Evaluating peer-assisted learning using simulated data: A propensity Score study

Juan David Cobo, Francis Duah and Boza Tasić.

Toronto Metropolitan University.

Peer-assisted learning (PAL) is a widely used support strategy in post-secondary education, aimed at enhancing student success in challenging courses such as undergraduate mathematics. PAL involves collaborative learning facilitated by trained peers. Despite its popularity, observational studies evaluating PAL in mathematics education often lack rigour due to selection bias, unbalanced covariates, and unbalanced treatment assignments, hindering valid causal inference. This paper explores the use of Propensity Score Analysis (PSA) to estimate causal effects in observational studies. PSA balances covariates across treatment groups, reducing selection bias and enabling more reliable treatment effect estimates. We examine the effectiveness of PSA in evaluating PAL research using a simulated dataset based on an observational study of a post-secondary mathematics course, with the goal of estimating the impact of PAL participation on student grade performance and comparing PSA methods.

Keywords: propensity score, peer-assisted learning, undergraduate mathematics, covariates

Introduction

Peer-assisted learning (PAL), also known in some regions as supplemental instruction (SI), is a widely adopted academic support strategy in post-secondary education. PAL embeds trained peers into challenging courses to facilitate collaborative, student-led review sessions outside of formal lectures. It is particularly prevalent in undergraduate mathematics, where attrition and underperformance remain persistent concerns for students, educators, and institutions alike (Gueudet, 2008).

The motivation for this study stems from a well-documented methodological gap in the PAL literature. While PAL programmes enjoy broad institutional support, the evidence base for their effectiveness is diminished by the observational nature of most studies. Students who choose to attend PAL sessions are not randomly assigned as they choose, or self-select, to participate. This self-selection introduces systematic differences between attendees and non-attendees across achievement and demographic variables, creating the conditions for selection bias. As a result, naive comparisons of grade outcomes between the two groups risk conflating the effects of PAL attendance with pre-existing differences in student preparation, motivation, or prior academic achievement.

This problem is not unique to mathematics education; selection bias in voluntary programmes has been widely addressed in the social sciences, medicine, and econometrics through PSA. Yet within the PAL literature, the method has received insufficient methodological attention. Many studies rely on simple regression models or raw group comparisons that fail to account for covariate imbalance, leaving

the causal status of their findings uncertain (McCarthy, Smuts, Cosser, 1997). This also matters to educational stakeholders for institutional decision-making. As without credible evidence of effectiveness, it is difficult to justify the sustained investment that PAL programmes require.

Although randomized control trials (RCTs) are considered the "gold standard" for evaluating causal effects, they are rarely used in PAL research. RCTs require random assignment of participants to treatment or control groups. In theory, randomization leads to balanced covariates between groups, thereby supporting causal conclusions. However, several ethical and practical constraints make RCTs unsuitable for SI research. For instance, withholding a potentially beneficial intervention such as SI from a control group may be viewed as unethical (Ehly & Topping, 2009). Moreover, implementing RCTs requires large sample sizes, substantial financial resources, and places additional burdens on already time-constrained students

Introduction to Propensity Score Analysis (PSA)

Propensity Score Analysis (PSA), introduced by Rosenbaum and Rubin (1983), offers a principled framework for addressing selection bias in observational research. By estimating each student's probability of attending PAL given observed background characteristics, PSA can be used to match, weight, or stratify participants in ways that produce more comparable treatment and control groups, enabling more credible causal inference without the need for randomisation. The propensity score (PS) is defined to be the conditional probability of receiving treatment given the observed covariates (specifically, $P(T = 1 \mid X)$).

The PS can be estimated using a logistic regression model to predict a student's participation based on their covariates. It is important to note that a propensity score can be assigned to a participant regardless of whether they receive treatment or not. The PS is used, in general, as a *matching* score to pair treatment and non-treatment units with similar PS to produce two balanced groups. The purpose of producing the balanced groups is to estimate the counterfactual outcomes of the participants. In the context of PAL, the counterfactual outcome would be the final grade a student who *did* attend PAL sessions would have received - had they not attended. Since this data cannot be directly observed, it is estimated through quasi-experimental designs. Specifically, the counterfactual outcome for a treatment participant (PAL attendee) is approximated by the final grade of a matched control participant (non-attendee). The difference in final grades can, in theory, be attributed to the treatment, as the covariates are balanced. Once propensity scores are calculated and assigned, the next step is to determine how to match treatment participants to control participants. However, when choosing covariates to include in the model, Kainz et al. (2017) provide practical recommendations for covariate selection and balance assessment in propensity score methods. Their results emphasize that effective covariate selection is crucial. This means, including covariates strongly associated with the outcome, rather than all available variables, helps to minimize bias and improve causal inferences.

