Unifying Design-based Inference:
A New Variance Estimation Principle

Joel A. Middleton*

April 4, 2021

*Charles and Louise Travers Department of Political Science, *University of California, Berkeley.*
email: joel.middleton@gmail.com

1 Introduction

In Unifying paper 1 of 4 a generalized sandwich variance estimators was proposed that is applicable to virtually any experimental design, any linear estimator and any variance bound (defined therein), subsuming other well known sandwich variance estimators (which go by names such as "robust", "cluster robust", "heteroskedastic consistent", "sandwich", "White", "Huber-White", "HC", "CR", etc.). This paper presents two novel classes of variance estimators with superior properties, in the absence of parametric or semi-parametric assumptions.

The first new class of estimator is the Obloženè Chlebìžky (OC) variance estimators as a novel alternative to the generalized sandwich in Paper 1 of 4. That the OC concept is unlikely to arise from other, more standard, frameworks is manifestly true in light of the 40 year lacuna since White (1980). For any member of the generalized sandwich variance estimator class, there is an OC with the same expected value. The this alternative replaces a random matrix at the center with a nonrandom one. The second type of estimator is guaranteed conservative for the variance of the estimator and is based upon a similar principle of replacing a random matrix with its (nonrandom) expectation.

There is heuristic appeal in eliminating variance components from variance estimators, because it may mean that the OC may be more precise than the sandwich upon which it is based, though this may need not be true in general. In simulations based on a real randomized experiment, reductions in variance of variance estimators is substantial, and guaranteed conservative variance estimators do not suffer from small sample bias the way so-called robust, sandwich-type variance estimators do.

2 Notation

Consider a randomized experiment with k treatment arms. The Neyman causal model (NCM) assumes that the units in the experimental study represent a finite population of size n. For a given outcome measure, call it y, each unit, i, responds with one of k possible values in $\{y_{1i}, y_{2i}, ..., y_{ki}\}$, depending on their treatment assignment. The possible responses are referred to as the *potential outcomes*. In the NCM these values are considered (nonrandom) constants, which stands in contrast to other, more common, formulations where potential outcomes are assumed to be sampled from some (possibly nonparametric) joint distribution.

The only random elements in the NCM are the treatment assignment indicators $\{R_{1i}, R_{2i}, ..., R_{ki}\}$, and they determine which potential outcome will be observed by the researcher. Since a unit, i, can only be assigned to one arm of the experiment these random indicators must sum to unity, i.e., $R_{1i} + R_{2i} + ... + R_{ki} = 1$. A standard representation of the *observed* response for the ith unit under the NCM would be,

$$Y_i^{obs} = y_{1i}R_{1i} + y_{2i}R_{2i} + ... + y_{ki}R_{ki},$$

which is itself random, due to the assignment indicators. The observed data can then be represented as $\{Y_i^{obs}, R_{1i}, R_{2i}, ..., R_{ki}, x_i\}_{\forall i}$, where x_i is an additional vector of k covariates. Like the potential outcomes, x_i is considered to be nonrandom, and, unlike the potential outcomes, does not depend on the assignment, which might be ensured, in practice, by collecting the covariates before the assignment.

Ideally, a researcher might like to know the difference between responses under various arms for a given individual, i

for example, $y_{2i} - y_{1i}$, or perhaps $\frac{1}{2}(y_{2i} - y_{1i} + y_{4i} - y_{3i})$ and so on, which are examples of different *contrasts* between potential outcomes.

However, it is clear from the definition of Y_i^{obs} that individual treatment effects are not identified since only one of the two potential outcomes can be observed. This is known as *fundamental problem of causal inference* (Holland, 1986). As a result, we often study averages of these values over the units of study. To continue with the above examples, a researcher might be satisfied to estimate averages over these individual treatment effects. Continuing the example, we have the average treatment effects $n^{-1}\sum_i (y_{2i} - y_{1i})$ and $n^{-1}\sum_i \frac{1}{2}(y_{2i} - y_{1i} + y_{4i} - y_{3i})$, respectively.

To simplify notation, let y_1, y_2,...,y_k represent length n vectors of potential outcomes associated with each of the arms, with the ith element of each corresponding to the ith unit. Next, define

$$y := (y_1' \ y_2' \ \cdots \ y_k')',$$

which is the length kn, representing all k potential outcomes for each of the n units.

Next, if we let 1_n be a n-length vector of ones, then a $kn \times k$ *intercept matrix* can be defined as,

$$\mathbb{1} := \begin{bmatrix} 1_n & & & \\ & 1_n & & \\ & & \ddots & \\ & & & 1_n \end{bmatrix},$$

which, for example, allows us to express a k-length vector of means of the arms as as $n^{-1}\mathbb{1}'y$ or, equivalently, $(\mathbb{1}'\mathbb{1})^{-1}\mathbb{1}'y$.

Given a length-k *contrast vector*, c, we can contrast the means for the various arms. In keeping with the running example with four assignment arms, if one is interested in the average difference in responses to the first two treatments, then $c = (-1 \ 1 \ 0 \ 0)'$ and we have $n^{-1}c'\mathbb{1}'y = n^{-1}\sum_i (y_{2i} - y_{1i})$. Likewise, if one chooses $c = (-\frac{1}{2} \ \frac{1}{2} \ -\frac{1}{2} \ \frac{1}{2})'$ then $n^{-1}c'\mathbb{1}'y = n^{-1}\sum_i \frac{1}{2}(y_{2i} - y_{1i} + y_{4i} - y_{3i})$. Sensible contrasts such as these meet the conditions $\sum_{j=1}^k c_j = 0$ and $\sum_{j=1}^k |c_j| = 2$, though this would not be strictly necessary, mathematically speaking.

