The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation

Kosuke Imai
Princeton University
Gary King **Clayton Nall**
Harvard University

July 19, 2007

Cluster-Randomized Experiments (CREs)

- Problem of many field experiments:
  - unit of randomization = clusters of individuals
  - unit of interest = individuals

Gosnell (1927)	city blocks	individuals
Gerber & Green (2000)	households	individuals
Wantchekon (2003)	villages	individuals
Arceneaux (2005)	precincts	individuals
Guan & Green (2006)	dorm rooms	individuals

- CREs among political science field experiments: 68% (out of 28)
- Public health & medicine: CREs have "risen exponentially since 1997" (Campbell, 2004)
- Economics (firms – products)
- Education (classrooms – students)
- Psychology (groups – individuals)
- Sociology (neighborhoods – households)

Design and Analysis of CREs

- Cluster randomization → loss of efficiency & specialized methods
- Prop. of polisci CREs which completely ignore the design: ≈ 50%
- Prop. of polisci CREs which use *design-based* analysis: 0%
- Prop. of polisci CREs which make more assumptions than necessary: 100%

- Matched-Pair Designs (MPDs) to improve efficiency:
  1. Pair clusters based on the similarity of background characteristics
  2. Within each pair, randomly assign one cluster to the treatment group and the other to the control group

- Use of MPDs in CREs:
  - Prop. of public health CREs: ≈ 50% (Varnell *et al.*, 2004)
  - Prop. of polisci CREs: 0%

Methodological Recommendations Against MPDs

- "Analytical limitations" of MPDs (Klar and Donner, 1997):
  1. restriction of prediction models to cluster-level baseline risk factors
  2. inability to test for homogeneity of causal effects across clusters
  3. difficulties in estimating the intracluster correlation coefficient
- In 10 or fewer pairs, MPDs can lose power (Martin *et al.* 1993)
- Echoed by other researchers and clinical standard organizations
- These claims are all unfounded!

- No formal definition of causal effects to be estimated
- No formal evaluation of the existing estimators for MPDs

Contributions of Our Paper

- Conclusion: **pair-matching should be used whenever feasible**
  - MPDs improve bias, efficiency, and power
  - Not pairing = throwing away one's data!
- Show that "analytical limitations" do not exist or are irrelevant
- Show that power calculations rely on unrealistic assumptions
- Existing estimator is based on a highly restrictive model
- Formally define causal quantities of interest
- Propose new simple design-based estimators and s.e.'s
- Offer power and sample size calculations
- Extend the estimator to CREs with unit-level noncompliance
- Clarify the assumptions about interference

Running Example: Seguro Popular de Salud (SPS)

- Evaluation of the Mexican universal health insurance program
- Aim: "provide social protection in health to the 50 million uninsured Mexicans" (Frenk *et al.*, 2003)
- A key goal: reduce out-of-pocket health expenditures
- Sounds obvious but not easy to achieve in developing countries
- Individuals must affiliate in order to receive SPS services
- $12,824$ "health clusters"
- 100 clusters nonrandomly chosen for randomized evaluation
- Pairing based on population, socio-demographics, poverty, education, health infrastructure etc. (King *et al.*, 2007)
- "Treatment clusters": encouragement for people to affiliate
- Data: aggregate characteristics, surveys of $32,000$ individuals

Causal Quantities of Interest

Quantities		Clusters	Units within Clusters	Inferential Target
ψ_S	SATE	Observed	Observed	Observed sample
ψ_C	CATE	Observed	Sampled	Population within observed clusters
ψ_U	UATE	Sampled	Observed	Observable units within pop. of clusters
ψ_P	PATE	Sampled	Sampled	Population

- Sample Average Treatment Effect (SATE):
  $\psi_S \equiv \mathbb{E}_{\mathcal{S}}(Y(1) - Y(0)) = \frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{2}\sum_{i=1}^{n_{jk}}(Y_{ijk}(1) - Y_{ijk}(0))$
- Cluster Average Treatment Effect (CATE):
  $\psi_C \equiv \mathbb{E}_{\mathcal{C}}(Y(1) - (0)) = \frac{1}{N}\sum_{k=1}^{m}\sum_{j=1}^{2}\sum_{i=1}^{N_{jk}}(Y_{ijk}(1) - Y_{ijk}(0))$
- Unit Average Treatment Effect (UATE): $\psi_U \equiv \mathbb{E}_{\mathcal{U}}(Y(1) - Y(0))$
- Population Average Treatment Effect (PATE): $\psi_P \equiv \mathbb{E}_{\mathcal{P}}(Y(1) - Y(0))$

