Adaptive Design of Clustered Experiments

Carlos Gonzalez-Perez, University of Oxford

Econometrics & Development Reading Group

April 22, 2026

Outline

Outline

Motivation

- Surveys and experiments have become **larger**
  - Not just by adding more units, but by spanning **larger geographical regions**
  - Either physically (provinces, schools) or figuratively (markets, industries)

- **Why?** [Muralidharan and Niehaus, 2017]
  - **External Validity.** We want our results to be representative of the population of interest
  - Get better sense of implementation dynamics
  - Governmental support

Motivation

- Ideally, draw units directly from the population of interest, but
  - **Logistically infeasible** and **very costly**
  - Imagine a researcher going all around the country interviewing people
- **Solution?**
  - Run **multistage/stratified** surveys and experiments [Sedgwick, 2015], [Barcaroli et al., 2022]
  - **Structure**
    - Split the population of interest into *independent* clusters
    - Randomly select a number of clusters K
    - Randomly select a set of units N from each of the K clusters
    - Tower the process if needed

Motivation

- **The more clusters, the better?**
  - Even if clusters are selected at random from the population, observations may be highly **correlated within clusters**, so we may **reduce the variance** of our estimators by increasing the number of clusters
  - **Clusters are costly**: partnership with local authorities, travel expenses, additional staff, location specific infrastructure...
- When there is a hard **budget**, a clear **trade-off** emerges
  - **clusters** (uncorrelated but expensive) vs **units** (correlated but cheap)
  - The trade-off is mediated by the relative **within-cluster correlation**
- **Research Question**
  - **Can we design optimal multistage sampling experiments under unknown within-cluster correlation?**

Some Examples

- **Development Economics**
  - [Mbiti et al., 2019] (QJE) randomly sample 10 districts from mainland Tanzania, and then randomly allocate treatment (combination of school grants and teacher incentives) to 35 schools within each district
  - Others: [Karing, 2024] (QJE), [Heß et al., 2021] (REStud)
- **Survey Design** (prevalent approach)
  - [Sedgwick, 2015] Survey on British public opinion on the use of personal medical data. Randomly select postcodes from the UK (proportionally to size), and then randomly select households from postcodes
  - Others: [Osei and Zhuang, 2020], [Lin and Ho, 2003]
- **Stratification** seems **prevalent** in RCT and survey designs, but **"no clear guidelines"** driving the decision process

Common Problems in Cluster Selection

- Selection seems based on *feel*, logistics and **common misconceptions**
  - UN Handbook of Household Sample Surveys in Developing and Transition Economies, [Tam, 2005]: Sample **as many clusters as possible**,
  - [Sedgwick, 2015]: **Fix number of clusters** in advance,
  - [Barcaroli et al., 2022]: **Guess** the correlation, do minmax or estimate it from previous studies
- **My Goal?** Provide **theoretical guarantees** for an **intuitive and easy to implement** (adaptive) design

Main Contributions

- **Framework** for optimal K vs N selection in stratified surveys and experiments under budget constraints and linear costs
- Characterize the **oracle policy** under known correlation
- Propose a **simple algorithm** under **unknown cluster correlation**
  - Show its excess variance is **negligible** compared to the variance of the oracle
  - Show its excess variance is **unimprovable** through matching lower bounds
  - Show its excess variance is **non-trivial** when compared to optimal static policies
- Algorithm uses a **two-stage approach**
  - **First stage** sequentially samples **clusters** (w few obs)
  - **Second stage** samples as many **units** as budget permits from explored clusters

Outline

Model

- Focus on **survey design** problem. Experimental designs on their simplest difference-in-means version are completely equivalent

- (Nested) **random effects** model

$$y_{ki} = \mu + \varepsilon_k + \varepsilon_{ki}, \qquad \mathbb{E}[\varepsilon_k] = \mathbb{E}[\varepsilon_{ki}] = 0, \qquad \mathbb{E}[\varepsilon_k^2] = \rho, \ \ \mathbb{E}[\varepsilon_{ki}^2] = \sigma$$

