Do Extracurricular Activities give '*Extra*' on Academic Performance? Evidence from Propensity Score Matching Methods

B M Nozaleda[12*]

[1]College of Human Kinetics, Cagayan State University-Carig Campus, Philippines
[2]College of Education, University of the Philippines-Diliman, Quezon City, Philippines

Corresponding author's email: bmnozaleda@up.edu.ph

Abstract. This study compares different statistical methods to determine whether participating in extracurricular activities helps improve students' academic performance. Utilizing a dataset of 1,000 students, the study balances students who did and did not take part in extracurriculars by adjusting for factors like study hours and attendance. It compares Nearest Mahalanobis Distance, Nearest Neighbor Matching (with and without a caliper), Optimal Pair Matching, Optimal Full Matching, Coarsened Exact Matching (CEM), and Inverse Probability Weighting (IPW) based on covariate balance, sample retention, and average treatment effect. Results reveal that IPW performs best in the covariates balance, reducing nearly all standardized mean differences to near zero while retaining the majority of the dataset. Nearest Neighbor Matching with Caliper and Optimal Pair Matching also perform well with significant treatment effect estimates and relatively strong model fits. However, each method involves trade-offs in which IPW excels in covariate balance but has a higher AIC, a slight compromise in model fit, while Nearest Neighbor Matching with Caliper offers a balance between precision, model fit, and sample retention. In contrast, CEM provides strong covariate balance for categorical variables but results in significant sample loss, demonstrating the trade-off between strict matching criteria and practical applicability. Conversely, Nearest Neighbor Matching without Caliper performed poorly in balancing covariates. As evidenced by the average treatment effect estimates derived from the propensity score matching (PSM) methods, this study concludes that participation in extracurricular activities has a positive and significant impact on students' academic performance, with study hours, attendance, and resource accessibility emerging as critical factors as well. The novelty of this study is in comparing multiple statistical matching approaches side by side in an educational context, providing guidance for researchers and policymakers.

Keyword: Academic Performance; Extracurricular Activities; Inverse Probability Weighting; Observational Study; Propensity Score Matching; Propensity Score Analysis.

1. Introduction

Propensity score matching (PSM) has emerged as a powerful tool for reducing confounding in observational studies, allowing researchers to estimate causal effects more reliably by creating comparable treatment and control groups [1], [2], [3]. The propensity score, defined as the conditional probability of treatment assignment given observed covariates, can be employed through various methods such as matching, stratification, and inverse probability weighting [4], [5]. Among these, matching techniques are particularly prominent, as they facilitate direct comparisons between treated and untreated groups by minimizing differences in baseline characteristics.

This paper specifically focuses on the implementation and comparison of various Propensity Score Matching (PSM) methods, including Nearest Mahalanobis distance, Nearest Neighbor Matching (with and without a caliper), Optimal Pair Matching, Optimal Full Matching, Coarsened Exact Matching (CEM), and Inverse Probability Weighting (IPW). The comparative evaluation of these methods is essential, as previous research highlights trade-offs between bias reduction, precision, and computational efficiency [6]. For instance, caliper matching is particularly effective in mitigating bias by restricting the allowable differences in propensity scores between matched pairs, while nearest neighbor matching often achieves greater precision in treatment effect estimation by pairing treated and control units with minimal propensity score differences [7]. It is hoped that the findings will offer helpful insights into the trade-offs between covariate balance, sample retention, and treatment effect to the methodological rigor of studies employing matching techniques in observational research.

To contextualize, the application of PSM in this study addresses a practical and widely debated educational question, the impact of extracurricular activity participation on students' academic performance. Researchers argue that such activities enhance cognitive and social skills [8], whereas critics caution against potential distractions from academic priorities [9], [10]. By employing matching algorithms, this analysis aims to isolate the causal effect of extracurricular activities on final exam scores, providing clarity to these competing claims.

The remainder of this paper is structured as follows: the next section presents the case study, detailing the dataset, variables, and context of the analysis. Subsequent sections describe the methodological framework, including the implementation of various PSM techniques, statistical software, and report the results of balance assessments and treatment effect estimation.

Case Study

Extracurricular activities have long been recognized as avenues for holistic student development, offering opportunities to enhance cognitive, social, and emotional skills [11]. Empirical research has highlighted both positive and negative relationships between extracurricular participation and academic outcomes. For instance, King et al. [12] found that students involved in extracurricular activities often achieve better academic performance, attributing this to improved time management and heightened motivation. On the other hand, Liang et al. [13] and Bacon and Lord [14] observed that excessive participation might detract from study time, potentially diminishing academic focus. These mixed findings underscore the need for a nuanced approach that accounts for confounding factors when evaluating the impact of extracurricular activities on academic achievement.

The theoretical framework for this study is grounded in developmental systems theory, which emphasizes the dynamic interplay between individual characteristics and environmental contexts in shaping outcomes [15]. Participation in extracurricular activities can be conceptualized as a proximal process that nurtures skills critical for academic success, such as teamwork, discipline, and goal-setting. This study posits that the effect of extracurricular involvement on academic performance is mediated by a range of covariates, necessitating careful control to isolate causal effect.

The dataset analyzed in this study includes detailed information on 6,607 students, sourced from a publicly available repository. To make the analysis manageable, a subset of 1,000 students using stratified random sampling was done, preserving the proportions of participants vs. non-participants and of gender from the full dataset. The treatment variable, participation in extracurricular activities, is binary, dividing students into treated (participants) and control (non-participants) groups. The outcome variable is final exam scores, which serves as a quantitative measure of academic performance.

The analysis incorporates key covariates that influence academic performance. Study hours, attendance rates, and sleep hours capture individual effort and habits that contribute to academic success. External influences are measured through variables such as access to educational resources (categorized as Low, Medium, and High), motivation level (categorized as Low, Medium, and High), and teacher quality (categorized as Low, Medium, and High), describing institutional and contextual variations. Additionally, gender is included as a demographic variable to account for potential gender-based differences in academic outcomes.

