# Center for Assessment Summer Internship 2021

## Further Explorations of Propensity Score Weighting

Allie Cooperman

*Supporting Research by Damian Betebenner and Adam Van Iwaarden*

July 2, 2021

# Presentation Overview

1. Simulation comparing propensity score weighting and Ho's (2021) Fair Trend metric.

2. Considerations for incorporating a method like propensity score weighting into reproducible reports for large-scale assessment data.

3. Reflections on the internship experience.

# Refresher on Propensity Score Weighting

- Propensity score weighting (PSW; Rosenbaum & Rubin, 1983) can be used in quasi-experimental and observational studies to reduce the confounding effects of selection bias on parameter estimation (Bishop et al., 2018; Desai & Franklin, 2019; Leite et al., 2015).

- When analyzing educational assessment data, PSW may help to balance samples on a set of demographic characteristics before estimating differences in aggregated scale scores (Thoemmes & Kim, 2011).

- There are numerous PSW methods available (e.g., Desai & Franklin, 2019; Lee et al., 2010; Leite et al., 2015; Li et al., 2013), and their efficacy depends on the model specification and selected covariates (e.g., Bishop et al., 2018; Thoemmes & Kim, 2011).

# The Fair Trend Metric

Ho (2021) proposed a Fair Trend metric that uses a series of regressions to create adjusted scores for a cohort of "academic peers" that are comparable to students in the year of interest.

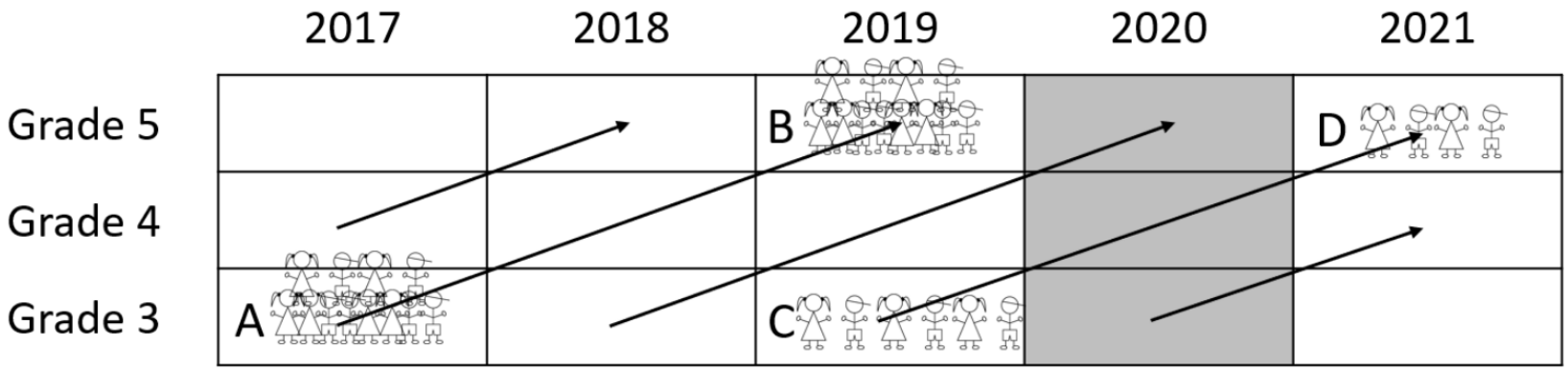


Figure 1: Replication of Figure 2 from Ho (2021, p. 5) demonstrating the cohorts used in the Fair Trend calculation.

# Simulation Design

- A small-scale, Monte Carlo simulation was designed to compare PSW and Fair Trend when estimating the difference in mean scale scores between two cross-sectional cohorts.

- Simulated data were generated from a longitudinal mixed-effects model with random intercepts for students within schools (Curran et al., 2021; DeBruine & Barr, 2021; Holmes Finch et al., 2019).

- Parameters were chosen to mirror values from `sgpData_LONG_COVID` (a simulated data set in the *SGPdata* package; Betebenner et al., 2021).