Next define an $n \times n$ diagonal matrix that has all n assignment indicators for treatment arm 1 on the diagonal,

$$\mathbf{R}_1 := \begin{bmatrix} R_{11} & & & & & \\ & R_{12} & & & & \\ & & \ddots & & & \\ & & & R_{1i} & & \\ & & & & \ddots & \\ & & & & & R_{1n} \end{bmatrix},$$

and define $\mathbf{R}_2$, $\mathbf{R}_3$, ..., $\mathbf{R}_k$ analogously. Arange these matrices to create the diagonal $kn \times kn$ matrix

$$\mathbf{R} := \begin{bmatrix} \mathbf{R}_1 & & & \\ & \mathbf{R}_2 & & \\ & & \ddots & \\ & & & \mathbf{R}_k \end{bmatrix}$$

3

and note the a $kn \times kn$ diagonal matrix of assignment probabilities can be written as $\boldsymbol{\pi} := \mathrm{E}[\mathbf{R}]$, with the first n diagonal elements representing probabilities of assignment to arm 1, then the next n diagonal elements are probabilities of assignment to arm 2 and so on.

In this alternative notation the researcher can be said to observe the assignment, $\mathbf{R}$, the observed vector of outcomes, $\mathbf{R}y$, and also a matrix of l pre-treatment covariates, $\mathbf{x}$, which has size $n \times l$. In a randomized experiment $\boldsymbol{\pi}$ is also observed (known) in many cases. When intractable analytically, however, it might be estimated to arbitrary precision by repeating the original randomization until a target level of precision is achieved.

For covariate adjusted estimators, it will also be convenient to define the $kn \times (k + l)$ matrix,

$$
\varkappa := \begin{bmatrix} 1_n & & & & \mathbf{x} \\ & 1_n & & & \mathbf{x} \\ & & \ddots & & \vdots \\ & & & 1_n & \mathbf{x} \end{bmatrix}.
$$

Remark 1. *For some cases, such adjusting for covariates separately by arm, it might be useful to define $\varkappa$ with $\mathbf{x}$ matrices arranged along a block-diagonal. In that case, it is prudent to stipulate that $\mathbf{x}$ have columns that sum to zero to avoid problems of coefficient interpretation (cf. Lin, 2013; Middleton, 2018). This will be discussed further in paper 3 of 4.*

3 The class of WLS estimators

This paper will discuss the class of estimators that can be characterized as weighted least squares (WLS), which includes the difference-of-means, OLS and Hajek estimators as special cases. For the purposes of giving asymptotic variance expressions for this class, the Horvitz-Thompson estimator is also introduced. Paper 3 of 4 will introduce a wider class of estimators that will include these as special cases, and specifically examine their properties as point estimators. This paper is primarily concerned with a new approach to variance bound estimation, irrespective of what the estimator *estimates*.

As in Paper 1 of 4, *linear estimators* are defined as having the form,

$$
\widehat{\delta}_c := c'\mathbf{W}\mathbf{R}y, \tag{1}
$$

where $\mathbf{W}$ a matrix with kn columns and k rows if it is an unadjusted esimator and $k + l$ rows if it is a covariate adjusted estimator. The length of the contrast vector, c, is equal to the number of rows in $\mathbf{W}$. The first k entries of c are the contrast values, followed by l zeros in the case of a covariate adjusted estimator.

Definition 3.1 (Horvitz-Thompson estimator). *The Horvitz-Thompson estimator written as in equation (1) with,*

$$
\mathbf{W} = \mathbf{w}^{\text{HT}} := \left(\mathbb{1}'\mathbb{1}\right)^{-1} \mathbb{1}'\boldsymbol{\pi}^{-1}.
$$

Definition 3.2 (WLS estimators). *WLS estimators can be written as in equation (1) with,*

$$
\mathbf{W} = \mathbf{W}^{\text{WLS}} := \left(\varkappa'\mathbf{m}\mathbf{R}\varkappa\right)^{-1} \varkappa'\mathbf{m}
$$

where the $kn \times kn$ matrix $\mathbf{m}$ must have nonegative entries on the diagonal (with some strictly positive) and zeros elsewhere. Also, define

$$\mathbf{w}^{\mathrm{WLS}} := \mathbf{W}^{\mathrm{WLS}}|_{\mathbf{R}=\boldsymbol{\pi}}$$
$$= \left(\varkappa' \mathbf{m} \boldsymbol{\pi} \varkappa\right)^{-1} \varkappa' \mathbf{m},$$

which is $\mathbf{W}^{\mathrm{WLS}}$ with $\boldsymbol{\pi} = \mathrm{E}\left[\mathbf{R}\right]$ replacing $\mathbf{R}$.

Remark 2. *WLS gives OLS as a special case when $\mathbf{m} = \mathbf{i}_{kn}$ ($\mathbf{i}_{kn}$ is the identity matrix). If $\mathbf{m} = \mathbf{i}_{kn}$ and, in addition, $\varkappa = \mathbb{1}$ (there are no covariates), WLS is equivalent to the difference-of-means. If $\mathbf{m} = \boldsymbol{\pi}^{-1}$ and $\varkappa = \mathbb{1}$, then it is the Hajek estimator. The covariate adjusted WLS with $\mathbf{m} = \boldsymbol{\pi}^{-1}$ will be discussed further in paper 3 of 4, because it is algebraically equivalent to the generalized regression estimator introduced there.*

Theorem 3.3. *For a suitably chosen constant, a, and vector of constants, z_c, the first-order Taylor approximation for a WLS can be written as,*

$$\widehat{\delta}_c^{\mathrm{T(WLS)}} = a + n 1'_{(k+l)} \mathbf{w}^{\mathrm{HT}} \mathbf{R} z_c,$$

with

$$z_c := \boldsymbol{\pi} \mathrm{diag}\left(u\right) \mathbf{w}^{\mathrm{WLS}\prime} c,$$

and where

$$u := y - \varkappa b^{\mathrm{WLS}}$$

is the "true" residual and

$$b^{\mathrm{WLS}} := \mathbf{w}^{\mathrm{WLS}} \boldsymbol{\pi} y$$

is the "true" WLS coefficient.