Interference in CREs under MPDs

- What is *interference*?: one's (potential) outcome depends on treatment assignment of others as well as her own
- Disease contagion, social pressure, help from families and friends

1. Among individuals in the same cluster
2. Between clusters in different pairs
3. Between treatment and control clusters in the same pair

- (1) is allowed in CREs as a consequence of treatment
- (1) is not allowed in individual randomized trials
- (2) is not allowed in CREs under MPDs
- (3) is allowed:
  - *with-interference* causal effects
  - *no-interference* causal effects

Design-based Analysis of CREs under MPDs

- Existing Model-based approach: assume DGP for observed data
- Randomness comes from the assumed model
- If the model is correct, inference is valid
- If the model is incorrect, inference is invalid

- Our Design-based approach (Fisher and Neyman)
- Randomness comes from:
  1. randomization of treatment assignment
  2. random sampling of clusters and units within clusters
- Avoids modeling assumptions

Definition of Estimators

- "A good estimator for one ATE is automatically a good estimator for the other" (Imbens, 2004)
- Does not apply to CREs
- Our estimator:

$$\hat{\psi}(w_k) \equiv \frac{1}{\sum_{k=1}^{m} w_k} \sum_{k=1}^{m} w_k \left\{ Z_k \left(\frac{\sum_{i=1}^{n_{1k}} Y_{i1k}}{n_{1k}} - \frac{\sum_{i=1}^{n_{2k}} Y_{i2k}}{n_{2k}} \right) + (1 - Z_k) \left(\frac{\sum_{i=1}^{n_{2k}} Y_{i2k}}{n_{2k}} - \frac{\sum_{i=1}^{n_{1k}} Y_{i1k}}{n_{1k}} \right) \right\}$$

	SATE	CATE	UATE	PATE
Point estimator	$\hat{\psi}(n_{1k} + n_{2k})$	$\hat{\psi}(N_{1k} + N_{2k})$	$\hat{\psi}(n_{1k} + n_{2k})$	$\hat{\psi}(N_{1k} + N_{2k})$
Variance	$\mathrm{Var}_a(\hat{\psi})$	$\mathrm{Var}_{au}(\hat{\psi})$	$\mathrm{Var}_{ap}(\hat{\psi})$	$\mathrm{Var}_{aup}(\hat{\psi})$
Identified	no	no	YES	YES

Bias

- Bias expression for SATE ($E_a\{\hat{\psi}(n_{1k}+n_{2k})\}-\psi_S$):

$$\frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{2}\left\{\left(\frac{n_{1k}+n_{2k}}{2}-n_{jk}\right)\sum_{i=1}^{n_{jk}}\frac{Y_{ijk}(1)-Y_{ijk}(0)}{n_{jk}}\right\}$$

- Conditions for unbiasedness:
  1. Exact match on sample cluster sizes: $n_{1k}=n_{2k}$ for all k
  2. Exact match on within-cluster SATEs: $\sum_{i=1}^{n_{1k}}(Y_{i1k}(1)-Y_{i1k}(0))/n_{1k}=\sum_{i=1}^{n_{2k}}(Y_{i2k}(1)-Y_{i2k}(0))/n_{2k}$ for all k
- Match on cluster sizes and important covariates!
- Bias for CATE ($E_{au}(\hat{\psi}(N_{1k}+N_{2k}))-\psi_C$):

$$\frac{1}{N}\sum_{k=1}^{m}\sum_{j=1}^{2}\left\{\left(\frac{N_{1k}+N_{2k}}{2}-N_{jk}\right)E_u(Y_{ijk}(1)-Y_{ijk}(0))\right\}$$

- Additional condition for UATE & PATE: cluster sizes $\perp\!\!\!\perp$ ATEs

Existing Estimator

- Estimator based on harmonic mean weights and associated variance estimator (Donner, 1987): $w_k=n_{1k}n_{2k}/(n_{1k}+n_{2k})$
- No formal justification in the literature (weighted one-sample t-test)
- Assumed unrealistic unit-level model: for $t=0,1$,

$$Y_{ijk}(t)\overset{\text{i.i.d.}}{\sim}N(\mu_t,\sigma)$$

1. Normality
2. I.I.D. across units within each cluster, and across clusters & pairs
3. Equal variances for potential outcomes