Research Questions

Due to the gap in the literature surrounding propensity score analysis for the purpose of mathematics education, this paper aims to address the following research questions:

1. Do PSA and multiple linear regression (MLR) yield similar estimates of the treatment effect of PAL on student grade performance?

2. What are the limitations of PSA in evaluating PAL?

Addressing these questions is important for both the mathematics education research community and practitioners designing or evaluating peer learning interventions. A clearer understanding of when and how PSA can improve causal inference will support more rigorous programme evaluation across the field.

Methodology

Data source and simulation

This study draws on a simulated dataset modelled after an observational study of a Complex Variables mathematics course conducted by Duah (2017) at a UK post-secondary institution. The original study examined the impact of PAL attendance on student academic performance using achievement and demographic variables as predictors. The simulated dataset preserves the key distributional properties and covariate structure reported in that study, while enabling controlled variation in sample-size conditions that reflect the diversity of SI research contexts.

The covariates included in the simulation are: Complex Variables final grade (outcome), Calculus grade, Geometry/Vectors/Complex Numbers (GVCN) prerequisite grade, lecture attendance, gender, academic programme type, and student status (domestic/international). A student is defined as a PAL participant if they attended at least one session. Whereas the original source paper by Duah (2017) uses PAL attendance as an ordinal variable, this paper treats it as a binary variable.

The treatment criteria for PAL attendance are contested because there is no agreed standard for what counts as "participation." Dawson et al.'s (2017) systematic review found that the threshold used to classify a student as a PAL participant varied widely across studies and was often chosen without clear justification. Some studies set the bar at just one session, others at three, five, or even twelve and with some treating attendance as a continuous variable instead.

This matters because the threshold choice directly shapes who counts as "treated" in any analysis. A student who attends two sessions would be a participant in some studies and a non-participant in others, making cross-study comparisons unreliable and effect size estimates difficult to interpret consistently. It also raises a deeper question about what level of attendance is needed to produce an academic benefit, a current limitation of the binary treatment definition.

Sample size configurations

To assess the robustness of PSA across realistic research conditions, four sample size configurations were examined: equal and small (N = M = 61), equal and large (N = M = 500), unequal and small (N = 90, M = 60), and unequal and large (N = 1000, M = 500). These configurations reflect the wide variation in group sizes observed across the PAL literature and allow for systematic comparison of PSA performance under various conditions. However, for the sake of brevity, only the equal and small (N=M=61) configuration will be discussed in detail.

Descriptive statistics and pre-matching imbalance

The variables were generated using a truncated normal distribution (Johannes, 2021) to preserve the maximums, minimums, means, and standard deviations of the original

dataset (Duah, 2017). Figure 1 shows the basic descriptive statistics of the variables used in this study.

Variable	Minimum	Maximum	Mean	SD
PAL Attendance	0	11	2.29	2.81
Lecture Attendance	0	6	3.70	1.38
Complex Variables Grade	23	98	70.24	17.62
Calculus Marks	32	93	63.62	13.13
GVCN Marks	29	95	66.56	11.91

Figure 1: Descriptive statistics of variables. SD = Standard Deviation.

Moreover, using Love Plots from the Cobalt Package, one can visualise pre- and post-matching imbalance to assess whether PS matching leads to improvements in covariate balance (Greifer, 2026). Propensity score matching showed improvements in balance diagnostics on some covariates, but not others (see Figure 2). The figure shows that one-to-one matching methods (such as optimal, nearest, and quick matching) did not improve due to the lack of pruning of control units, since the two groups were the same size. However, other configurations in which the control group was larger than the treatment group saw pruning of control units to ensure strong matches with the treatment units.