Remark 3. *By "true" it is meant to suggest that these are nonrandom quantities that, for large enough samples, represent approximately the centers of the distributions of observed residuals and WLS coefficient, respectively.*

Remark 4. *While the vectors z_c, u and b^{WLS} are not directly observable, the result is useful because it shows that the first-order Taylor approximation of a linear estimator is a Horvitz-Thompson estimator with contrast vector $c = n 1_{(k+l)}$ and "outcome" vector z_c. Expressing its variance will provide a basis for an asymptotically valid approximation. The constant a is unimportant for the purposes of variance expressions, and so its explicit form is not given.*

Corollary 3.3.1 (The Obložene Chlebìžky Principle)**.** *In Theorem 3.3 we could have also given an equivalent expression for, z_c, as,*

$$z_c^{equiv} := \boldsymbol{\pi} \, \mathrm{diag}\left(c' \mathbf{w}^{\mathrm{WLS}}\right) u$$
$$= z_c.$$

Remark 5. *Corollary 3.3.1 gives the key insight in the development of the Obložene Chlebìžky variance estimators. Residuals need not be in the center of the sandwich.*

4 Variance bounds and their estimation

As in Paper 1 of 4, define 1_{kn} as a kn-length vector of ones and the $kn \times kn$ "design matrix" as,

$$\mathbf{d} := \mathrm{V}\left(1'_{kn}\boldsymbol{\pi}^{-1}\mathbf{R}\right). \tag{2}$$

This is a variance-covariance matrix of weighted treatment assignments.

An exact expression for first-order Taylor approximations of linear estimators can be written as,

$$\mathrm{V}\left(\widehat{\delta}_c^{\mathrm{T(WLS)}}\right) = z_c'\mathbf{d}z_c. \tag{3}$$

Definition 4.1 (Bounding Matrix). *Let y be an outcome vector and let the contrast vector be $c = n1_k$. Then the associated Horvitz-Thompson estimator is $n1_k\mathbf{w}^{\mathrm{HT}}\mathbf{R}y$ with variance $y'\mathbf{d}y$. Then, the arbitrary $kn \times kn$ matrix, $\tilde{\mathbf{d}}$, is a "bounding matrix" if, for all $y \in \mathbb{R}^{kn}$, $y'\mathbf{d}y \leq y'\tilde{\mathbf{d}}y$.*

Remark 6. *For more on bounding matrices, see Paper 1 of 4. Unless otherwise noted, the Generalized Neyman Bounding matrix, $\tilde{\mathbf{d}} = \tilde{\mathbf{d}}^{\mathrm{N}}$, will be assumed.*

Now, define the $2n \times 2n$ matrix of probabilities and joint probabilities of assignment,

$$\mathbf{p} := \mathrm{E}\left[\mathbf{R}1_{2n}1'_{2n}\mathbf{R}\right].$$

Next define an inverse probability weighted version of bounding matrix, $\tilde{\mathbf{d}}$, as

$$\tilde{\mathbf{d}}_{/\mathbf{p}} := \tilde{\mathbf{d}}/\mathbf{p} \tag{4}$$

with / denoting element-wise division defined such that division by zero equals zero.

Hypothetically, if the observed outcome vector was $\mathbf{R}z_c$, then the Horvitz-Thompson estimator with $c = n1_k$ has variance $z_c'\tilde{\mathbf{d}}z_c$, which could be estimated unbiasedly by

$$\widehat{\mathrm{V}}\left(\widehat{\delta}_{n1_k}^{\mathrm{HT}}\right) := z_c'\mathbf{R}\tilde{\mathbf{d}}_{/\mathbf{p}}\mathbf{R}z_c. \tag{5}$$

Of course, the vector z_c is defined by quantities which are not directly observed. However, we might identify random quantities that approximate z_c and u as

$$
\begin{aligned}
Z_c :=& \boldsymbol{\pi}\mathrm{diag}\left(U\right)\mathbf{W}^{\mathrm{WLS}'}c \\
=& \boldsymbol{\pi}\mathrm{diag}\left(c'\mathbf{W}^{\mathrm{WLS}}\right)U
\end{aligned}
$$

and

$$U := y - \mathbf{x}\mathbf{W}^{\mathrm{WLS}}\mathbf{R}y,$$

respectively.

Definition 4.2 (The Generalized Sandwich Variance Estimator). *The Generalized Sandwich Variance Estimator is,*

$$
\begin{aligned}
\widehat{\tilde{\mathrm{V}}}^{\mathrm{GS}}\left(\widehat{\delta}_c^{\mathrm{T(WLS)}}\right) :=& Z_c'\mathbf{R}\tilde{\mathbf{d}}_{/\mathbf{p}}\mathbf{R}Z_c \\
=& c'\mathbf{W}^{\mathrm{WLS}}\,\mathrm{diag}\left(\mathbf{R}U\right)\left(\boldsymbol{\pi}\tilde{\mathbf{d}}_{/\mathbf{p}}\boldsymbol{\pi}\right)\mathrm{diag}\left(\mathbf{R}U\right)\mathbf{W}^{\mathrm{WLS}'}c \\
=& y'\left\{\mathbf{M}\,\mathrm{diag}\left(c'\mathbf{W}^{\mathrm{WLS}}\right)\left(\boldsymbol{\pi}\tilde{\mathbf{d}}_{/\mathbf{p}}\boldsymbol{\pi}\right)\,\mathrm{diag}\left(c'\mathbf{W}^{\mathrm{WLS}}\right)\mathbf{M}\right\}y \\
=& y'\mathbf{O}_{(0)}y
\end{aligned}
\tag{6}
$$

where $\mathbf{M} := (\mathbf{R} - \mathbf{R}_x\mathbf{W}^{\mathrm{WLS}}\mathbf{R})$ *is a "residual maker" matrix and* $\mathbf{O}_{(0)}$ *denotes the random matrix in curly brackets in the line just above with the subscript* (0) *differentiating* $\mathbf{O}_{(0)}$ *from* $\mathbf{O}_{(1)}$ *and* $\mathbf{O}_{(2)}$, *introduced below.*

The third line uses the Obloženè Chlebìžky principle from Corollary 3.3.1 and the equality $\mathbf{R}U = \mathbf{M}y$.