  - Model is parameterized by $\omega = (\mu, \rho, \lambda) \in \Omega$ with Ω compact, $\rho > 0$ where $\boldsymbol{\lambda = \sigma/\rho}$ is the **relative cluster correlation**
  - **cross-cluster independence**, i.e. $\text{CoVar}(y_{ki}, y_{lj}) = 0$,
  - homogeneous total variance, i.e. $\text{Var}(y_{ki}) = \rho(1 + \lambda)$, (*heterogeneity is coming*)
  - **within-cluster equicorrelation**, i.e. $\text{CoVar}(y_{ki}, y_{kj}) = \rho$

Designer's Problem

- **Estimator**

$$\hat{\mu}(K,N) = \frac{1}{KN}\sum_{k=1}^{K}\sum_{i=1}^{N} y_{ki}$$

- **Minimize $\hat{\mu}$ MSE subject to a budget constraint and linear costs**
  - $\hat{\mu}$ is unbiased under stratified randomization, so simply minimize variance

$$\min_{K\geq 1, N\geq 1} W(K,N,\omega)$$
$$\text{st } FK + VKN \leq B$$

- where
  - $W(K,N,\omega) = \text{Var}(\hat{\mu}(K,N))$
  - F is the **fixed cost** of sampling an additional cluster
  - V is the **variable cost** of sampling an additional unit
  - $B > F + V$ is the hard **budget**

Designer's Problem

- Solving from the FOC, we define the **optimal map** as

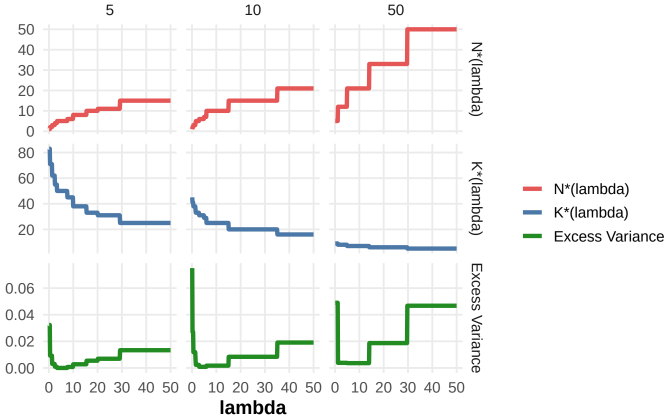
$$K^*(\omega) = \frac{B}{F + VN^*(\omega)}, \qquad N^*(\omega) = \min\left\{\max\left\{1, \sqrt{\lambda\frac{F}{V}}\right\}, \frac{B-F}{V}\right\}$$

- The optimal map depends on ω only via λ and it is very sensitive to λ mispecification

- For some $\omega = \omega_0$, the oracle variance

$$W(N^*(\lambda_0), K^*(\lambda_0), \omega_0) = \frac{\rho_0\left(\sqrt{V\lambda_0} + \sqrt{F}\right)^2}{B} \sim \boldsymbol{O(1/B)}$$

Comparative Statics of λ



$N^*(\lambda)$ (top), $K^*(\lambda)$ (middle) and excess variance (bottom) for different F/V ratios.
$B = 1000, V = 2,\ \omega = (\mu = 0, \rho = 1, \lambda = 5)$

Algorithm Heuristics

- For some estimate $\hat{\lambda}$, select $(K^*(\hat{\lambda}), N^*(\hat{\lambda}))$, but **how to learn** λ?
- $\lambda = \sigma/\rho$ can only be estimated under K **asymptotics** because ρ is an incidental parameter
  - If I had a *learning budget* $B_p < B$ to estimate λ, (in principle) I would set $N = N_p$ small and K as large as possible
  - Learning λ is only a means to an end, (in principle) I would never sample $K \geq K^*(\hat{\lambda})$ clusters

Adaptive Cluster Selection Algorithm (ACS)