# Simulation Design

Four methods were examined:

- The Fair Trend metric
- Estimating propensity score weights with a gradient boosting method, and applying the weights to a non-hierarchical OLS regression ("PSW with OLS")
- Estimating propensity score weights with a gradient boosting method, and applying the weights to a two-level random intercept regression ("PSW with MLM")
- Computing the raw difference between mean scale scores (i.e., not applying any data adjustment)

# Simulation Results

Table 1: Average bias and root mean-squared error (RMSE) results for estimated scale score differences across 100 simulation trials

| | **Raw Bias** | **Absolute bias** | **RMSE** |
|---|---|---|---|
| **Fair Trend** | 0.558 | 0.582 | 0.669 |
| **PSW with OLS** | 0.319 | 0.420 | 0.504 |
| **PSW with MLM** | 0.287 | 0.403 | 0.489 |
| **Raw Difference** | 0.748 | 0.764 | 0.885 |

# Simulation Results

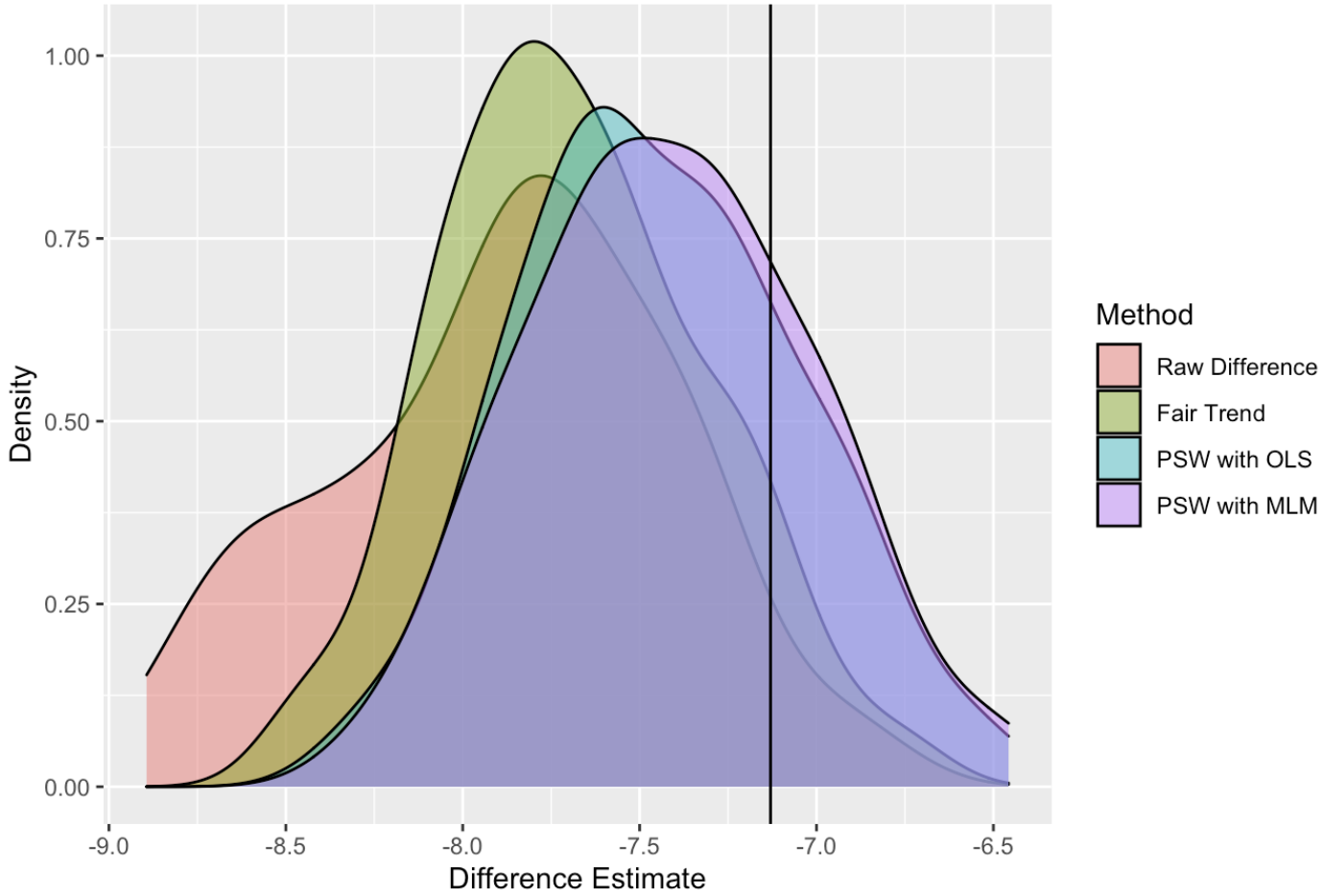


Figure 2: Density plot of estimated differences across simulation trials.

# Simulation Summary

- In the current simulation, using PSW produced the most accurate estimates of the aggregated scale score differences between cohorts.
- Both PSW and Fair Trend improved upon computing the simple raw mean difference.
- These results provide preliminary evidence that adjusting cross-sectional samples to achieve better balance on important characteristics may be beneficial when evaluating educational assessment data.
- Numerous modifications and extensions of the simulation design may lead to differing patterns of results, and warrant further exploration.