Remark 7. *The generalized sandwich is equivalent to Eicker-Huber-White variance estimator under Bernoulli designs and* $\mathbf{m} = \mathbf{i}_{kn}$. *Likewise it is algebraically equivalent to "cluster robust" standard errors (CR0) under Bernoulli assignment of clusters and* $\mathbf{m} = \mathbf{i}_{kn}$ *(see Paper 1 of 4 for more details). Refinements for degrees-of-freedom and leverage are easily accommodated, and simulations will compare sandwiches HC0, HC1, and HC2 to the parallel OC estimators.*

5 Obloženè Chlebìžky variance bound estimators

The last line of equation (6) has used the Obloženè Chlebìžky principle from Corrolary 3.3.1 to rearrange terms in the generalized sandwich. Written like this, it is easy to see that the expectation of the generalized sandwich is,

$$\mathrm{E}\left[\widehat{\mathrm{V}}^{\mathrm{GS}}\left(\widehat{\delta}_c^{\mathrm{T(WLS)}}\right)\right] = y'\mathrm{E}\left[\mathbf{O}_{(0)}\right]y$$
$$= y'\mathbf{o}_{(0)}y \tag{7}$$

with $\mathbf{o}_{(0)} := \mathrm{E}\left[\mathbf{O}_{(0)}\right]$. The matrix $\mathbf{o}_{(0)}$ can be computed for small samples. For larger samples, it may be simulated to arbitrary precision by drawing from the randomization distribution.

5.1 OC0: A "baseline" variance estimator that is not invariant

Definition 5.1 (OC0). *Next, an estimator with the same expected value as the Generalized Sandwich in Equation* (6), *is the OC0 estimator,*

$$\widehat{\mathrm{V}}^{\mathrm{OC0}}\left(\widehat{\delta}_c^{\mathrm{T(WLS)}}\right) := y'\mathbf{R}\mathbf{o}_{(0/\mathbf{p})}\mathbf{R}y,$$

with $\mathbf{o}_{(0/\mathbf{p})} := \mathbf{o}_{(0)}/\mathbf{p}$ *where, as above, "/" is element-wise division with division by zero resolving to zero.*

The key insight with respect to OC0 is that it is an unbiased estimator of the expected value of the Generalized Sandwich, i.e.,

$$\mathrm{E}\left[\widehat{\mathrm{V}}^{\mathrm{OC0}}\left(\widehat{\delta}_c^{\mathrm{T(WLS)}}\right)\right] = \mathrm{E}\left[\widehat{\mathrm{V}}^{\mathrm{GS}}\left(\widehat{\delta}_c^{\mathrm{T(WLS)}}\right)\right],$$

because $\mathrm{E}\left[\mathbf{R}\mathbf{o}_{(0/\mathbf{p})}\mathbf{R}\right] = \mathbf{o}_{(0)}$.

OC0 is a Horvitz-Thompson estimator, and therefore lacks invariance to location shifts in the outcome variable. Likewise, its variance may depend on the scaling of y and so it may also be less precise than covariate-adjusted alternatives. In the next sections, refinements that are invariant are considered.

5.2 OC1: Covariate-adjustment achieves invariance, but introduces bias

To improve the precision of OC0, consider the covariate adjusted alternative,

$$\widehat{\widehat{V}}^{\text{OC1}}\left(\widehat{\delta}_c^{\text{T(WLS)}}\right) := y'\left\{\mathbf{M}\mathbf{o}_{(0\!/\!\mathbf{P}}\mathbf{M}\right\}y \tag{8}$$

$$= y'\mathbf{O}_{(1)}y \tag{9}$$

with $\mathbf{M} := (\mathbf{R} - \mathbf{R}_{\!\!/\!\!}\mathbf{W}^{\text{WLS}}\mathbf{R})$, as above, the residual maker for WLS. Due to the covariate adjustment, OC1 may be more precise than OC0, and OC1 is also invariant to location shifts in y. It is, nonetheless, biased for the mean of the Generalized Sandwich. In the next section, a bias correction is considered.

5.3 OC2: An unbiased and invariant OC variance estimator

Next the bias of OC1 relative to the expected value of the Generalized Sandwich variance,

$$\mathbf{Bias}^{\text{(GS)}}\left[\widehat{\widehat{V}}^{\text{OC1}}\left(\widehat{\delta}_c^{\text{T(WLS)}}\right)\right] = \mathrm{E}\left[\widehat{\widehat{V}}^{\text{OC1}}\left(\widehat{\delta}_c^{\text{T(WLS)}}\right) - \widehat{\widehat{V}}^{\text{GS}}\left(\widehat{\delta}_c^{\text{T(WLS)}}\right)\right]$$

$$= y'\left\{\mathrm{E}\left[\mathbf{M}\mathbf{o}_{(0\!/\!\mathbf{P}}\mathbf{M} - \mathbf{R}\mathbf{o}_{(0\!/\!\mathbf{P}}\mathbf{R}\right]\right\}y. \tag{10}$$

Next, define the degree 4, $kn \times kn \times kn \times kn$ tensor,

$$\mathfrak{B}^{abcd} := \left(\mathbf{M}^{ab}\cdot\mathbf{M}^{cd} - \mathbf{R}^{ab}\cdot\mathbf{R}^{cd}\right)\circ\left(\mathbf{p}^{ad}\cdot\mathbf{p}^{bc}\right)^{\circ-\frac{1}{2}},$$

with a, b, c and d giving labels for dimensions, $\cdot$ represents the tensor multiplication using Einstein's convention of inner product for dimensions with identical labels and outer product otherwise, $\circ$ is element-wise (Hadamard) multiplication and $(.)^{\circ-\frac{1}{2}}$ is element wise exponentiation by $-\frac{1}{2}$, with division by zero resolving to zero.