2. Research Method

As previously mentioned, the matching methods employed in this case study are Nearest Mahalanobis Distance, Nearest Neighbor Matching (with and without caliper), Optimal Pair Matching, Optimal Full Matching, Coarsened Exact Matching (CEM) based on the categorical variables, and Inverse Probability Weighting (IPW).

Yasunaga [16] broadly categorizes matching algorithms into two types: nearest neighbor matching (also known as greedy matching) and optimal matching. In nearest neighbor matching, a participant is randomly selected from the treatment group and paired with a participant from the control group who has the closest propensity score. Nearest neighbor matching is often implemented with a caliper, ensuring that the propensity scores of matched pairs fall within a predefined range. Narrower caliper widths result in more closely matched pairs but reduce the overall number of matches. Research has demonstrated that using a caliper width of 0.25 times the standard deviation of the logit of the propensity score can eliminate 98% of the bias associated with measured covariates [17], [18]. Another study recommended a caliper width of 0.2 times the standard deviation of the propensity score logit as an effective choice [19], [20]. Meanwhile, Optimal Pair Matching seeks to minimize the total within-pair differences in propensity scores while retaining all treated units, thereby achieving balanced groups while preserving the sample size [17], [18]. Optimal Full Matching further improves upon this by using a weighted structure to match treated and control units in an optimal manner, but at the cost of extreme weights that can influence the effective sample size [20], [21].

Moreover, Coarsened Exact Matching (CEM), in contrast, uses a preprocessing step to group observations into coarsened strata based on selected covariates. In this study, categorical variables—motivation, access to resources, teacher quality, and sex—were used in the coarsening procedure. Observations from the treated and control groups that fell into the same strata were matched exactly, thereby achieving strong covariate balance but often at the cost of reduced sample size [22]. Finally, Inverse Probability Weighting (IPW) reweights observations based on their propensity scores to create a pseudo-population where the distribution of covariates is balanced across treated and control groups [23].

The analysis for this study was conducted using RStudio version 2024.12.0+467. Several R packages were utilized to implement the matching methods and evaluate covariate balance and treatment effects. The *MatchIt* package was employed for the implementation of propensity score matching methods, including Nearest Neighbor Matching (with and without caliper), Optimal Pair Matching, and Optimal Full Matching. Coarsened Exact Matching (CEM) was performed using the cem package, which allows for exact matching based on coarsened variables.

To assess covariate balance, the cobalt package was used to generate Love plots and calculate absolute standardized mean differences (ASMDs). For estimating treatment effects after matching, the *marginaleffects* package was applied, facilitating computation of average treatment effects with appropriate weighting. Additionally, visualization and data management were supported by the *ggplot2* package for graphical representation and the *dplyr* package for data manipulation.

3. Result and Discussion

3.1 Results of the Covariance Balance Assessment

The balance assessment results across different matching methods are presented in Figures 1 through 6. These figures show the absolute standardized mean differences (ASMDs) for each covariate before and after matching indicating the degree of balance achieved by each method.

Figure 1 illustrates the results for Nearest Mahalanobis Distance. Based on the figure, the balance assessment reveals that all covariates, both in the unadjusted and adjusted data, are within the commonly accepted threshold of 0.1 ASMD. However, matching resulted in an improvement in balance for study hours and attendance, with closer to zero ASMD in the adjusted data compared to the unadjusted data. In contrast, for sleep, the balance did not improve after matching, as the ASMD for the unadjusted is lower than the adjusted data.

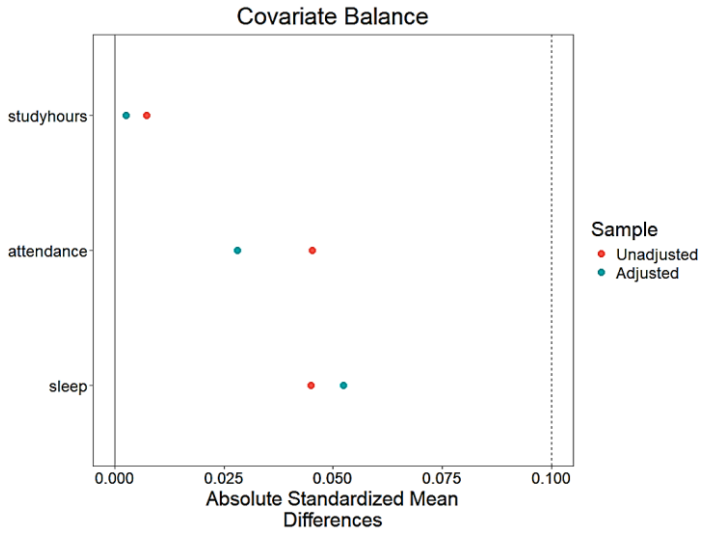


Figure 1. Covariate balance for Nearest Mahalanobis Distance

Moving on, figure 2 displays the balance assessment for Nearest Neighbor Matching without (a) and with caliper (b), respectively. Nearest Neighbor Matching without caliper was applied to match treated and control units by identifying the closest untreated neighbor for each treated individual based on their propensity scores. This approach does not impose any restrictions on the allowable differences in propensity scores between matched pairs, providing a straightforward method for matching (Stuart, 2010). However, the absence of restrictions allows for the inclusion of poor matches, particularly when large differences in propensity scores exist between treated and control units. The balance assessment revealed that many covariates still exhibited absolute standardized mean differences (ASMDs) exceeding the commonly accepted threshold of 0.1 even after matching. Notable examples of poorly balanced covariates include attendance, sleep, resources, and teacher quality, where ASMDs remained significantly above the threshold. These results suggest that this matching method, while simple, may not adequately address the differences between treated and control groups in this dataset. The persistence

of imbalance highlights the limitations of Nearest Neighbor Matching without caliper in ensuring robust covariate balance, particularly when large propensity score differences are present.