Algorithm ACS

Input $B, F, V, N_p, f_\omega, \Omega$

Initialize $k = 0,\ \hat{\lambda}_0 = \lambda_m$

while $k + 1 \leq \min\left\{K^*(\hat{\lambda}_k), \left\lfloor \frac{B}{F+VN_p} \right\rfloor\right\}$

  explore an additional cluster and **observe** $(y_{ki})_{N_p}$

  update $k = k + 1$ and $\hat{\lambda}_k$ via MLE of f_ω

set $K = k$ and **recover** $\hat{\lambda}_{K_p - 1}$

sample $N = \max\{0, N^*(\hat{\lambda}_{K_p-1}) - N_p\}$

Outline

Stopping Time and Excess Variance

- For stopping time τ, $\texttt{ACS}$ approximately implements $(K^*(\hat{\lambda}_\tau), N^*(\hat{\lambda}_\tau))$
- The expected **excess variance** of our policy

$$\mathbb{E}_{\texttt{ACS}}\left[W(K^*(\hat{\lambda}_\tau), N^*(\hat{\lambda}_\tau), \omega_0)\right] - W(K^*(\lambda_0), N^*(\lambda_0), \omega_0) \approx$$
$$\frac{1}{2} W_{\lambda\lambda}(\lambda_0, \omega_0)\ \mathbb{E}_{\texttt{ACS}}\left[(\hat{\lambda}_\tau - \lambda)^2\right], \qquad W_{\lambda\lambda}(\lambda_0, \omega_0) \sim \boldsymbol{O(1/B)}$$

- How can we control for $\mathbb{E}_{\texttt{ACS}}\left[(\hat{\lambda}_\tau - \lambda)^2\right]$, when $\mathbb{E}_{\texttt{ACS}}$ is a (complicated) law?

Deviation Control

- Ideally, we would like to construct a sequence $(D_k(\alpha))_k$ st

$$\mathbb{P}_{\texttt{ACS}}(\mid \hat{\lambda}_\tau - \lambda \mid \leq D_\tau(\alpha)) \geq 1 - \alpha \tag{1}$$

- But it is not trivial to fine tune $D_k(\alpha)$ so delivers the desired coverage when integrated over $\mathbb{P}_{\texttt{ACS}}$, instead construct $(D_k(\alpha))_k$ such that

$$\mathbb{P}_{\texttt{ACS}}(\mid \hat{\lambda}_\tau - \lambda \mid \leq D_\tau(\alpha)) \geq \boldsymbol{\mathbb{P}(\mid \hat{\lambda}_k - \lambda \mid \leq D_k(\alpha)\ \forall k)} \geq 1 - \alpha \tag{2}$$

- Sadly, **frequentist CI do not satisfy** the **time uniformity** property

Anytime Confidence Sequence [Howard et al., 2021]

- Let $y_j = (y_{ij})_{N_p}$ and consider $f_\omega(y_j) \in C^3(\omega)$ single-peaked at $\hat{\omega}$

Anytime Confidence Sequence [Howard et al., 2021]

- Let $y_j = (y_{ij})_{N_p}$ and consider $f_\omega(y_j) \in \mathcal{C}^3(\omega)$ single-peaked at $\hat{\omega}$
- For smooth prior w full-support $\Pi(\omega)$, define the **LR/mixing martingale**

$$M_k(\omega) := \int \prod_{j=1}^{k} \frac{f_\theta(y_j)}{f_\omega(y_j)} \Pi(d\theta)$$

Anytime Confidence Sequence [Howard et al., 2021]

- Let $y_j = (y_{ij})_{N_p}$ and consider $f_\omega(y_j) \in C^3(\omega)$ single-peaked at $\hat{\omega}$
- For smooth prior w full-support $\Pi(\omega)$, define the **LR/mixing martingale**

$$M_k(\omega) := \int \prod_{j=1}^{k} \frac{f_\theta(y_j)}{f_\omega(y_j)} \Pi(d\theta)$$