With this tensor notation, the bias in Equation (10) can be written,

$$\mathbf{Bias}^{\text{(GS)}}\left[\widehat{\widehat{V}}^{\text{OC0.B}}\left(\widehat{\delta}_c^{\text{T(WLS)}}\right)\right] = y'\left\{\left(\mathrm{E}\left[\mathfrak{B}^{abcd}\right]\circ\left(\mathbf{p}^{ad}\cdot\mathbf{p}^{bc}\right)^{\circ\frac{1}{2}}\right)\cdot\mathbf{o}^{bc}_{(0\!/\!\mathbf{P}}\right\}y$$

$$= y'\left\{\left(\mathfrak{b}^{abcd}\circ\left(\mathbf{p}^{ad}\cdot\mathbf{p}^{bc}\right)^{\circ\frac{1}{2}}\right)\cdot\mathbf{o}^{bc}_{(0\!/\!\mathbf{P}}\right\}y$$

with $\mathfrak{b}^{abcd} := \mathrm{E}\left[\mathfrak{B}^{abcd}\right]$, which can be estimated to arbitrary precision using repeated randomizations.

Next taking the tensor SVD we can decompose as follows,

$$\mathfrak{b}^{abcd} = -\mathfrak{u}^{ade}\cdot\lambda^e\cdot\mathfrak{u}^{bce}$$

$$= -\mathfrak{u}^{ade}_{(0<\lambda<1)}\cdot\lambda^e_{(0<\lambda<1)}\cdot\mathfrak{u}^{bce}_{(0<\lambda<1)} \quad -\mathfrak{u}^{ade}_{(\lambda\geq1)}\cdot\lambda^e_{(\lambda\geq1)}\cdot\mathfrak{u}^{bce}_{(\lambda\geq1)}$$

$$= \mathfrak{b}^{abcd}_{(0<\lambda<1)} \quad + \quad \mathfrak{b}^{abcd}_{(\lambda\geq1)},$$

with λ^e representing a length k^2n^2 vector (tensor of degree 1) of singular values and $\mathfrak{u}^{ade}$ as a $kn \times kn \times k^2n^2$ tensor. In the second line, the tensor is decomposed into two terms. The first term, with subscripts $(0 < \lambda < 1)$, is the tensor constructed from the slices of $\mathfrak{u}$ corresponding to singular values greater than 0 but less than 1. The second term, with subscripts $(\lambda \geq 1)$, is constructed from the slices of $\mathfrak{u}$ corresponding to singular values greater than or equal to 1. Note that the orthogonality property, such that the tensor product $\mathfrak{u}^{ade}\cdot\mathfrak{u}^{adf}$ results in the identity matrix, $\mathbf{i}^{ef}$.

Conjecture 5.2. *The singular values, λ^e, are bounded by zero and one, i.e., $0 \le \lambda^e \le 1$.*

Lemma 5.3. *The infinite tensor series*

$$
\begin{aligned}
\mathfrak{b}^{abcd}_{(\infty)} \quad &:= \quad \mathfrak{b}^{abcd}_{(0<\lambda<1)} \quad - \quad \mathfrak{b}^{aefd}_{(0<\lambda<1)} \cdot \mathfrak{b}^{ebcf}_{(0<\lambda<1)} \quad + \quad \mathfrak{b}^{aefd}_{(0<\lambda<1)} \cdot \mathfrak{b}^{eghf}_{(0<\lambda<1)} \cdot \mathfrak{b}^{gbch}_{(0<\lambda<1)} \\
&\quad - \quad \mathfrak{b}^{aefd}_{(0<\lambda<1)} \cdot \mathfrak{b}^{eghf}_{(0<\lambda<1)} \cdot \mathfrak{b}^{gijh}_{(0<\lambda<1)} \cdot \mathfrak{b}^{ibcj}_{(0<\lambda<1)} \quad + \quad \dots
\end{aligned}
$$

has closed form,

$$
\mathfrak{b}^{abcd}_{(\infty)} = -\mathfrak{u}^{ade}_{(0<\lambda<1)} \cdot \phi^e_{(0<\lambda<1)} \cdot \mathfrak{u}^{bce}_{(0<\lambda<1)}
$$

where $\phi^e_{(0<\lambda<1)} := \lambda^e_{(0<\lambda<1)}/(1 - \lambda^e_{(0<\lambda<1)})$ with / representing element-wise division.

Theorem 5.4. *An unbiased estimator of the bias of OC1 as an estimator of the mean of the Generalized Sandwich is,*

$$
\begin{aligned}
\widehat{\mathrm{Bias}}^{(\mathrm{GS})}\left[\widehat{\mathrm{V}}^{\mathrm{OC1}}\left(\widehat{\delta}^{\mathrm{T(WLS)}}_c\right)\right] :=& y'\mathbf{M}\left\{\left(\mathfrak{b}^{abcd}_{(\infty)} \circ \left(\mathbf{p}^{ad \circ -\frac{1}{2}} \cdot \mathbf{p}^{bc \circ \frac{1}{2}}\right)\right) \cdot \mathbf{o}^{bc}_{(0)/\mathbf{P}}\right\}\mathbf{M}y \qquad (11) \\
&+ y'\mathbf{R}\left\{\left(\mathfrak{b}^{abcd}_{(\lambda \ge 1)} \circ \left(\mathbf{p}^{ad \circ -\frac{1}{2}} \cdot \mathbf{p}^{bc \circ \frac{1}{2}}\right)\right) \cdot \mathbf{o}^{bc}_{(0)/\mathbf{P}}\right\}\mathbf{R}y,
\end{aligned}
$$

where $\mathbf{M}$ is the residual-maker, as above.