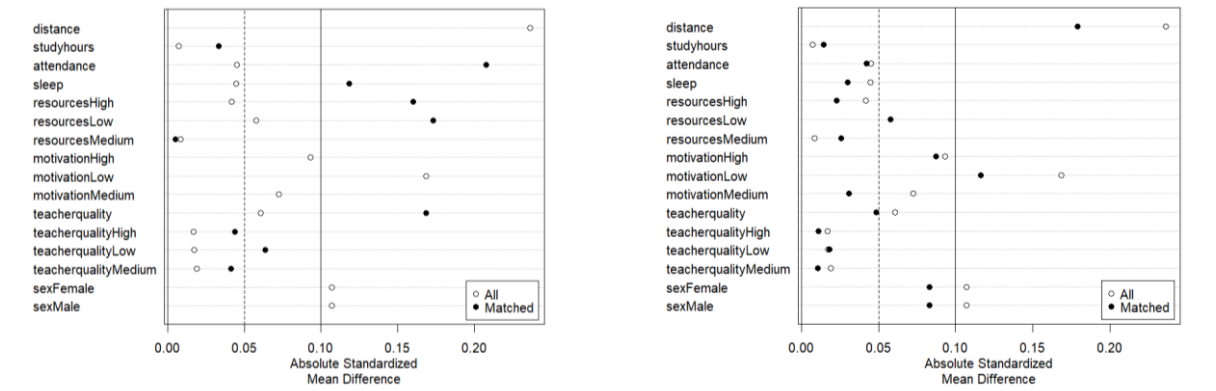


Figure 2. Covariate balance for Nearest Neighbor Matching without Caliper (left) and with caliper (right)

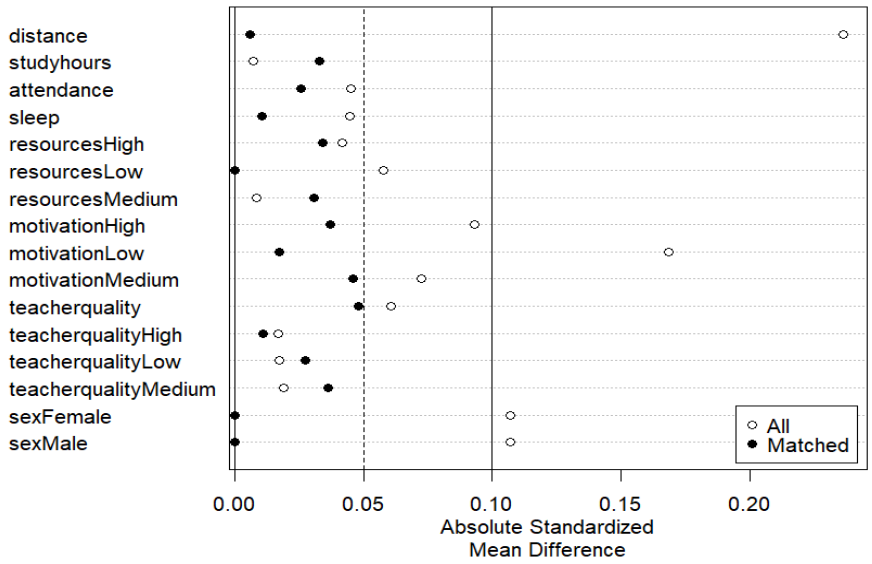


Figure 3. Covariate balance for Optimal Pair Matching

To address the limitations observed in Nearest Neighbor Matching without caliper, a caliper of 0.2 was introduced to restrict the maximum allowable difference in propensity scores between matched treated and control units, following the guideline recommended by Austin [19]. The caliper acts as a quality control mechanism, ensuring that only pairs with small differences in propensity scores are included while discarding poor matches that could bias the results. The balance assessment (Figure 2a) following the application of a caliper demonstrated substantial improvements. Most covariates achieved ASMDs below the 0.1 threshold, indicating enhanced comparability between treated and control groups. However, low motivation remained above the threshold. Overall, the use of a caliper significantly improved the quality of matches compared to the no-caliper method.

Meanwhile. the balance assessment for Optimal Pair Matching shown in figure 3 indicates a noticeable improvement in covariate balance between treated and control groups. The absolute standardized mean differences (ASMDs) for most covariates in the matched data are substantially reduced compared to the unmatched data. Key covariates such as attendance, sleep, and sex , signified good balance after matching. In the case of the optimal full matching (Figure 4), its noticeable that the balance is not as good as optimal pair matching. Two covariates, sleep and teacher quality, exceeded the threshold of 0.1 ASMD.