- Under the model, the estimator is UMVUE
- The model assumes there is no point of matching to begin with!
- Unless these assumptions are met, the estimator is invalid

Variance Identification and Estimation

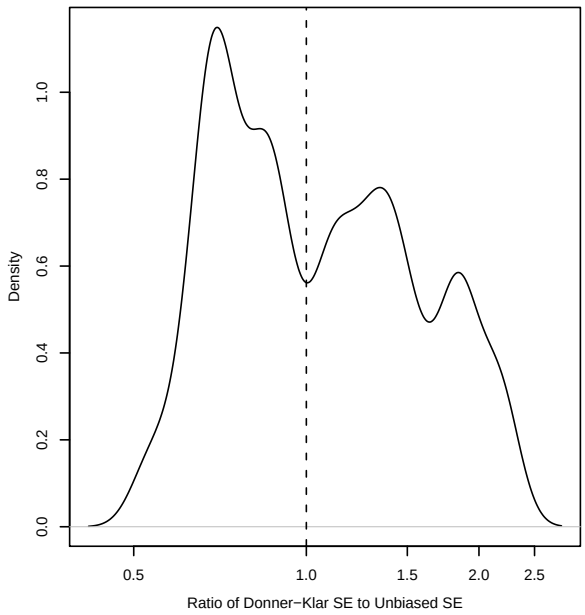
- Our general unbiased variance estimator for $\hat{\psi}(\tilde{w}_k)$:

$$\hat{\sigma}(\tilde{w}_k) \equiv \frac{m}{(m-1)n^2}\sum_{k=1}^{m}\left[\tilde{w}_k\left\{Z_k\left(\frac{\sum_{i=1}^{n_{1k}} Y_{i1k}}{n_{1k}} - \frac{\sum_{i=1}^{n_{2k}} Y_{i2k}}{n_{2k}}\right) + (1-Z_k)\left(\frac{\sum_{i=1}^{n_{2k}} Y_{i2k}}{n_{2k}} - \frac{\sum_{i=1}^{n_{1k}} Y_{i1k}}{n_{1k}}\right)\right\} - \frac{n\hat{\psi}(\tilde{w}_k)}{m}\right]^2$$

  where $\tilde{w}_k$ is the normalized weights, $\tilde{w}_k \equiv n w_k / \sum_{k=1}^{m} w_k$
- $E_a(\hat{\sigma}(\tilde{w}_k))$ is the sharp upper bound of SATE variance
- $E_{au}(\hat{\sigma}(\tilde{w}_k))$ is the sharp upper bound of CATE variance
- $E_{ap}(\hat{\sigma}(\tilde{w}_k))$ is UATE variance
- $E_{apu}(\hat{\sigma}(\tilde{w}_k))$ is PATE variance

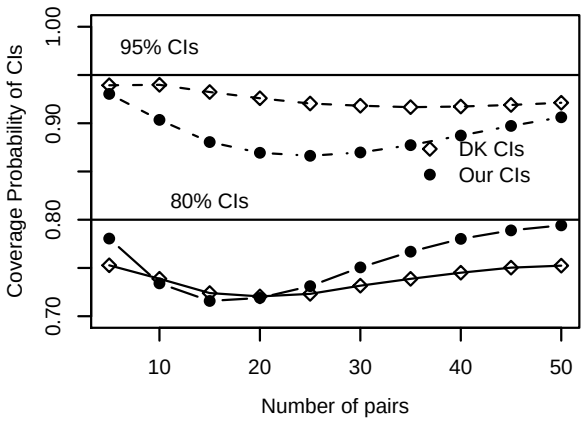
Illustration using SPS Data

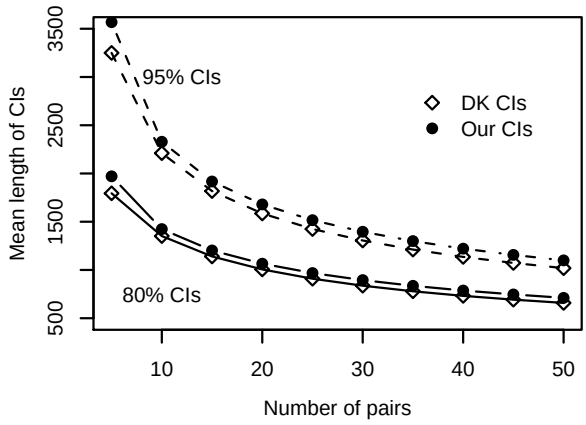
- The direction of bias for DK's s.e. is indeterminate: from 3 times larger to 3 times smaller.