- Consider the **sequence of confidence intervals**

$$C_k(\alpha)\{\omega : M_k(\omega) < 1/\alpha\} \qquad \Lambda_k(\alpha) = \{\lambda : \exists(\mu,\rho),\ (\mu,\rho,\lambda) \in C_k(\alpha)\}$$

$$D_k(\alpha) = \sup_{\lambda \in \Lambda_k(\alpha)} |\ \hat{\lambda}_k - \lambda\ |$$

Anytime Confidence Sequence

- **Martingale Interpretation**
  - Classical LR intuition. If the data looks more like the mixing (numerator) than it does for a candidate ω (denominator), then $M_k(\omega)$ grows (evidence against ω) and it is excluded from $\{\omega : M_k(\omega) < 1/\alpha\}$
- **Theorem 5.1. Anytime Confidence Sequences**

$$\mathbb{P}(\lambda \in \Lambda_k(\alpha) \ \ \forall k) \geq 1 - \alpha$$

Algorithm Analysis in 3 Steps

- Define the **good event** $\mathcal{E} = \{\lambda \in \Lambda_k(\alpha)\ \forall k\}$, $\mathbb{P}(\mathcal{E}) \geq 1 - \alpha$ (Thm 5.1)

Algorithm Analysis in 3 Steps

- Define the **good event** $\mathcal{E} = \{\lambda \in \Lambda_k(\alpha)\ \forall k\}$, $\mathbb{P}(\mathcal{E}) \geq 1 - \alpha$ (Thm 5.1)
- **Lemma 5.2. Algorithm Performance**

$$\Delta_{\texttt{ACS}} \mid \mathcal{E} \leq \frac{1}{2} W_{\lambda\lambda}(\lambda_0, \omega_0) \mathbb{E}_{\texttt{ACS}}[D_\tau(\alpha)^2 \mid \mathcal{E}\ ] + R_\tau$$

Algorithm Analysis in 3 Steps

- Define the **good event** $\mathcal{E} = \{\lambda \in \Lambda_k(\alpha)\ \forall k\}$, $\mathbb{P}(\mathcal{E}) \geq 1 - \alpha$ (Thm 5.1)
- **Lemma 5.2. Algorithm Performance**

$$\Delta_{\texttt{ACS}} \mid \mathcal{E} \leq \frac{1}{2} W_{\lambda\lambda}(\lambda_0, \omega_0) \mathbb{E}_{\texttt{ACS}}[D_\tau(\alpha)^2 \mid \mathcal{E}\ ] + R_\tau$$

- **Lemma 5.3. Upper Bound on** $D_k(\alpha)$

$$D_k(\alpha) \leq \sqrt{\frac{2\ln(1/\alpha) + 3\ln k + O(1)}{k\ q_k(\hat{\omega}_k)}}, \quad \text{for } k \text{ large}$$

Algorithm Analysis in 3 Steps

- Define the **good event** $\mathcal{E} = \{\lambda \in \Lambda_k(\alpha)\ \forall k\}$, $\mathbb{P}(\mathcal{E}) \geq 1 - \alpha$ (Thm 5.1)
- **Lemma 5.2. Algorithm Performance**

$$\Delta_{\texttt{ACS}} \mid \mathcal{E} \leq \frac{1}{2} W_{\lambda\lambda}(\lambda_0, \omega_0) \mathbb{E}_{\texttt{ACS}}[D_\tau(\alpha)^2 \mid \mathcal{E}\ ] + R_\tau$$

- **Lemma 5.3. Upper Bound on $D_k(\alpha)$**

$$D_k(\alpha) \leq \sqrt{\frac{2\ln(1/\alpha) + 3\ln k + O(1)}{k\ q_k(\hat{\omega}_k)}}, \quad \text{for } k \text{ large}$$

- **Lemma 5.4. Linear Stopping Time**

$$\mathbb{P}_{\texttt{ACS}}(\tau \in \Theta(B) \mid \hat{\lambda}_k \in [\lambda_m, \lambda_M]\ ) = 1$$