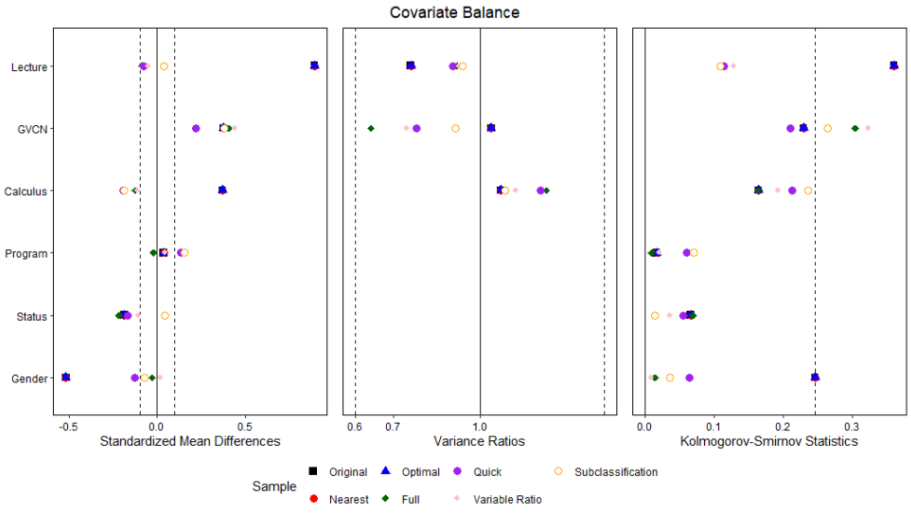

Figure 2: Love plot showing covariate balance. Small equal configuration (N = M = 61)

Main Findings

PSA and MLR Yielded Consistent Treatment Effect

Across all four sample-size configurations, PSA and MLR produced treatment-effect estimates that were directionally consistent and numerically comparable. PSA-derived average treatment effect on the treated (ATT) estimates ranged from approximately 11.2 to 17.3 grade points, with strong agreement across matching methods within each configuration. For example, in the equal small-sample configuration (N = M = 61), ATT estimates across five methods fell within a narrow range of 14.0 to 15.8 points, suggesting that the findings are robust to the choice of matching strategy. The MLR model yielded regression coefficients in a similar range (2.26–2.69), and the unadjusted Welch t-test produced results that fell within the confidence intervals of

every matched estimator. The congruence in results across methods strengthens confidence in the estimated treatment effect, suggesting that PSA is a viable method to diminish biases stemming from observational studies.

Limitations of Propensity Score Analysis

PSA's primary limitation in this context is its sensitivity to sample size. In smaller configurations (N=M=61 and N=90, M=60), achieving adequate covariate balance without substantially reducing the effective sample size proved difficult, as the limited pool of potential matches constrained how well methods create balanced groups. Moreover, a natural limitation of PSA is the collection of covariates. This is because the propensity score matching relies heavily on the collected covariates, and it may often be difficult, expensive, or unfeasible to obtain all the related variables that relate to the outcome measure. These findings underscore that PSA's theoretical advantages for bias reduction do not always translate into practice under the sample size conditions common in PAL research.

A second limitation concerns the binary treatment definition that PSA typically requires. Classifying students as simply "attendees" or "non-attendees" hides potentially important dose-response relationships. For example, whether attending three sessions produces a different effect than attending ten. As noted in the previous section, the threshold used to define participation is itself contested in the literature, which compounds this problem. While it is possible to apply PSA to clusters defined by attendance levels, thereby approximating a dose-response model, this approach requires sufficient data within each stratum to be viable. In practice, MLR handles continuous treatment variables more naturally, which represents one area where it retains an advantage over PSA for SI evaluation.

Discussion and conclusion

Implications for practice and policy

The findings carry several implications for mathematics education researchers and university mathematics teachers. Under the simulated conditions examined here, PAL attendance was associated with a meaningful positive effect on final grade performance, on the order of 11 to 18 grade points depending on configuration and method. While these are simulation-based results and should not be read as empirical confirmation of PAL's effectiveness, they are grounded in realistic data distributions drawn from Duah's thesis (2017) and are consistent with the broader PAL and SI literature.

More importantly for practitioners, this study demonstrates that the choice of evaluation methodology matters. Studies relying on unadjusted group comparisons or standard regression models risk producing misleading conclusions when treatment and control groups differ systematically, as is typical in voluntary PAL programmes where self-selection is the norm. PSA offers a principled, transparent approach to addressing this problem, and its application is feasible using open-source software (R) that is accessible to most education researchers.