Conjecture 5.5. *The second term in Equation (11) is zero for all randomizations. Hence,*

$$
\widehat{\mathrm{Bias}}^{(\mathrm{GS})}\left[\widehat{\mathrm{V}}^{\mathrm{OC1}}\left(\widehat{\delta}^{\mathrm{T(WLS)}}_c\right)\right] = y'\mathbf{M}\left\{\left(\mathfrak{b}^{abcd}_{(\infty)} \circ \left(\mathbf{p}^{ad \circ -\frac{1}{2}} \cdot \mathbf{p}^{bc \circ \frac{1}{2}}\right)\right) \cdot \mathbf{o}^{bc}_{(0)/\mathbf{P}}\right\}\mathbf{M}y
$$

is an unbiased estimator of the bias OC1 relative to the Generalized Sandwich.

Definition 5.6 (OC2). *The OC2 variance estimator is,*

$$
\begin{aligned}
\widehat{\mathrm{V}}^{\mathrm{OC2}}\left(\widehat{\delta}^{\mathrm{T(WLS)}}_c\right) :=& y'\mathbf{M}\mathbf{o}_{(0)/\mathbf{P}}\mathbf{M}y - y'\mathbf{M}\left\{\left(\mathfrak{b}^{abcd}_{(\infty)} \circ \left(\mathbf{p}^{ad \circ -\frac{1}{2}} \cdot \mathbf{p}^{bc \circ \frac{1}{2}}\right)\right) \cdot \mathbf{o}^{bc}_{(0)/\mathbf{P}}\right\}\mathbf{M}y \\
=& y'\left\{\mathbf{M}\left(\left(\mathbf{i}^{ab}_{kn} \cdot \mathbf{i}^{cd}_{kn} - \mathfrak{b}^{abcd}_{(\infty)} \circ \left(\mathbf{p}^{ad \circ -\frac{1}{2}} \cdot \mathbf{p}^{bc \circ \frac{1}{2}}\right)\right) \cdot \mathbf{o}^{bc}_{(0)/\mathbf{P}}\right)\mathbf{M}\right\}y \\
=& y'\mathbf{O}_{(2)}y
\end{aligned}
$$

Remark 8. *If true, Conjecture (5.5) implies that OC2 is unbiased as an estimator of the Generalized Sandwich variance. The presence of the residual-maker matrix, $\mathbf{M}$, means that OC2 is invariant to location shifts in the outcome, y.*

5.4 Comparing the precision of OC2 and GS

Using tensor notation, the variance of the OC2 variance estimator can be written,

$$
\mathrm{V}\left(\widehat{\mathrm{V}}^{\mathrm{OC2}}\right) = y^a \cdot y^b \cdot \left\{\mathrm{E}\left[\mathbf{O}^{ab}_{(2)} \cdot \mathbf{O}^{cd}_{(2)}\right] - \mathrm{E}\left[\mathbf{O}^{ab}_{(2)}\right] \cdot \mathrm{E}\left[\mathbf{O}^{cd}_{(2)}\right]\right\} \cdot y^c \cdot y^d.
$$

Similarly, the variance of the GS variance estimator can be written,

$$
\mathrm{V}\left(\widehat{\mathrm{V}}^{\mathrm{GS}}\right) = y^a \cdot y^b \cdot \left\{\mathrm{E}\left[\mathbf{O}^{ab}_{(0)} \cdot \mathbf{O}^{cd}_{(0)}\right] - \mathrm{E}\left[\mathbf{O}^{ab}_{(0)}\right] \cdot \mathrm{E}\left[\mathbf{O}^{cd}_{(0)}\right]\right\} \cdot y^c \cdot y^d.
$$

So, with Conjecture (5.5), we have the difference of variances,

$$\text{V}\left(\widehat{\widehat{\text{V}}}^{\text{GS}}\right) - \text{V}\left(\widehat{\widehat{\text{V}}}^{\text{OC2}}\right) = y^a \cdot y^b \cdot \left\{ \text{E}\left[\mathbf{O}^{ab}_{(0)} \cdot \mathbf{O}^{cd}_{(0)}\right] - \text{E}\left[\mathbf{O}^{ab}_{(2)} \cdot \mathbf{O}^{cd}_{(2sim)}\right] \right\} \cdot y^c \cdot y^d.$$

with $\text{E}\left[\mathbf{O}^{ab}_{(0)} \cdot \mathbf{O}^{cd}_{(0)}\right]$ and $\text{E}\left[\mathbf{O}^{ab}_{(2)} \cdot \mathbf{O}^{cd}_{(2sim)}\right]$ computable in small samples or to arbitrary precision through repeated randomizations.

6 Guaranteed conservative variance estimator

From first principles, the exact variance of a linear estimators of the form given in equation (1) is,

$$
\begin{aligned}
\text{V}\left(\widehat{\delta}_c\right) &= \text{E}\left[y'\mathbf{RW}'cc'\mathbf{WR}y\right] - \text{E}\left[y'\mathbf{RW}'c\right]\text{E}\left[c'\mathbf{WR}y\right] \\
&= y'\left\{ \text{E}\left[\mathbf{RW}'cc'\mathbf{WR}\right] - \text{E}\left[\mathbf{RW}'c\right]\text{E}\left[c'\mathbf{WR}\right] \right\}y \\
&= y'\mathbf{g}y
\end{aligned}
$$

where $\mathbf{g}$ is defined as the matrix inside the curly brackets in the line above. The matrix, $\mathbf{g}$, may be computed exactly for small studies or to arbitrary precision by repeating the randomization until desired precision is achieved.

As before, some terms in this quadratic are not observed, and a suitable bound can be found generalizing one of the methods from Paper 1 of 4 (Neyman, Aronow-Samii or Middleton). Algorithm 4.7 from Paper 1 of 4 will be slightly modified and applied to the simulations below.