Upper Bound Theorem

Theorem 5.5. Upper Bound on ACS

Under the assumptions in Lemmas 5.2, 5.3 and 5.4, and $0 < q_k(\omega) < \infty$ for $\omega \in \Omega$, wp at least $1 - \alpha$, there exists a finite constant C such that

$$\Delta_{\texttt{ACS}} \lesssim C \cdot \frac{\ln(1/\alpha) + \ln B}{B^2}$$

- The excess variance of `ACS` is **asymptotically negligible** compared to the oracle variance $O(1/B) >> O(\ln B/B^2)$

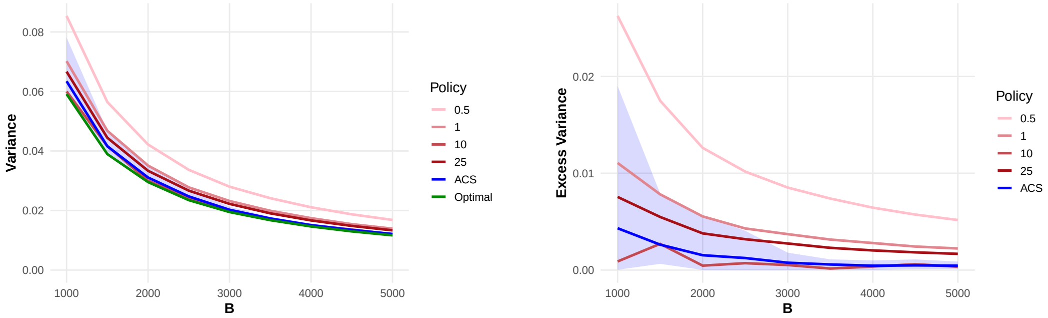
Outline

Simulations

Policy	K	N	Excess Variance	$Q_{0.05}$	$Q_{0.95}$
Oracle	55	8	0	-	-
ACS	**54.4**	**8.62**	**.0022**	**.000011**	**.0095**
Fixed .5	83	2	.013	-	-
Fixed 1	76	3	.0055	-	-
Fixed 10	50	10	.00046	-	-
Fixed 25	40	15	.0038	-	-

Table: Performance comparison across policies. $\omega(\mu = 0, \rho = 1, \lambda = 5)$, $B = 2000, F = 20, V = 2$. 1,000 replications.

Simulations: Performance as a Function of B



(a) Variance (b) Excess Variance

Figure: Variance (left) and Excess Variance (right) as a function of budget B across policies. $\omega = (0, 1, 5), F = 20, V = 2$. 1,000 replications.

Simulations: Design and Estimators

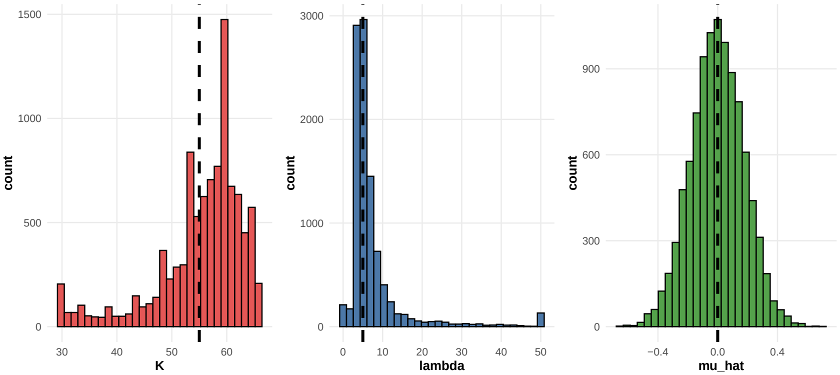


Figure: $K_{\texttt{ACS}}$ (left), $\hat{\lambda}_{\tau}$ (middle) and $\hat{\mu}(K_{\texttt{ACS}}, N_{\texttt{ACS}})$ (right), $\omega = (0, 1, 5), B = 2000, F = 20, V = 2$. Dashed line equals oracle values. 10,000 replications