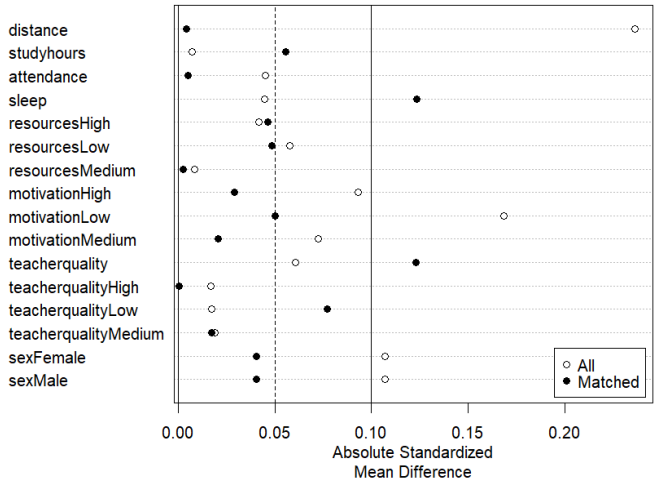


Figure 4. Covariate balance for Optimal Full Matching

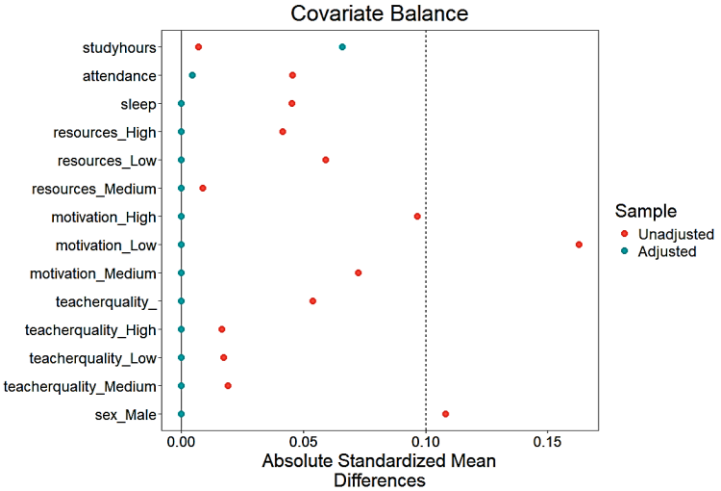


Figure 5. Covariate balance for Coarsened Exact Matching

Moving on, Figure 5 depicts the results for Coarsened Exact Matching (CEM). In the Coarsened Exact Matching (CEM) procedure, categorical variables were utilized to create bins for exact matching between treated and control units. The categorical variables used for this process were motivation level, teacher quality, resources, and sex with each of their respective categories forming distinct bins. It further shows a marked improvement in absolute standardized mean differences (ASMDs) for most covariates after matching, with the adjusted data demonstrating substantially lower ASMDs compared to the unadjusted data. It can be noted though that continuous covariates have higher ASMDs than the unadjusted data which could be a result of the coarsening process. Overall, CEM effectively balances

all categorical covariates and demonstrates its utility in improving comparability between treated and control groups, though the trade-off between balance and sample retention should be considered which will be shown in the succeeding sections.

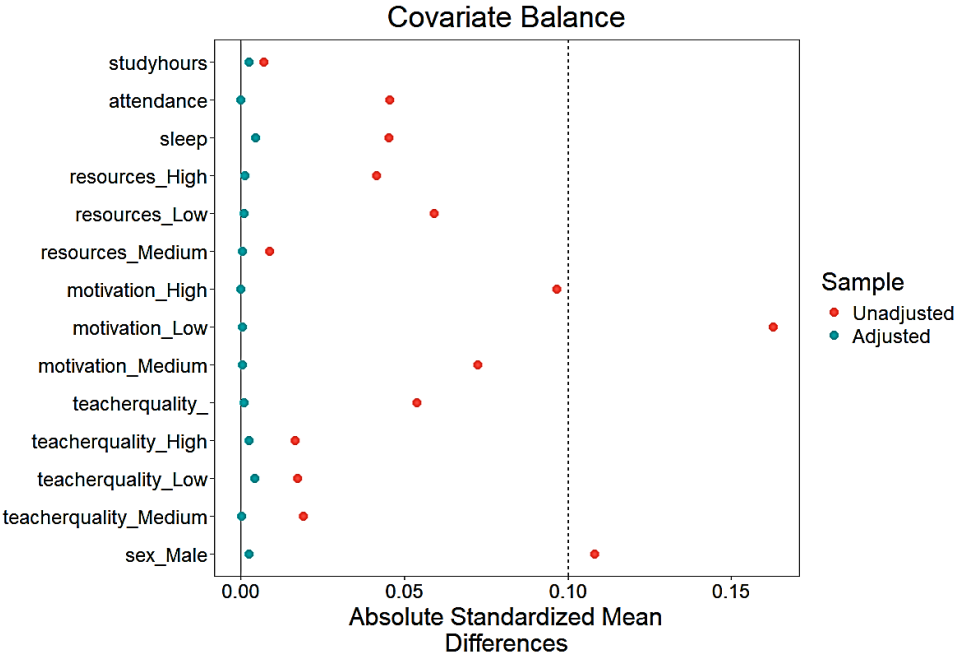


Figure 6. Covariate balance for Inverse Probability Weighting

Lastly, Figure 6 provides the balance assessment for Inverse Probability Weighting (IPW). The balance assessment for Inverse Probability Weighting (IPW) reweights observations based on propensity scores. The plot indicates that, after weighting, absolute standardized mean differences (ASMDs) for nearly all covariates have been reduced to near zero. Covariates such as motivation, sleep, and sex, show significant improvement, indicating that the weighting scheme effectively accounts for differences between treated and control groups.

In summary, the overall covariate balance assessment across all matching methods in reducing absolute standardized mean differences (ASMDs) between treated and control groups is shown in Figure 7. The plot demonstrates substantial variation in performance, with some methods achieving excellent balance while others leave notable residual imbalances.

Inverse Probability Weighting (IPW) stands out as the most effective method, with nearly all covariates achieving ASMDs close to zero. Optimal Pair Matching and Coarsened Exact Matching (CEM) also perform well, with most covariates falling below the 0.1 threshold, though slight residual imbalances remain for certain variables such as teacher quality and sleep hours. In contrast, Nearest Neighbor Matching without a caliper perform poorly, with several covariates showing large ASMDs well above the 0.1 threshold. Introducing a caliper to Nearest Neighbor Matching significantly improves balance, although residual imbalances persist for those with low motivation.