With Conjecture

7 Data from Paluck and Green (2009)

Paluck and Green (2009) pair-randomized 14 villages (7 pairs) in post-genocide Rwanda to receive one of two possible radio programs. Half of the villages were exposed to a program aimed at "discouraging blind obedience and reliance on direction from authorities and promoting independent thought and collective action in problem solving" (treatment). The other half were exposed to a radio program about health (control).

The analysis of the properties of the OC and GS variance estimators are compared in the spirit of (permutation/rerandomization-based) simulation. The analysis is simulation in the sense that it makes untestable assumptions about the outcomes that would have been observed under alternative assignments (e.g., the values of missing potential outcomes). The appeal of using real data, rather than generating data from a contrived DGP, however, is that it may provide a more realistic data set and the analysis can reflect the design actually used.

7.1 Randomization method

The original paired-cluster design was simulated and the estimator is OLS as in the original paper. With 7 pairs this results in $2^7 = 128$ possible randomizations.

7.2 Data

Analysis used the same 497 cases analyzed in model 1 of Table 4 in Paluck and Green (2009). Number of units by village pair and assignment arm are given in Table 1. The average cluster size was 35.5. The minimum and maximum cluster sizes were 20 and 43, respectively.

Five cases had missing values for *age*, which were mean-imputed for this analysis so that the same units to be used in both specifications. By contrast, the original analysis dropped these five cases when *age* was included in the OLS.

Table 1: Number of units by village pair and original assignment

	Village Pair						
	1	2	3	4	5	6	7
Control	37	39	39	39	33	37	43
Treatment	33	37	36	37	38	20	29

7.3 Outcome measure

The outcome is a post-treatment measure of social distance. The original effect estimates using OLS were presented in Table 4 of Paluck and Green (2009). The social distance measure was an index created by combining four survey items, each of which could take on integer values of 1,2,3 or 4. The resulting measure, an average of the items, also ranged from 1 to 4, but could take on 13 possible values.

Additionally the outcome, y, was location shifted by subtracting the mid-point on the scale, $(\max(y)\text{-}\min(y))/2$, such that the scale ranged from -1.5 to 1.5 rather than 1 to 4. One motivation for this is that non-invariant estimators such as OCJ and GCJ can, themselves, have high variance when the mean of y is shifted away from zero.

Another motivation is that when scale is bounded, and thus $\max(y)$ and $\min(y)$ are known, shifting by the scale midpoint introduces an invariance property. For example, subtracting off the scale midpoint results in the same observed values whether a researcher records the original outcomes on a scale of integers $\{0, 1, 2, 3\}$ or $\{1, 2, 3, 4\}$.

Another option might be to subtract the observed sample mean from each outcome. However, this leads to a shift that may depend on the randomization, when there are treatment effects, and so the practice can introduce biases.

7.4 Variance estimators compared

The Generalized Neyman bounding matrix, given in section 4.1 of Paper 1 of 4 (Middleton, 2020), $\tilde{\mathbf{d}}^{\text{N}}$, is not applicable in this case because of unobservable cells in main block diagonal of the design matrix, $\mathbf{d}$. Therefore simulations use the bounding matrix, $\tilde{\mathbf{d}}^{\text{M}}$, which is based on Algorithm 4.7 in Paper 1 of 4. It provides a tighter bound than the Aronow-Samii bounding matrix because clusters were pair-randomized, a result that is similar to the one given for pair-randomization of units in the example provided in section 4.4 of that paper.

The variance estimators are based upon the Generalized Sandwich (GS) given in equation (6) in Section 4, the Obložené Chlebìžky-based, given in Section **??** ($J = 10$), and the Guaranteed Conservative, given in Section 6 ($J = 10$).

7.5 Covariate specifications

The OC and GS variance estimators were examined with and without covariates. The "no covariates" OLS included only intercepts, so that $x = 1$. This is equivalently the difference-of-means. The second specification included the additional covariates in column 3 of Table 4 in the original: *displaced by violence*, *sex*, *age*, and *radio lisening habits*.

For both of these specifications, classic Cluster Robust (CR) variance estimators also included fixed-effects for cluster-pairs as in columns 2 and 3 of Table 4 of Paluck and Green (2009). The typical motivation for including the fixed effects for blocks/pairs is to obtain standard errors that reflect this design feature. These are unnecessary in the case of OC and GC variance estimators because the paired-cluster design is already accounted for in the design matrix, **d**.

7.6 Simulation assumptions

The assumptions of NCM apply, for example, assuming that potential outcomes are fixed (non-stochastic) and depend only on assignment. Also the sharp null, $y_{0i} = y_{1i}$ for all i, is invoked in order to impute missing potential outcomes.[1] The simulation also proceeds as if the survey response is not stochastic and whether a unit participates in the survey is not dependent on treatment assignments.[2]

7.7 Simulation results

Tables 3 and 2 give simulation results for OLS with and without covariates, respectively, using the data originally reported in Paluck and Green (2009). Comparing rMSE of variance estimators:

Comparisons:

- Generalized Sandwich is less biased and has a smaller $\text{SE}[\widehat{\widetilde{V}}]$ compared to the classical Cluster Robust variance estimators

- OC estimators have smaller $\text{SE}[\widehat{\widetilde{V}}]$ compared to GS, while having he same expected value as GS by construction

- Only GCJ is unbiased for the variance of the OLS coefficient. It has slightly higher $\text{SE}[\widehat{\widetilde{V}}]$ than the GS, though the rMSE is lower because GS is biased.

---

[1]Simulating under the sharp null seems reasonable for two reasons. First, the estimated average effects (-.029 to -.041, depending on specification) are small relative to the range of the measure (1.0% to 1.4%, respectively) and also to its standard deviation (3.6% to 5.1%, respectively). Second, and more importantly, simulating under the sharp null (or constant treatment effects more generally) is useful because the variance and its (identified) bound will be equal. The implication is that any *downward* bias of the variance estimator that is attributable to borrowing the variance of the first-order approximation, and/or to using a "plug-in" estimator thereof, will be evident in simulation. By contrast, downward bias can be masked (i.e., offset) under heterogeneous effects because the *bound* will tend to be conservative in that case. So, simulating under the sharp null gives a "worst-case" assessment of anti-conservative bias.