Monte Carlo Evidence

- Setup:
  - Use population cluster sizes
  - Out-of-pocket health expenditure variable (peso)
  - Use cluster-specific sample mean and variances as truth
- CATE: ours (bias=0, RMSE=6), DK (bias=21, RMSE=22)
- PATE: confidence interval comparison





Relative Efficiency of MPDs

- Compare with Completely-Randomized Designs (CRDs)
- Relative efficiency of MPDs over CRDs:

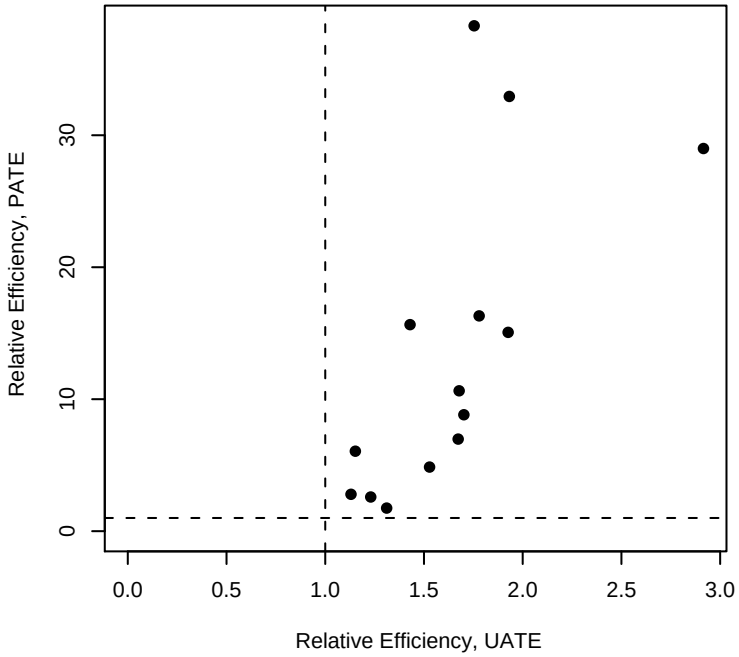
$$\frac{\mathrm{Var}_{ac}(\hat{\tau}(\tilde{w}_j))}{\mathrm{Var}_{ap}(\hat{\psi}(\tilde{w}_k))} = \left\{1 - \frac{2\mathrm{Cov}_p(\tilde{w}_k \overline{Y_{jk}(1)}, \tilde{w}_k \overline{Y_{j'k}(0)})}{\sum_{t=0}^{1} \mathrm{Var}_p(\tilde{w}_k \overline{Y_{jk}(t)})}\right\}^{-1}$$

- Greater (positive) correlation within pair $\rightarrow$ greater efficiency

- MPDs vs. Stratified Designs (CRDs within pre-defined strata)
- MPDs can improve efficiency within strata

Illustration Using SPS Data

- UATE: MPDs are between 1.1 and 2.9 times more efficient
- PATE: MPDs are between 1.8 and 38.3 times more efficient!



Power and Sample Size Calculations under MPDs

- Statistical power: prob. of rejecting the null when the null is false
- Assume equal cluster size for planning purposes
- UATE ($H_0 : \psi_U = 0$ and $H_A : \psi_U = \psi$):

$$1 + \mathcal{T}_{m-1}(-t_{m-1,\alpha/2} \mid d_U\sqrt{m}) - \mathcal{T}_{m-1}(t_{m-1,\alpha/2} \mid d_U\sqrt{m}),$$

  where $d_U \equiv \psi/\sqrt{\mathrm{Var}(D_k)}$.
- PATE ($H_0 : \psi_P = 0$ and $H_A : \psi_U = \psi$):

$$1+\mathcal{T}_{m-1}\left(-t_{m-1,\alpha/2} \,\middle|\, \frac{d_P\sqrt{m}}{\sqrt{1+\pi/\bar{n}}}\right) - \mathcal{T}_{m-1}\left(t_{m-1,\alpha/2} \,\middle|\, \frac{d_P\sqrt{m}}{\sqrt{1+\pi/\bar{n}}}\right)$$

  where $d_P \equiv \psi/\sqrt{\mathrm{Var}_p\{E_u(D_k)\}}$ and π is the ratio of between-cluster and within-cluster variances.
- Sample size calculation: what sample size do I need in order to achieve a certain level of power under a particular H_A?