# Practical Applications of Propensity Score Weighting

How do we incorporate a method like propensity score weighting (if at all) into yearly reports of educational achievement and growth?

# Practical Applications of Propensity Score Weighting

- A long-term goal is to compile a set of R and Markdown tools that researchers and policymakers can use to produce yearly assessment reports.
- The technical details and applications of propensity score weighting seem well-suited for an appendix file in this report structure.
- Analyses can be presented with and without propensity score weighting.
- The file is structured to be as **dynamic** as possible, requiring analysts to set only a handful of parameter values before generating the output.

# Practical Applications of Propensity Score Weighting

Proposed general structure for the propensity score appendix file:

1. Examine covariate distributions for each sample.
2. Estimate propensity scores and weights (potentially using multiple methods).
3. Conduct diagnostic checks to ensure that methodological assumptions (e.g., covariate balance, common support) are met.
4. Use weights in the analysis of interest.
5. Perform sensitivity analyses.

Bishop et al., 2018; Desai & Franklin, 2019; Leite et al., 2015

# Considerations for Creating Reproducible Reports

- It can be difficult to build a dynamic report that can easily take different numbers and types of variables - but there are a lot of helpful online resources!
- How do we appropriately balance diving "into the weeds" on the technical details with presenting the key take-away points?
- How do we automate, or partially automate, model re-specifications if covariate balance is not achieved?
- Should flags for covariate imbalance be incorporated earlier in the main report?

# Internship Take-Aways

This internship has been an invaluable experience, helping me to better understand

- The applications of technical quantitative methods to large-scale assessment and education policy.
- How to best communicate statistical findings.
- The importance of reproducibility and open-source work (e.g., using tools like GitHub).
- And so much more!

# Thank You!

A massive **thank you** to Damian Betebenner, Adam Van Iwaarden, Nathan Dadey, the full Center for Assessment staff, and the 2021 Center Interns for their guidance, support, and encouragement during this internship!

# References


- Bishop, C. D., Leite, W. L., & Snyder, P. A. (2018). Using propensity score weighting to reduce selection bias in large-scale data sets. *Journal of Early Intervention, 40*(4), 347-362. https://doi.org/10.1177/1053815118793430
- Betebenner, D. W., Van Iwaarden, A. R., & Domingue, B. (2021). SGPdata: Exemplar data sets for student growth percentile (SGP) analyses. R package version 25.1-0.0. https://centerforassessment.github.io/SGPdata/
- Cefalu, M., Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B. A., & Burgette, L. (2021). twang: Toolkit for weighting and analysis of nonequivalent groups. R package version 2.1. https://CRAN.R-project.org/package=twang
- Curran, P. J., McGinley, J. S., Serrano, D., & Burfeind, C. (2012). A multivaraite growth curve model for three-level data. In H. Cooper (Ed.), *APA handbook of research methods in psychology: Vol. 3. Data analysis and research publication*. American Psychological Association. https://doi.org/10.1037/13621-017

# References


- DeBruine, L. M., & Barr, D. J. (2021). Understanding mixed-effects models through data simulation. *Advances in Methods and Practices in Psychological Science, 4*(1), 1-15. https://doi.org/10.1177/2515245920965119
- Desai, R. J., & Franklin, J. M. (2019). Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: A primer for practitioners. *The British Medical Journal, 367*, 15657. https://doi.org/10.1136/bmj.l5657
- Ho, A. (2021, Feburary 26). *Three test-score metrics that all states should report in the COVID-19-affected spring of 2021.* Harvard Graduate School of Education. https://scholar.harvard.edu/files/andrewho/files/threemetrics.pdf
- Holmes Finch, W., Bolin, J. E., & Kelley, K. (2019). *Multilevel modeling using R*. CRC Press, Taylor & Francis Group.

# References


- Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine, 29*(3), 337-346. https://doi.org/10.1002/sim.3782
- Leite, W. L., Jimenez, F., Kaya, Y., Stapleton, L. M., MacInnes, J. W., & Sandbach, R. (2015). An evaluation of weighting methods based on propensity scores to reduce selection bias in multilevel observational studies. *Multivariate Behavioral Research, 50*(3), 265-284. https://doi.org/10.1080/00273171.2014.991018
- Li, F., Zaslavsky, A. M., & Landrum, M. B. (2013). Propensity score weighting with multilevel data. *Statistics in Medicine, 32*(19), 3373-3387. https://doi.org/10.1002/sim.5786
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55.
- Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research, 46*(1), 90-118. https://doi.org/10.1080/00273171.2011.540475