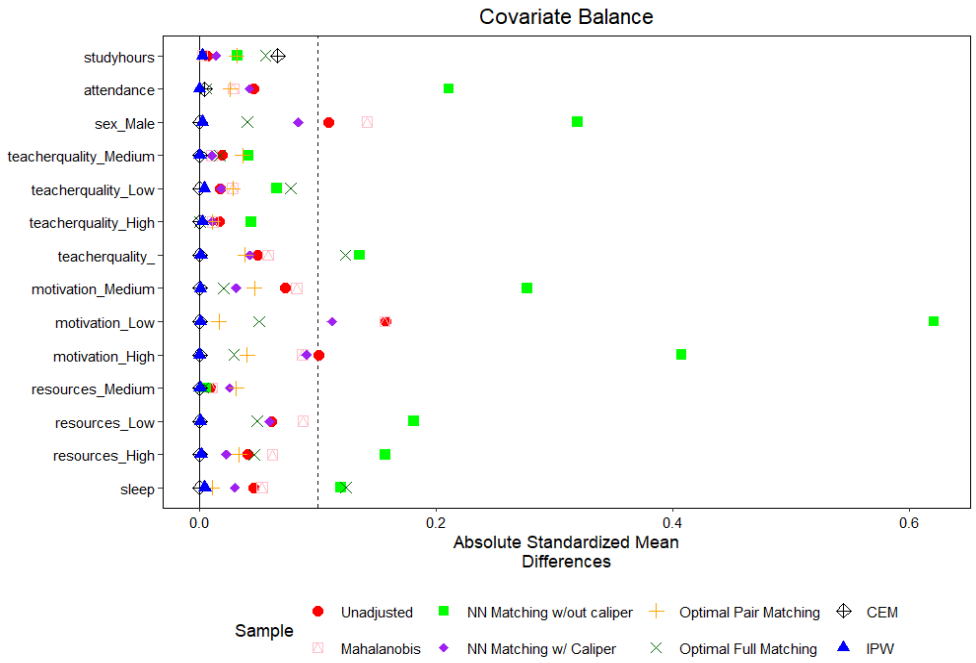


Figure 7. Covariate Balance Across Matching Methods

3.2 Sample Retention across Matching Methods

To further evaluate the matching methods, data on sample retention after matching is shown in table 1. The table reveals differences among the methods, which, when considered alongside the balance assessment, highlight trade-offs between retaining observations and achieving covariate balance. Nearest Mahalanobis Distance retained all treated and control units, but the balance assessment in some covariates indicated poor performance. Similarly, Nearest Neighbor Matching without caliper preserved the entire dataset but failed to adequately improve balance for many covariates, with several remaining far above the 0.1 ASMD threshold.

Table 1. Sample Sizes Across Matching Methods

	Mahalanobis		**NN w/out caliper**		**NN w/ caliper**		**Optimal Pair Matching**		**Optimal Full Matching**		**Coarsened Exact Matching**		**Inverse Probability Weighting**	
	Cont	*Treat*	*Cont*	*Treat*	*Cont*	*Treat*	*Cont*	*Treat*	*Cont*	*Treat*	*Cont*	*Treat*	*Cont*	*Treat*
All	390	610	390	610	390	610	390	610	390	610	390	610	390	610
Matched	390	390	390	390	388	388	390	390	197.46*	610	20	20	382.34*	604.85*
Unmatched	0	220	0	220	2	222	0	220	0	0	940	20	0	0
Discarded	0	0	0	0	0	0	0	0	0	0	0	0	0	0

**Effective sample size*

Introducing a caliper of 0.2 in Nearest Neighbor Matching resulted in the exclusion of two treated and control units, improving balance significantly by discarding poor matches with large propensity score differences. This adjustment led to better overall covariate balance, though low motivation

remained imbalanced. Optimal Pair Matching retained the full treated sample while matching it strictly to the control units, achieving good balance across most covariates, including attendance and sex.

Meanhwile, Optimal Full Matching retained all observations through a weighting mechanism, achieving good balance across most covariates. However, the effective sample size (ESS) for the control group was reduced to 197.46, indicating potential issues with extreme weights and their impact on representativeness. Meanwhile, Coarsened Exact Matching (CEM), which was based on the four categorical variables, resulted in a total of 40 matched units, equally split between 20 treated and 20 control observations. The remaining 940 control units and 20 treated units were unmatched. This is expected as this method is strict on achieving exact matches while considerably reducing the sample size.

Finally, Inverse Probability Weighting (IPW) retained most of the dataset while achieving the best overall covariate balance, with nearly all ASMDs reduced to near zero. Its ESS values of 382.34 for control units and 604.85 for treated units further underscore its efficiency in balancing covariates without discarding any observations.