A key step is that covariate balance should be routinely assessed and reported as along with reported results and for programme evaluation. Balance diagnostics such as SMDs and Love plots provide a clear visual indication of the imbalance between attendees and non-attendees and are straightforward to produce and interpret.

Institutions investing in PAL programmes should expect this level of methodological transparency from evaluators.

Limitations

Several limitations should be acknowledged. This study uses simulated rather than real-world data, which means the full complexity of student experience, including unmeasured confounders such as motivation, study habits, and socioeconomic background, is not captured. PSA can only balance observed covariates; unobserved confounding remains a threat to causal inference in any observational design. Additionally, the binary treatment definition used here may obscure important dose-response relationships, a concern compounded by the lack of consensus in the literature on what threshold of attendance constitutes meaningful participation. Finally, small and imbalanced sample sizes that are common in educational studies due to issues of student *buy-in* constrain PSA's ability to achieve adequate covariate balance, and researchers working under these conditions should exercise caution in interpreting ATT estimates.

Conclusion

This paper makes the case that Propensity Score Analysis is a valuable and underutilised tool for evaluating PAL programmes in undergraduate mathematics. When implemented carefully with appropriate covariate selection, rigorous balance assessment, and transparent reporting, PSA can substantially strengthen the credibility of causal claims drawn from observational data. The consistency observed between PSA and MLR estimates across configurations is encouraging, suggesting that both methods, when applied rigorously, can lead to similar conclusions regarding PAL effectiveness.

The primary takeaways for mathematics education researchers are threefold: (1) observational evaluations of PAL should routinely assess and report covariate balance; (2) PSA should be considered alongside standard regression as an analytical option for quasi-experimental PAL research; and (3) sample size is a critical design consideration as larger samples enable more reliable matching and more credible causal inference. Future work should apply these methods to real-world datasets across diverse PAL contexts to further validate and extend these findings.

References

Dawson, P., van der Meer, J., Skalicky, J., & Cowley, K. (2014). On the effectiveness of supplemental instruction: A systematic review of supplemental instruction and peer-assisted study sessions literature between 2001 and 2010. *Review of Educational Research, 84*(4), 609–639. https://doi.org/10.3102/0034654314540007

Duah, F. K. (2017). *Students as partners and students as change agents in the context of university mathematics* (Doctoral thesis, Loughborough University). https://hdl.handle.net/2134/25556

Duah, F., Croft, T., & Inglis, M. (2014). Can peer-assisted learning be effective in undergraduate mathematics? *International Journal of Mathematical Education in Science and Technology, 45*(4), 552–565. https://doi.org/10.1080/0020739X.2013.855329

Ehly, S., & Topping, K. J. (2009). Introduction to peer-assisted learning. In K. J. Topping & S. Ehly (Eds.), *Peer-assisted learning* (pp. 1–23). Routledge.

Greifer, N. (2025, May 29). Estimating effects after matching. In *MatchIt R package vignettes*. Comprehensive R Archive Network. https://cran.r-project.org/web/packages/MatchIt/vignettes/estimating-effects.html

Greifer, N. (2025, May 29). MatchIt: Getting started. In *MatchIt R package vignettes*. Comprehensive R Archive Network. https://cran.r-project.org/web/packages/MatchIt/vignettes/MatchIt.html

Greifer, N. (2026, January 23). Using love.plot() to generate love plots. In *cobalt R package vignettes*. Comprehensive R Archive Network. https://cran.r-project.org/web/packages/cobalt/vignettes/love.plot.html

Gueudet, G. (2008). Investigating the secondary–tertiary transition. *Educational Studies in Mathematics, 67*(3), 237–254.

JohannesNE. (2021, February 24). Simulate constrained normal on lower or upper bound in R. *Stack Exchange*. https://stats.stackexchange.com/q/510821

Kainz, K., Chalfin, A., & Rotz, D. (2017). Improving causal inference in educational research: The promise and challenges of propensity score methods. *Review of Research in Education, 41*(1), 128–158. https://doi.org/10.3102/0091732X16686755

McCarthy, A., Smuts, B., & Cosser, M. (1997). Assessing the effectiveness of supplemental instruction: A critique and a case study. *Studies in Higher Education, 22*(2), 221–231. https://doi.org/10.1080/03075079712331381054

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