Outline

Conclusion

- Proposed a **framework** for optimal K vs N selection in stratified surveys and experiments under budget constraints and linear costs
- Characterized the **oracle design** under known correlation
- Proposed an **optimal algorithm** with **unknown correlation**
  - Showed its excess variance is **negligible** compared to the variance of the oracle
  - Showed its excess variance is **unimprovable** through matching lower bounds
  - Showed its excess variance is **non-trivial** when compared to optimal static policies

Thank you!

carlos.gonzalezperez@economics.ox.ac.uk

carlosgonzalezperez@fas.harvard.edu

Outline

Intuition for Lower Bound

- Show that `ACS` excess variance is **essentially unimprovable** by deriving **matching lower bounds** on the problem up to logarithmic terms
- **Proof Structure**
  - Consider two DGPs $\overline{\omega}, \underline{\omega}$ such that $\overline{\omega} = (\mu, \overline{\rho}, \sigma/\overline{\rho})$, $\underline{\omega} = (\mu, \underline{\rho}, \sigma/\underline{\rho})$
  - where $\overline{\rho} = \rho + \delta$ and $\underline{\rho} = \rho - \delta$ with $\delta \sim O(B^{-1/2})$
  - δ is **small enough** such that it is difficult to distinguish across instances
  - δ is **large enough** such that misidentification is costly
  - Finally, argue that any policy which does well under $\overline{\omega}$ cannot do very well under $\underline{\omega}$ and viceversa

Graphical Representation of Lower Bound

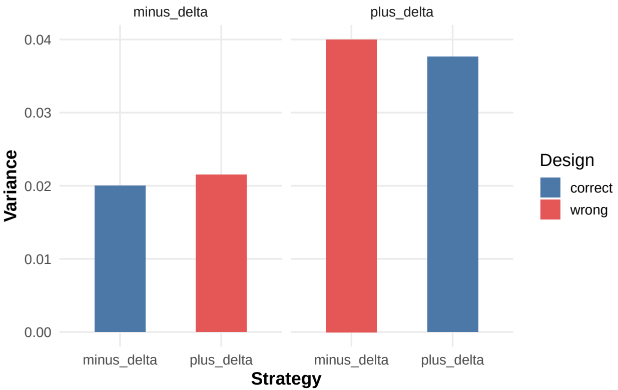


Figure: (True) Variance $W(K,N)$ under $\underline{\omega}$ (left) and $\overline{\omega}$ (right). $\rho = 1$, $\sigma = 5$, $\delta = 0.5$, $B = 2000$, $F = 20$, $V = 2$

Lower Bound Theorem

Theorem 6.1. Lower Bound on Excess Variance

For any policy π, such that $P_\pi\left(\sum_k^K (F + VN_k) \leq B\right) = 1$ there exists a DGP $\omega_0 \in \Omega$ such that

$$\Delta_\pi \gtrsim \frac{C_0\sqrt{\lambda_0 FV}}{B^2} \sim O(1/B^2)$$

- `ACS` is **nearly-optimal** up to logarithmic and constant terms

Policy Comparison

- The excess variance of `ACS` $O(\ln B/B^2)$ is **not trivial**
- **Lemma 6.2.**
  - Any static policy which takes a **mispecified** $\tilde{\lambda} = \lambda + b$,
  - and the **minmax optimal** policy over $\lambda \in [\lambda_m, \lambda_M]$
  - accrue an excess variance of $\boldsymbol{O(1/B)} >> O(\ln B/B^2)$

Outline

Literature Review in Survey Design

- Large literature on cluster vs unit selection in **survey design** [Biemer and Lyberg, 2003], [Bethel, 1989], [Barcaroli et al., 2022], where despite
  - Highlighting the role of **survey costs** and **within-cluster correlation**, and
  - Impressive theoretical and algorithmic development
- Always takes the **correlation parameter as given**

Literature Review in Economics

- Little econometrics literature on K vs N
  - Similar spirit to **covariate** based **stratified treatment allocation** [Cytrynbaum, 2021], [Cytrynbaum, 2024], [Bai, 2022], [Bai et al., 2024]
  - Similar in form to **adaptive experimental design** in economics [Kasy and Sautmann, 2021], [Cesa-Bianchi et al., 2025], [Adusumilli, 2026], [Imbens et al., 2025], [Li et al., 2026]
    - Other references: **classic** literature on (panel data) econometrics with **nested random effects** [Wooldridge, 2003], bandit perspective on **optimal sampling designs** [Lattimore and Szepesvári, 2020], [Kiefer and Wolfowitz, 1960]

Bibliography I

Adusumilli, K. (2026).
How to sample and when to stop sampling: The generalised wald problem and minimax policies.
Review of Economic Studies, 93(1):1–34.