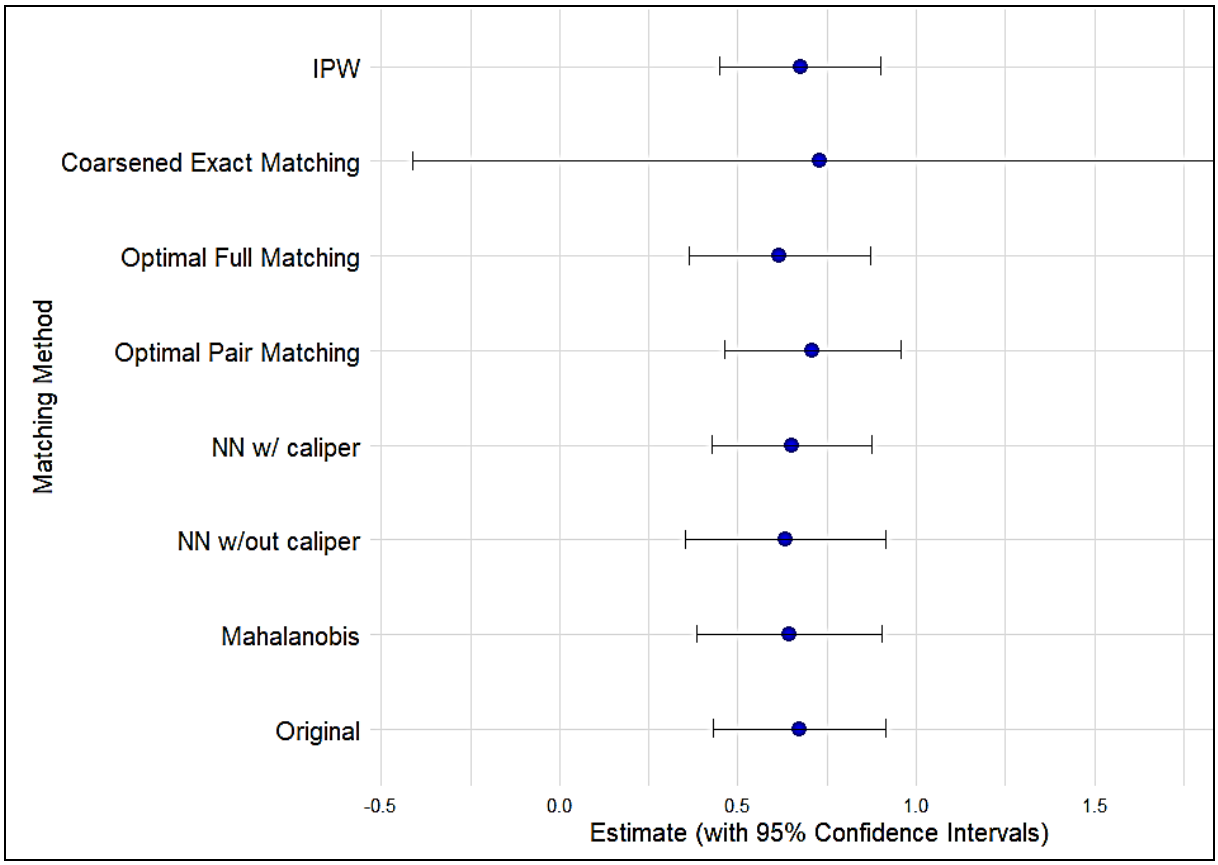
Therefore, based on the results of the covariate balance assessment and sample retention across matching methods, Inverse Probability Weighting (IPW) emerges as the best method. It achieves superior covariate balance, with nearly all ASMDs reduced to near zero, and retains a substantial effective sample size (ESS) of 382.34 for control units and 604.85 for treated units. Optimal Full Matching and Coarsened Exact Matching (CEM) also perform well in terms of covariate balance, with most covariates falling below the 0.1 ASMD threshold. However, CEM's stringent exact matching approach leads to significant sample loss, matching only 40 observations, while Full Matching retains all observations but suffers from reduced ESS due to extreme weights. In contrast, Nearest Neighbor Matching without caliper perform poorly, failing to adequately improve balance for many covariates, and are thus identified as the least effective methods.

3.3 Estimates of the Average Treatment Effect

To evaluate the impact of extracurricular activities on students' academic performance, a regression-based approach in R using linear modeling combined with weights derived from various matching methods was implemented. This procedure estimates the treatment effect while accounting for differences between treated and control groups as adjusted by each matching approach. Specifically, the researcher fitted a linear regression model using the *lm()* function, where the treatment variable (treatment), indicating participation in extracurricular activities, was included alongside the covariates (e.g., study hours, attendance, sleep, resources, motivation, and teacher quality) to model the outcome variable, final exam scores (outcome).

To evaluate the impact of extracurricular activities on students' academic performance, a regression-based approach in R using linear modeling combined with weights derived from various matching methods was implemented. This procedure estimates the treatment effect while accounting for differences between treated and control groups as adjusted by each matching approach. Specifically, the researcher fitted a linear regression model using the *lm()* function, where the treatment variable (treatment), indicating participation in extracurricular activities, was included alongside the covariates (e.g., study hours, attendance, sleep, resources, motivation, and teacher quality) to model the outcome variable, final exam scores (outcome).

Table 2. Estimated Average Treatment Effect Across Matching Methods



	Original	Mahalanobis	NN w/out caliper	NN w/ caliper	Optimal Pair Matching	Optimal Full Matching	Coarsened Exact Matching	Inverse Probability Weighting
Estimate	0.673	0.644	0.633	0.651	0.71	0.617	0.731	0.675
Std. Error	0.124	0.133	0.143	0.115	0.127	0.13	0.584	0.115
Pr(>|z|)	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.211	<0.001
AIC	4121.74	3178.24	3157.37	2948.51	3120.81	4216.42	176.196	4085.42

For each matching method, the weights generated during the matching process were applied to the regression model using the weights argument. The model was run for both the matched and unmatched datasets to provide a comparative analysis of the treatment effects. After fitting the model, the *marginaleffects* package was utilized to compute the Average Treatment Effect on the Treated (ATT). Specifically, the *avg_comparisons()* function was used to estimate the ATT by comparing the predicted outcomes for treated and control groups, incorporating the subclass structure (if applicable) via the *vcov = ~subclass* argument and using the generated weights.

Table 2 summarizes the treatment effect estimates, standard errors, significance levels, and model fit (as measured by AIC) across the original unadjusted data and the various matching methods. This allows for a direct comparison of how each method influences the estimated effect of extracurricular participation on exam scores, as well as their relative efficiency and model fit.

The significant treatment effect estimates range from 0.617 (Optimal Full Matching) to 0.71 (Optimal Pair Matching), indicating that participating in extracurricular activities is associated with an increase in exam scores across all methods. The original unadjusted data produced an estimate of 0.673, which falls within the range of estimates provided by the matching methods, suggesting that adjusting for covariates via matching has a modest impact on the effect size. Optimal Pair Matching yielded the highest estimated effect (0.71), while Optimal Full Matching provided the lowest (0.617). Both Nearest

Neighbor with caliper (0.651) and Inverse Probability Weighting (0.675) produced estimates close to the original.