[2]Simulating as if survey response is non-stochastic and not affected by treatment assignments also seems reasonable for two reasons. First, an assumption about the independence of assignment and response would also be required for consistency in the original analysis given by Paluck and Green (2009). Second, similar to the case of the sharp null, if survey response was independent of assignment but stochastic, then variance estimators would be conservativeness relative to the assumption of response being non-random. Again, with this assumption, simulation results give a "worst-case" assessment of anti-conservative bias.

Table 2: Comparing Variance Estimators for OLS, No Covariates

	Classical Cluster Robust			Gen. Sandwich	Obloženè Chlebìžky			Guaranteed Conservative		
	CR0	CR1	CR2	GS0	OC1	OCJ.B	OCJ	GC0.B	GCJ.B	GCJ
$\mathrm{E}[\widehat{\mathbb{V}}/\mathbb{V}]$	0.429	0.462	0.818	0.842	0.841	0.842	0.842	0.858	0.860	1.000
$\mathrm{SE}[\widehat{\mathbb{V}}/\mathbb{V}]$	0.097	0.104	0.167	0.144	0.123	0.123	0.123	0.128	0.129	0.132
$\mathrm{Bias}[\widehat{\mathbb{V}}/\mathbb{V}]$	-0.571	-0.538	-0.182	-0.158	-0.159	-0.158	-0.158	-0.142	-0.140	0.000
$\mathrm{rMSE}[\widehat{\mathbb{V}}/\mathbb{V}]$	0.579	0.548	0.247	0.214	0.201	0.200	0.200	0.191	0.190	0.132
$\mathrm{CV}[\widehat{\mathbb{V}}]$	0.225	0.225	0.204	0.171	0.146	0.147	0.146	0.149	0.150	0.132
rMSE ratio:										
vs. GS0	2.712	2.566	1.155	1.000						
vs. OCJ	2.892	2.737	1.232	1.066	1.005	1.000	1.000			
vs. GCJ	4.392	4.156	1.870	1.619	1.526	1.519	1.519	1.451	1.442	1.000
CV ratio:										
vs. GS0	1.317	1.317	1.191	1.000						
vs. OCJ	1.537	1.537	1.389	1.167	0.999	1.000	1.000			
vs. GCJ	1.707	1.707	1.543	1.296	1.110	1.111	1.111	1.132	1.134	1.000

References

Aronow, Peter M. and Cyrus Samii. 2012. Conservative variance estimation for sampling designs with zero pairwise inclusion probabilities. *Survey Methodology* **39**(1): 231-241.

Aronow, Peter M. and Cyrus Samii. 2017. Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics* .

Aronow, Peter M. and Joel A. Middleton. 2015. A class of unbiased estimators of average treatment effect in randomized experiments. *Journal of Causal Inference* **1**(1): 135-154.

Holland, P.W. 1986. Statistics and Causal Inference, *Journal of the American Statistical Association*, vol. 81, no. 396: 945-968.

Horvitz, D.G. and Thompson, D.J. 1952. A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47**: 663-684.

Lin, Winston. 2013. Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique. *Annals of Applied Statistics* **7**(1): 295-318

Middleton, Joel A. 2018. A unified theory of regression adjustment for design-based inference. *arXiv:1803.06011*

Middleton, Joel A. 2020. Unifying Design-based Inference: On bounding and estimating the variance of any linear estimator in any experimental design. *Working paper at* www.joelmiddleton.com.

Middleton, Joel A. and Peter M. Aronow. 2015. Unbiased Estimation of the Average Treatment Effect in Cluster-Randomized Experiments. *Statistics, Politics and Policy* **1**:

Table 3: Comparing Variance Estimators for OLS, With Covariates

	Classic Cluster Robust			Gen. Sandwich	Obloženè Chlebìžky			Guaranteed Conservative		
	CR0	CR1	CR2	GS0	OC1	OCJ.B	OCJ	GC0.B	GCJ.B	GCJ
$E[\widehat{V}/V]$	0.431	0.464	0.819	0.802	0.797	0.802	0.802	0.853	0.861	1.000
$SE[\widehat{V}/V]$	0.113	0.121	0.187	0.131	0.094	0.095	0.095	0.118	0.121	0.125
$Bias[\widehat{V}/V]$	-0.569	-0.536	-0.181	-0.198	-0.203	-0.198	-0.198	-0.147	-0.139	0.000
$rMSE[\widehat{V}/V]$	0.580	0.550	0.260	0.237	0.224	0.220	0.220	0.189	0.184	0.125
$CV[\widehat{V}]$	0.261	0.261	0.228	0.163	0.118	0.119	0.118	0.138	0.140	0.125
rMSE ratio:										
vs. GS0	2.447	2.318	1.095	1.000						
vs. OCJ	2.640	2.500	1.181	1.079	1.019	1.000	1.000			
vs. GCJ	4.638	4.393	2.075	1.895	1.790	1.758	1.757	1.506	1.472	1.000
CV ratio:										
vs. GS0	1.606	1.606	1.400	1.000						
vs. OCJ	2.207	2.207	1.923	1.374	0.998	1.003	1.000			
vs. GCJ	2.089	2.089	1.820	1.300	0.944	0.949	0.946	1.106	1.121	1.000

Neyman, Jerzy Splawa, D. M. Dabrowska, and T. P. Speed. [1923.] 1990. On the application of probability theory to agricultural experiments: Essay on principles, section 9. *Statistical Science* **5**: 465-480.

Paluck, Elizabeth Levy, and Donald P. Green. 2009. Deference, Dissent, and Dispute Resolution: An Experimental Intervention Using Mass Media to Change Norms and Behavior in Rawanda. *American Political Science Review* **103**(4): 622-644.