions and should be interpreted carefully [10].

7 Conclusion

This article considered DB RCT methods for ATE estimation for discrete subgroups defined by pre-treatment sample characteristics. Our subgroup estimators derive from the Neyman–Rubin–Holland model that underlies experiments and are based on simple least squares regression methods. We considered ratio estimators due to the randomness of observed subgroup sample sizes in the treatment and control groups that were not conditioned on for the asymptotic analysis. The DB approach is appealing in that it applies to continuous, binary, and discrete outcomes, and is nonparametric in that makes no assumptions about the distribution of potential outcomes or the model functional form.

We developed a new finite population, unconditional CLT for our subgroup ATE estimators under the non-clustered RCT, allowing for baseline covariates to improve precision. The main difference between our CLT and prior full sample ones is that the asymptotic variance for the subgroup estimator is based on expected subgroup sizes rather than actual ones. Another difference is that the subgroup variance includes a finite sample adjustment (ϕ_k) that reflects the single treatment indicator shared by the subgroups. To apply the estimators in practice, we discussed simple consistent variance estimators using regression residuals that are asymptotically equivalent to robust variance estimators, but with finite sample degrees-of-freedom adjustments that derive directly from the experimental design. Our re-analysis of the NYC Voucher experiment demonstrated the simplicity of the methods, while maintaining statistical rigor.

A contribution of this work is that it provides a unified DB framework for subgroup analyses across a range of RCT designs. We discussed extensions of the asymptotic theory to blocked and clustered designs. We also discussed extensions to other commonly used estimators with random treatment-control sample sizes or summed weights: post-stratification estimators that average subgroup estimators to obtain overall effects, weighted estimators to adjust for data nonresponse, and estimators from BTs.

Our simulations for the non-clustered RCT show that the subgroup ATE estimators yield low bias and confidence interval coverage near nominal levels, although with slight over-coverage. This is somewhat surprising as the simulation literature on DB and robust variance estimators for clustered RCTs – that also applies to the subgroup context – shows the opposite issue of under-coverage [14,45].

Our simulations find very similar results using either actual or expected subgroup sample sizes for variance estimation. As demonstrated in several ways, this occurs because the difference between the observed subgroup proportions, π_k^1 and π_k^0, and their expected value, π_k, decreases exponentially with the overall sample size. This finding justifies the typical approach of using actual subgroup sample sizes for variance estimation, which blurs the distinction between a subgroup analysis conditional on the observed treatment-control subgroup sizes and an unconditional subgroup analysis considered here. Thus, a conditional analysis may be preferred due to its simplicity and parallel structure to the full sample DB analysis.

The free RCT-YES software (www.rct-yes.com), funded by the U.S. Department of Education, estimates ATEs for both full sample and baseline subgroup analyses using the DB methods discussed in this article using either R or Stata. The software applies actual sample sizes for the variance formulas for subgroup analyses and allows for general weights. The software also allows for multi-armed trials with multiple treatment condition.

Acknowledgements: The author would like to thank the two reviewers for very helpful suggestions and comments.

Funding information: Author states no funding involved.

Author contribution: The author confirms the sole responsibility for the conception of the study, presented results and manuscript preparation.

Conflict of interest: Author states no conflict of interest.

Data availability statement: The NYC Voucher data for the empirical analysis were obtained under a restricted data use license agreement with Mathematica. Per license requirements, these data cannot be shared with journal readers. However, to the best of my knowledge, these data can be obtained, and I would be happy to provide the SAS and R programs used for the analysis.

