SREE 2026 Workshop Proposal (2-hours)

Workshop Title: Matching Methods for Multilevel Data in Education Research: A Practical Guide and Illustration of Approaches in R

Instructors: Jordan Rickles, Ph.D. (UCLA) & Alberto Guzman-Alvarez, Ph.D. (American Institutes for Research)

Session Description

Propensity score matching has become a common tool for estimating causal effects over the past 25 years. However, most methodological guidance overlooks the multilevel nature common in education research where students are nested within classrooms, teachers within schools, and schools within districts. While some existing matching methods address multilevel settings (e.g., Page et al., 2020; Rickles & Seltzer, 2014), the work is limited to specific designs (e.g., two-level multisite or cluster designs) and existing reviews focus on medical or biological applications (e.g., Cafri et al., 2019; Chang & Stuart, 2022; Cui et al., 2025). Applied education researchers lack comprehensive guidance on applying matching to their desired multilevel study design. Studies that fail to address the multilevel structure can lead to biased effect estimates (Hong & Raudenbush, 2006; Miratrix et al., 2021).

To address this gap, we developed a typology and toolkit for designing and analyzing matching studies with multilevel data. The typology is rooted in the Potential Outcomes framework (Neyman, 1923), the literature on randomized experimental designs for multilevel data (Raudenbush & Schwartz, 2020), and the literature on propensity score matching.

In this workshop, we will walk participants through the analytic stages for three common multilevel designs using the Trends in International Mathematics and Science Study (TIMSS) 2015 Grade 4 Canada dataset: (a) multisite individual assignment design (MIAD), where individuals are assigned to treatment within sites (e.g., students assigned within schools); (b) cluster assignment design (CAD), where treatment is assigned at the group level (e.g., schools assigned to treatment or control); and (c) multisite cluster assignment design (MCAD), where clusters are assigned within sites (e.g., teachers assigned to professional development within schools, with students nested within teachers). Participants will leave with decision flowcharts to support design decisions and annotated R code that can be directly applied to their own research.

For each design, the toolkit provides a staged approach to designing and analyzing matching studies with multilevel data. We will walk researchers through the design and modeling decisions at each stage:

- **Stage 1:** Define the estimand (e.g., individual-average or site/cluster-average treatment effect)
- **Stage 2:** Define the distance measure (e.g., single-level or multilevel propensity score models)

- **Stage 3:** Implement a matching method (e.g., within-site, across-site, cluster-level, or sequential approaches)
- **Stage 4:** Assess the quality of the matched sample (e.g., balance of individual or site/cluster covariates)
- **Stage 5:** Analyze outcomes (e.g., single-level model or multilevel random intercept model)
- **Stage 6:** Sensitivity analysis (e.g., Rosenbaum bounds)

Throughout the workshop, we emphasize how design assumptions translate into analytic decisions rather than advocating a single "best" method.

Workshop Outline

Introduction (20 min): Overview of propensity score matching. The core challenge of multilevel data structures, consequences of ignoring nesting, an overview of the typology, and study context.

Design Walkthroughs (70 min): Full demonstration of the five-stage workflow for each design, with applied R examples using TIMSS data and attention to design-specific considerations, such as within-site versus across-site matching and site-level confounding in MIAD.

Hands-On Practice (30 min): Participants apply decision flowcharts and R code to their own data, with instructor support

Significance of the Topic and Support of Conference Theme

The SREE 2026 conference theme emphasizes public trust, accountability, and credible evidence in education research. Education policy and practice are increasingly informed by studies that draw on administrative data and naturally occurring variation across schools, districts, and programs, the kinds of multilevel observational studies where analytic choices are most consequential. When matching methods are applied without careful consideration of how treatment assignment and selection operate across levels, results may not generalize across settings, standard errors may be misaligned with the estimand of interest, and findings may be difficult to reconcile with related studies (Hong & Raudenbush, 2006; Miratrix et al., 2021). When findings cannot be reconciled across studies, the cumulative evidence base that policy depends on is weakened, posing direct risks to accountability and public confidence, particularly in high-stakes policy contexts.

This workshop supports the conference theme by providing applied, design-based guidance that makes the analytic choices in multilevel matching studies explicit, transparent, and aligned with the research design, promoting the kind of methodological transparency that is central to credible, reproducible, and policy-relevant evidence in education research.

Target Audience

This workshop is intended for education researchers and evaluators who conduct observational studies using multilevel data, including graduate students and early-career scholars seeking applied training in matching, as well as quantitative researchers working in education policy and program evaluation. Participants are expected to have basic familiarity with causal inference concepts such as potential outcomes and propensity scores.

Works Cited

Cafri, G., Wang, W., Chan, P. H., & Austin, P. C. (2019). A review and empirical comparison of causal inference methods for clustered observational data with application to the evaluation of the effectiveness of medical devices. *Statistical Methods in Medical Research*, *28*(10–11), 3142–3162. https://doi.org/10.1177/0962280218799540

Chang, T., & Stuart, E. A. (2022). Propensity score methods for observational studies with clustered data: A review. *Statistics in Medicine*, *41*(18), 3612–3626. https://doi.org/10.1002/sim.9437

Cui, C., Zhang, Y., Yang, S., Reich, B. J., & Gill, D. A. (2025). Matching estimators of causal effects in clustered observational studies. *Journal of Causal Inference*, *13*(1). https://doi.org/10.1515/jci-2024-0061

Hong, G., & Raudenbush, S. W. (2006). Evaluating Kindergarten Retention Policy: A Case Study of Causal Inference for Multilevel Observational Data. *Journal of the American Statistical Association*, *101*(475), 901–910. https://doi.org/10.1198/016214506000000447

Miratrix, L. W., Weiss, M. J., & Henderson, B. (2021). An applied researcher's guide to estimating effects from multisite individually randomized trials: Estimands, estimators, and estimates. *Journal of Research on Educational Effectiveness*, *14*(1), 270–308. https://doi.org/10.1080/19345747.2020.1831115

Neyman, J. 1923 [1990]. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science* 5 (4): 465–472.

Page, L. C., Lenard, M. A., & Keele, L. (2020). The design of clustered observational studies in education. *AERA Open*, *6*(3). https://doi.org/10.1177/2332858420954401

Raudenbush, S. W., & Schwartz, D. (2020). Randomized Experiments in Education, with Implications for Multilevel Causal Inference. *Annual Review of Statistics and Its Application*, *7*(Volume 7, 2020), 177–208. https://doi.org/10.1146/annurev-statistics-031219-041205

Rickles, J., & Seltzer, M. (2014). A two-stage propensity score matching strategy for treatment effect estimation in a multisite observational study. *Journal of Educational and Behavioral Statistics, 39*(6), 612–636. https://doi.org/10.3102/1076998614559748

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41