In contrast, Coarsened Exact Matching (CEM) produces an estimate of 0.731 with a relatively large standard error of 0.584 and a non-significant p-value (0.211), reflecting instability due to the drastic reduction in the matched sample size. Despite its high ATT estimate, CEM's extremely low AIC value of 176.196 highlights potential overfitting and suggests caution when interpreting these results.

Based on the model fit statistics, Nearest Neighbor Matching with Caliper (AIC = 2948.51) and Optimal Pair Matching (AIC = 3120.81) demonstrate the best performance among the methods, as they have the lowest AIC values. Notably, Optimal Pair Matching achieves the highest treatment effect estimate (0.71), though the significant estimates across the methods are largely comparable. However, due to the strict sample retention in Optimal Pair Matching, which limits flexibility and representativeness by requiring exact matches, Nearest Neighbor Matching with Caliper is a more practical choice.

If covariate balance is given the highest priority in selecting the superior method, Inverse Probability Weighting (IPW) emerges as the most favorable approach. IPW achieves nearly perfect covariate balance with all absolute standardized mean differences (ASMDs) reduced to near zero, ensuring that the treated and control groups are highly comparable. While its AIC value (4085.42) is higher than that of NN with Caliper or Optimal Pair Matching, the superior covariate balance achieved by IPW minimizes potential bias in the treatment effect estimation, making it an excellent alternative for ensuring internal validity.

3.4 Effect of Extracurricular Activities on Academic Performance

Across the methods chosen for estimating the treatment effect, significant and positive effects were observed, underscoring the potential benefits of participating in extracurricular activities. Specifically, the treatment effect estimates across NN with Caliper and IPE consistently indicate that students participating in extracurricular activities achieve higher final exam scores compared to their non-participating peers. For instance, the effect size estimated using NN with Caliper (0.651) and IPW (0.675) suggests that these activities provide students with measurable academic advantages. These findings align with existing literature, such as Knifsend & Graham [24], who reported positive associations between extracurricular participation and academic outcomes.

The results of the multiple linear regression analysis on the matched data provide further insights into the factors influencing academic performance, aligning with established literature on academic outcomes. Both methods revealed consistent patterns, with several covariates demonstrating strong associations with final exam scores. Study hours and attendance were positively and significantly associated with higher exam scores, corroborating findings from studies such as Kauffman et al. [25], which highlighted the direct role of time invested in studying and classroom participation in enhancing academic achievement. Similarly, Kim et al. [26] underscored the predictive power of attendance in academic outcomes, attributing this to increased engagement and exposure to course content.

Conversely, resources and motivation levels showed significant negative associations in certain categories. Students with low and medium levels of resources exhibited markedly lower exam scores than their peers with high resources. This finding aligns with Schmidt et al. [27] and Elenbaas & Killen [28], whose works emphasized the role of resource disparities in perpetuating academic inequalities. Similarly, Masa'deh et.al [29] and Oakes [30] found that access to adequate learning resources, such as books and technology, positively impacts academic performance, particularly in underprivileged groups. Moreover, low and medium motivation levels were significantly associated with poorer outcomes, reinforcing the conclusions of Ryan and Deci [31], who identified intrinsic and extrinsic motivation as critical drivers of student effort and persistence in academic tasks.

Notably, sleep hours and other covariates did not reach statistical significance which may imply that their direct effects on academic performance may be less pronounced or mediated by other factors. Although the positive effects of sleep on cognitive function and memory are well-documented in

systematic reviews [32], [33], this study's findings may reflect variability in sleep patterns or individual differences in sleep needs that offset its direct association with academic outcomes.

From a policy and educational perspective, the findings reinforce the argument for encouraging well-balanced extracurricular programs within schools. By encouraging student engagement through such activities, educational institutions can leverage the holistic benefits of extracurricular participation to enhance academic performance while addressing concerns about potential overcommitment.

4. Conclusion

This study concludes that participation in extracurricular activities has a positive and significant impact on students' academic performance, as evidenced by significant treatment effect estimates across various propensity score matching methods. Both Nearest Neighbor Matching with Caliper and Inverse Probability Weighting (IPW) emerge as superior methods, offering valid treatment effect estimates while balancing trade-offs between covariate balance and sample retention. Nearest Neighbor Matching with Caliper demonstrated strong model fit and reasonable covariate balance, making it a practical choice for many applications. On the other hand, IPW achieved near-perfect covariate balance, ensuring high internal validity but with slightly higher model fit metrics.

References

[1] M. Y. Q. Liau, E. Q. Toh, S. Muhamed, S. V. Selvakumar, and V. G. Shelat, "Can propensity score matching replace randomized controlled trials?," *World J Methodol*, vol. 14, no. 1, 2024.

[2] M. Maekawa, A. Tanaka, M. Ogawa, and M. H. Roehrl, "Propensity score matching as an effective strategy for biomarker cohort design and omics data analysis," *PLoS One*, vol. 19, no. 5, p. e0302109, 2024.

[3] K. Narita, J. D. Tena, and C. Detotto, "Causal inference with observational data: A tutorial on propensity score analysis," *Leadersh Q*, vol. 34, no. 3, p. 101678, 2023.

[4] V. Allan *et al.*, "Propensity score matching and inverse probability of treatment weighting to address confounding by indication in comparative effectiveness research of oral anticoagulants," *J Comp Eff Res*, vol. 9, no. 9, pp. 603–614, 2020.

[5] P. C. Austin, "Propensity Score Analysis With Baseline and Follow-Up Measurements of the Outcome Variable," *Pharm Stat*, 2024.

[6] S. Poletto, E. Longato, E. Tavazzi, and M. Vettoretti, "Comparing Propensity Score-Based Methods in Estimating the Treatment Effects: A Simulation Study," *arXiv preprint arXiv:2408.17385*, 2024.

[7] S. Mojarad, R. S. Baker, A. Essa, and S. Stalzer, "Replicating studying adaptive learning efficacy using propensity score matching and inverse probability of treatment weighting," *Journal of Interactive Learning Research*, vol. 32, no. 3, pp. 169–203, 2021.