Relative Power of MPDs

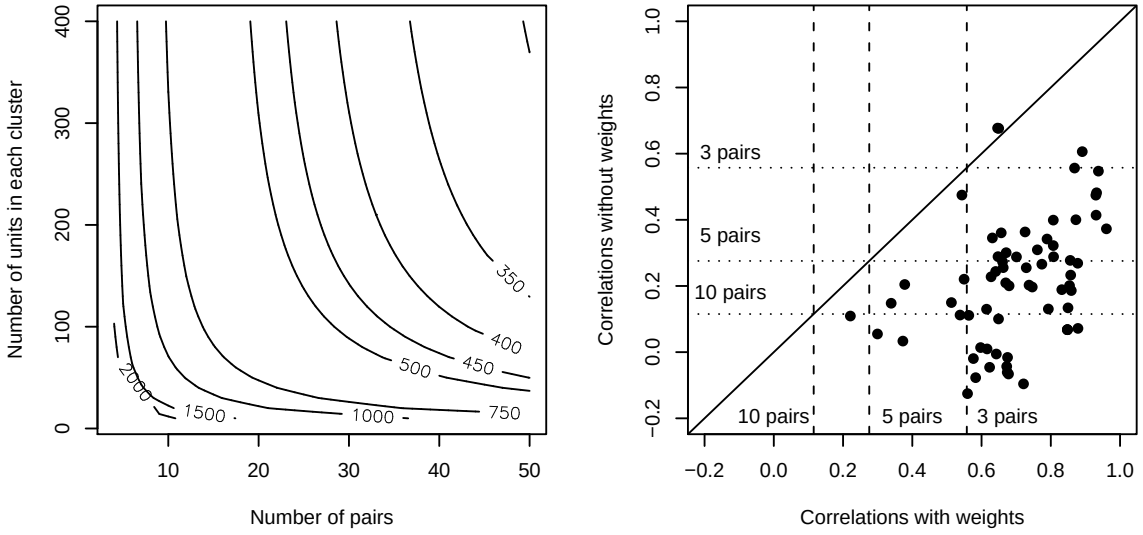
- When the number of pairs is fewer than 10, "the matched design will probably have less power than the unmatched design due to the loss of degrees of freedom" (Martin *et al.* 1993).
- Critical assumption: equal cluster sizes across all clusters
- In typical CREs, cluster sizes are different and observed
- Can match on cluster sizes:

$$\text{Corr}_p(\tilde{w}_k \overline{Y_{jk}(1)}, \tilde{w}_k \overline{Y_{j'k}(0)}) \geq \text{Corr}_p(\overline{Y_{jk}(1)}, \overline{Y_{j'k}(0)})$$

- Efficiency gain of MPDs is greater in CREs than in individual randomized experiments
- Thus, power of MPDs is also greater

Illustration Using SPS Data

- power=0.8 and size=0.95
- Sample size calculation using out-of-pocket health care expenditure
- Comparison of within-pair correlations with and without weights



Unit-Level Noncompliance in CREs

- No interference *between units* within (and across) clusters
  1. one's decision to comply doesn't depend on others' treatment assignment
  2. one's potential outcomes don't depend on others' treatment assignment and receipt
- Always-takers, compliers, and never-takers (Angrist *et al.* 1996)
- In SPS evaluation, the wealthy are never-takers (56%)
- Always-takers are those who travel and sign up for SPS (7%)
- No defier (monotonicity)
- Zero ITT effect on non-compliers (exclusion restriction)
- QoI: Complier Average Causal Effect or CACE (for SATE, CATE, UATE or PATE)
- We offer a consistent estimator and its valid s.e.