Bai, Y. (2022).
Optimality of matched-pair designs in randomized controlled trials.
American Economic Review, 112(12):3911–3940.

Bai, Y., Jiang, L., Romano, J. P., Shaikh, A. M., and Zhang, Y. (2024).
Covariate adjustment in experiments with matched pairs.
Journal of Econometrics, 241(1):105740.

Barcaroli, G., Fasulo, A., Guandalini, A., and Terribili, M. D. (2022).
Two-stage sampling design and sample selection with the r package r2beat.
arXiv preprint arXiv:2209.06532.

Bibliography II

Bethel, J. (1989).
Sample allocation in multivariate surveys.
Survey methodology, 15(1):47–57.

Biemer, P. P. and Lyberg, L. E. (2003).
Introduction to survey quality.
John Wiley & Sons.

Cesa-Bianchi, N., Colomboni, R., and Kasy, M. (2025).
Adaptive maximization of social welfare.
Econometrica, 93(3):1073–1104.

Cytrynbaum, M. (2021).
Optimal stratification of survey experiments.
arXiv preprint arXiv:2111.08157.

Bibliography III

Cytrynbaum, M. (2024).
Covariate adjustment in stratified experiments.
Quantitative Economics, 15(4):971–998.

Heß, S., Jaimovich, D., and Schündeln, M. (2021).
Development projects and economic networks: Lessons from rural gambia.
The Review of Economic Studies, 88(3):1347–1384.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2021).
Time-uniform, nonparametric, nonasymptotic confidence sequences.
The Annals of Statistics, 49(2):1055–1080.

Imbens, G., Qin, C., and Wager, S. (2025).
Admissibility of completely randomized trials: A large-deviation approach.
arXiv preprint arXiv:2506.05329.

Bibliography IV

Karing, A. (2024).
Social signaling and childhood immunization: A field experiment in sierra leone.
The Quarterly Journal of Economics, 139(4):2083–2133.

Kasy, M. and Sautmann, A. (2021).
Adaptive treatment assignment in experiments for policy choice.
Econometrica, 89(1):113–132.

Kiefer, J. and Wolfowitz, J. (1960).
The equivalence of two extremum problems.
Canadian Journal of Mathematics, 12:363–366.

Lattimore, T. and Szepesvári, C. (2020).
Bandit algorithms.
Cambridge University Press.

Bibliography V

Li, D., Raymond, L., and Bergman, P. (2026).
Hiring as exploration.
Review of Economic Studies, 93(2):1200–1240.

Lin, G. C. and Ho, S. P. (2003).
China's land resources and land-use change: insights from the 1996 land survey.
Land use policy, 20(2):87–107.

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., and Rajani, R. (2019).
Inputs, incentives, and complementarities in education: Experimental evidence from tanzania.
The Quarterly journal of economics, 134(3):1627–1673.

Bibliography VI

- Muralidharan, K. and Niehaus, P. (2017).
  Experimentation at scale.
  Journal of Economic Perspectives, 31(4):103–124.
- Osei, C. D. and Zhuang, J. (2020).
  Rural poverty alleviation strategies and social capital link: The mediation role of women entrepreneurship and social innovation.
  Sage Open, 10(2):2158244020925504.
- Sedgwick, P. (2015).
  Multistage sampling.
  Bmj, 351.
- Tam, B. (2005).
  Household sample surveys in developing and transition countries.

Bibliography VII

Wooldridge, J. M. (2003).
Cluster-sample methods in applied econometrics.
American Economic Review, 93(2):133–138.