References

[1] Neyman J. On the application of probability theory to agricultural experiments: essay on principles. Sect 9, Translated Stat Sci. 1990;5:465–72.
[2] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66:688–701.
[3] Rubin DB. Assignment to treatment group on the basis of a covariate. J Educ Stat. 1977;2:1–26.
[4] Holland PW. Statistics and causal inference. J Am Stat Assoc. 1986;81:945–60.
[5] Rothwell PM. Subgroup analyses in randomized controlled trials: importance, indications, and interpretation. Lancet. 2005;365:176–86.
[6] Schochet PZ, Puma M, Deke J. Understanding variation in treatment effects in education impact evaluations: an overview of quantitative methods (NCEE 2014-4017). Washington, DC: National Center for Education Evaluation and Regional Assistance; 2014.
[7] Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine-reporting of subgroup analyses in clinical trials. N Engl J Med. 2007;357(21):2189–94.
[8] Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. J Pharmacol Pharmacother. 2019;1:100–7.
[9] Mayer D, Peterson P, Myers D, Tuttle C, Howell W. School choice in New York City: an evaluation of the school choice scholarships program. Mathematica Policy Research, Washington, DC; 2002.
[10] Krueger AB, Zhu P. Another look at the New York City school voucher experiment. Am Behav Scientist 47(5):658–98.
[11] Bland JM. Cluster randomised trials in the medical literature: two bibliometric surveys. BMC Med Res Methodol. 2004;4:21.
[12] Schochet PZ. Statistical power for random assignment evaluations of education programs. J Educ Behav Stat. 2008;33:62–87.
[13] Pashley NE. Note on the delta method for finite population inference with applications to causal inference. Working Paper: Harvard University Statistics Department, Cambridge MA; 2019.
[14] Schochet PZ, Pashley NE, Miratrix LW, Kautz T. Design-based ratio estimators and central limit theorems for clustered, blocked RCTs. J Am Stat Assoc. 2022;117(540):2135–46.
[15] Yang L, Tsiatis A. Efficiency study of estimators for a treatment effect in a pretest-posttest trial. Am Statistician. 2001;55:314–21.
[16] Freedman D. On regression adjustments to experimental data. Adv Appl Math. 2008;40:180–93.
[17] Schochet PZ. Is regression adjustment supported by the Neyman model for causal inference? J Stat Plan Inference. 2010;140:246–59.
[18] Schochet PZ. Statistical theory for the RCT-YES software: design-based causal inference for RCTs: second edition (NCEE 2016–4011). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education; 2016.
[19] Aronow PM, Middleton JA. A class of unbiased estimators of the average treatment effect in randomized experiments. J Causal Inference. 2013;1:135–54.
[20] Lin W. Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique. Ann Appl Stat. 2013;7:295–318.
[21] Imbens G, Rubin D. Causal inference for statistics, social, and biomedical sciences: an introduction. Cambridge, UK: Cambridge University Press; 2015.

[22] Middleton JA, Aronow PM. Unbiased estimation of the average treatment effect in cluster-randomized experiments. Statistics Politics Policy. 2015;6:39–75.
[23] Li X, Ding P. General forms of finite population central limit theorems with applications to causal inference. J Am Stat Assoc. 2017;112:1759–69.
[24] Scott A, Wu CF. On the asymptotic distribution of ratio and regression estimators. J Am Stat Assoc. 1981;1981(112):1759–69.
[25] Miratrix LW, Sekhon JS, Yu B. Adjusting treatment effect estimates by post-stratification in randomized experiments. J R Stat Soc Ser B. 2013;75(2):369–96.
[26] Cochran W. Sampling techniques. New York: John Wiley and Sons; 1977.
[27] Lohr SL. Sampling: design and analysis. 2nd edn. Pacific Grove, CA: Duxbury Press; 2009.
[28] Thompson S. Sampling. Hoboken, NJ: John Wiley & Sons; 2012.
[29] Rubin DB. Which ifs have causal answers? Discussion of Holland's "statistics and causal inference". J Am Stat Assoc. 1986;81:961–2.
[30] Fraser DAS. Ancillaries and conditional inference. Stat Sci. 2004;19(2):333–69.
[31] Aronow PM, Green DP, Lee DKK. Sharp bounds on the variance in randomized experiments. Ann Stat. 2014;42:850–71.
[32] Wright T. On some properties of variable size simple random sampling and a limit theorem. Commun Stat Theory Methods. 1988;17(9):2997–3016.
[33] Rosenbaum P, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41–55.
[34] Rubin DB. Multiple imputation for nonresponse in surveys. NY: J. Wiley and Sons; 1987.
[35] Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. J Am Stat Assoc. 1952;47:663–85.
[36] Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. Stat Med. 2004;23(19):2937–60.
[37] Serfling RJ. Probability inequalities for the sum in sampling without replacement. Ann Stat. 1974;2:39–48.
[38] Greene E, Wellner JA. Exponential bounds for the hypergeometric distribution. Bernoulli. 2017;23(3):1911–50.
[39] Huber PJ. The behavior of maximum likelihood estimates under nonstandard conditions. Proced Fifth Berkeley Symp Math Stat Probability. 1967;1:221–33.
[40] White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica. 1980;1980(48):817–38.
[41] Su F, Ding P. Model-assisted analyses of cluster-randomized experiments. J R Stat Soc Ser B. 2021;83(5):994–1015.
[42] Pashley NE, Miratrix LW. Insights on variance estimation for blocked and matched pairs designs. J Educ Behav Stat. 2021;46(3):271–96.
[43] Liu H, Yang Y. Regression-adjusted average treatment effect estimates in stratified randomized experiments. Biometrika. 2020;107(4):935–48.
[44] Liang K, Zeger S. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73:13–22.
[45] Cameron AC, Miller DL. A practitioner's guide to cluster-robust inference. J Hum Resour. 2015;50:317–72.
[46] Bell R, McCaffrey D. Bias reduction in standard errors for linear regression with multi-stage samples. Surv Methodol. 2002;28:169–81.