[8] J. Murray and D. Cousens, "Primary school children's beliefs associating extra-curricular provision with non-cognitive skills and academic achievement," *Educ 3 13*, vol. 48, no. 1, pp. 37–53, 2020.

[9] B. Afalla, "Blending Extracurricular Activities with Academic Performance: Pain or Gain?," *Humanities and Social Sciences Reviews*, 2020.

[10] R. D. Heath, C. Anderson, A. C. Turner, and C. M. Payne, "Extracurricular activities and disadvantaged youth: A complicated—but promising—story," *Urban Educ (Beverly Hills Calif)*, vol. 57, no. 8, pp. 1415–1449, 2022.

[11] P. Buckley and P. Lee, "The impact of extra-curricular activity on the student experience," *Active Learning in Higher Education*, vol. 22, no. 1, pp. 37–48, 2021.

[12] A. E. King, F. A. E. McQuarrie, and S. M. Brigham, "Exploring the relationship between student success and participation in extracurricular activities," *SCHOLE: A Journal of Leisure Studies and Recreation Education*, vol. 36, no. 1–2, pp. 42–58, 2021.

[13] Q. Liang, W. Niu, L. Cheng, and K. Qin, "Creativity outside school: The influence of family background, perceived parenting, and after-school activity on creativity," *J Creat Behav*, vol. 56, no. 1, pp. 138–157, 2022.

[14] P. Bacon and R. N. Lord, "The impact of physically active learning during the school day on children's physical activity levels, time on task and learning behaviours and academic outcomes," *Health Educ Res*, vol. 36, no. 3, pp. 362–373, 2021.

[15] P. E. Griffiths and R. D. Gray, "Discussion: Three ways to misunderstand developmental systems theory," *Biol Philos*, vol. 20, pp. 417–425, 2005.

[16] H. Yasunaga, "Introduction to applied statistics—chapter 1 propensity score analysis," *Annals of Clinical Epidemiology*, vol. 2, no. 2, pp. 33–37, 2020.

[17] P. R. Rosenbaum, "Modern algorithms for matching in observational studies," *Annu Rev Stat Appl*, vol. 7, no. 1, pp. 143–176, 2020.

[18] P. R. Rosenbaum and D. B. Rubin, "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score," *Am Stat*, vol. 39, no. 1, pp. 33–38, 1985.
[19] P. C. Austin, "Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies," *Pharm Stat*, vol. 10, no. 2, pp. 150–161, 2011.
[20] P. C. Austin and E. A. Stuart, "Optimal full matching for survival outcomes: a method that merits more widespread use," *Stat Med*, vol. 34, no. 30, pp. 3949–3967, 2015.
[21] P. C. Austin and E. A. Stuart, "Estimating the effect of treatment on binary outcomes using full matching on the propensity score," *Stat Methods Med Res*, vol. 26, no. 6, pp. 2505–2525, 2017.
[22] J. E. Ripollone, K. F. Huybrechts, K. J. Rothman, R. E. Ferguson, and J. M. Franklin, "Evaluating the utility of coarsened exact matching for pharmacoepidemiology using real and simulated claims data," *Am J Epidemiol*, vol. 189, no. 6, pp. 613–622, 2020.
[23] F. Thoemmes and A. D. Ong, "A primer on inverse probability of treatment weighting and marginal structural models," *Emerging Adulthood*, vol. 4, no. 1, pp. 40–59, 2016.
[24] C. A. Knifsend and S. Graham, "Too much of a good thing? How breadth of extracurricular participation relates to school-related affect and academic outcomes during adolescence," *J Youth Adolesc*, vol. 41, pp. 379–389, 2012.
[25] C. A. Kauffman, M. Derazin, A. Asmar, and J. D. Kibble, "Relationship between classroom attendance and examination performance in a second-year medical pathophysiology class," *Adv Physiol Educ*, vol. 42, no. 4, pp. 593–598, 2018.
[26] A. S. N. Kim, S. Shakory, A. Azad, C. Popovic, and L. Park, "Understanding the impact of attendance and participation on academic achievement.," *Scholarsh Teach Learn Psychol*, vol. 6, no. 4, p. 272, 2020.
[27] W. H. Schmidt, N. A. Burroughs, P. Zoido, and R. T. Houang, "The role of schooling in perpetuating educational inequality: An international perspective," *Educational researcher*, vol. 44, no. 7, pp. 371–386, 2015.
[28] L. Elenbaas and M. Killen, "Children's perceptions of social resource inequality," *J Appl Dev Psychol*, vol. 48, pp. 49–58, 2017.
[29] R. Masa'deh, I. AlHadid, E. Abu-Taieh, S. Khwaldeh, A. Alrowwad, and R. S. Alkhawaldeh, "Factors influencing students' intention to use E-textbooks and their impact on academic achievement in Bilingual environment: An empirical study Jordan," *Information*, vol. 13, no. 5, p. 233, 2022.
[30] J. Oakes and M. Saunders, "Access to textbooks, instructional materials, equipment, and technology: Inadequacy and inequality in California's public schools," 2002.
[31] R. M. Ryan and E. L. Deci, "Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions," *Contemp Educ Psychol*, vol. 61, p. 101860, 2020.
[32] E. J. De Bruin, C. van Run, J. Staaks, and A. M. Meijer, "Effects of sleep manipulation on cognitive functioning of adolescents: A systematic review," *Sleep Med Rev*, vol. 32, pp. 45–57, 2017.
[33] J. C. Lo, J. A. Groeger, G. H. Cheng, D.-J. Dijk, and M. W. L. Chee, "Self-reported sleep duration and cognitive performance in older adults: a systematic review and meta-analysis," *Sleep Med*, vol. 17, pp. 87–98, 2016.