Empirical Analysis of SPS Data

- Average causal effects of SPS on the prob. of a household suffering from catastrophic health expenditures
- More than 30% of annual post-subsistence income (10% of all households)
- Its reduction is a major aim of SPS
- Predictions based on cluster-level baseline risk are straightforward
- Testing homogeneity of causal effects across pairs is also easy
- Loss of a cluster in follow-up results in loss of only one pair

		SATE		CATE		UATE		PATE	
All	ITT	$-.014$	$(\leq .007)$	$-.023$	$(\leq .015)$	$-.014$	$(.007)$	$-.023$	$(.015)$
	CACE	$-.038$	$(\leq .018)$	$-.064$	$(\leq .024)$	$-.038$	$(.018)$	$-.064$	$(.024)$
Male-	ITT	$-.016$	$(\leq .008)$	$-.025$	$(\leq .018)$	$-.016$	$(.008)$	$-.025$	$(.018)$
Headed	CACE	$-.042$	$(\leq .020)$	$-.070$	$(\leq .031)$	$-.042$	$(.020)$	$-.070$	$(.031)$

Concluding Remarks

- Field experiments often require cluster randomization
- Our recommendations: MPDs for CREs
  1. Select quantities of interest
  2. Identify pre-treatment covariates for matching
  3. Pair clusters based on the covariates and cluster sizes
  4. Randomize treatment within each pair
  5. Use design-based methods to analyze the data
- MPDs are preferred from perspectives of bias, efficiency, & power
- May affect CONSORT, Cochrane Collaboration, Council guidelines, etc.
- Our proposed estimators are design-based and avoid modeling assumptions
- Simple and require no simulation or numerical optimization
- R package **experiment** available at CRAN

Definition and Notation of MPDs

- Observed clusters: $2m$
- Number of pairs: m
- Number of observed units within the jth cluster in the kth pair: n_{jk}
- Population size of cluster: N_{jk}
- Total number of observed units: $n = \sum_{k=1}^{m}(n_{1k} + n_{2k})$
- Two clusters within each pair are randomly ordered
- Simple randomization of an indicator variable: Z_k
- $Z_k = 1$ ($Z_k = 0$): first (second) cluster gets treated
- Treatment variables: $T_{1k} = Z_k$ and $T_{2k} = 1 - Z_k$
- Potential outcomes for each individual: $Y_{ijk}(T_{jk})$
- Observed outcome: $Y_{ijk} = T_{jk}Y_{ijk}(1) + (1 - T_{jk})Y_{ijk}(0)$
- Cluster randomization: $(Y_{ijk}(1), Y_{ijk}(0)) \perp\!\!\!\perp Z_k$
- For now, consider the intention-to-treat (ITT) analysis

Alternative Estimators

- Unbiased estimator for SATE & UATE (but not for CATE & PATE)
- Problem: not invariant to constant shift
- Variance estimator is also not invariant

- Invariant Estimator with smaller bias
- Exact calculation of variance is impossible
- Standard variance estimator is not invariant

Inference under MPDs

- Many pairs:
  - No additional assumption: central limit theorem
  - $(1-\alpha)$ CI: $[\hat{\psi}(\tilde{w}_k) - z_{\alpha/2}\sqrt{\hat{\sigma}(\tilde{w}_k)},\ \hat{\psi}(\tilde{w}_k) + z_{\alpha/2}\sqrt{\hat{\sigma}(\tilde{w}_k)}]$
- Few pairs, many units:
  - CATE: $\tilde{w}_k D_k$ *is* normally distributed
  - SATE, UATE, & PATE: $\tilde{w}_k D_k$ is *assumed* to be normally distributed
  - $(1-\alpha)$ CI: $[\hat{\psi}(\tilde{w}_k) - t_{m-1,\alpha/2}\sqrt{\hat{\sigma}(\tilde{w}_k)},\ \hat{\psi}(\tilde{w}_k) + t_{m-1,\alpha/2}\sqrt{\hat{\sigma}(\tilde{w}_k)}]$
- Few pairs, few units:
  - For all quantities: $\tilde{w}_k D_k$ is *assumed* to be normally distributed
- No "Behrens-Fisher" problem unlike CREs under completely-randomized designs
- Irrelevance of intracluster correlation coefficient (ICC): "an estimate of ρ [ICC] is required to compute appropriate standard errors for the analyses in question" (Donner 1998).