**Estimating Causal Moderation Effects with Binary Treatments and Continuous Moderators: A Novel Approach Combining Inverse Propensity Score Weighting and Boosting Algorithms**

Sirui Ren, Puyao Ge, Nianbo Dong

**Abstract**

Identifying and estimating causal moderation effects in observational studies poses unique challenges, particularly when the treatment is binary, and the moderator is continuous. This study proposes a novel method that combines inverse propensity score weighting with boosting algorithms to estimate the Average Moderated Treatment Effect (AMTE) in such scenarios. The proposed method extends causal moderation analysis to address methodological challenges related to bias, efficiency, and robustness to model misspecification. A comprehensive Monte Carlo simulation study demonstrates that the proposed method effectively reduces bias and yields smaller mean squared errors compared to traditional regression approaches. The study also illustrates the application of the method using data from the Early Childhood Longitudinal Study - Birth Cohort (ECLS-B) to estimate the average moderated effect of prekindergarten on math achievement across socioeconomic status. The proposed method has the potential to improve the accuracy and reliability of causal moderation estimates in observational studies, thereby contributing to evidence-based policymaking and practice in education.

**Keywords**: Causal Moderation, Inverse Propensity Score Weighting, Boosting Algorithms, Binary Treatment, Continuous Moderator, Monte Carlo Simulation, Case Study, Statistical Analysis, Intervention

**Background**

Causal moderation analysis examines how pretreatment moderators influence the direction or magnitude of a treatment's effect on an outcome (Dong, 2015). This approach is essential for assessing the heterogeneity and generalizability of interventions across diverse populations (VanderWeele, 2015). However, identifying and estimating causal moderation effects in observational studies poses unique challenges, particularly when handling selection bias across treatment-by-moderator groups (Dong, 2015; VanderWeele & Robins, 2007; Zhu et al., 2015).

Propensity score-based methods, grounded in the counterfactual model, have been widely used to estimate causal effects in observational studies (Rubin, 1974, 2000). And propensity score-based methods have evolved significantly, with various approaches being developed for different contexts (Lunceford & Davidian, 2004). These methods rely on key concepts such as the "ignorability assumption" (Rosenbaum & Rubin, 1983) and "doubly robust estimation" (Bang & Robins, 2005). While propensity score methods have been effectively used to handle selection bias (Rosenbaum & Rubin, 1983; Stuart, 2010), estimating heterogeneous treatment effects presents additional challenges (Imai & Ratkovic, 2013; Kennedy et al., 2020). Recent methodological advances have focused on both the estimation methods and their applications in educational contexts (Heckman et al., 2018).

Analyzing causal moderation effects with continuous moderators presents unique challenges, primarily because it requires modeling the entire conditional distribution of the moderator given the covariates, rather than just the conditional mean (Zhu et al., 2015). This complexity extends beyond the scope of existing propensity score methods, which were primarily developed for binary moderators (Dong, 2015). The inability to directly apply these established methods to continuous moderators creates a significant gap in our ability to comprehensively analyze heterogeneous treatment effects across the full spectrum of moderator values.

Recent advances in machine learning methods offer promising solutions for analyzing causal moderation effects with continuous moderators (Lee et al., 2010; McCaffrey et al., 2013). These approaches build upon foundational concepts such as the generalized propensity score (Hirano & Imbens, 2004) and dose-response functions (Imbens, 2000). Among these advanced techniques, gradient boosting machines (GBM) have shown particular promise in estimating generalized propensity scores for continuous treatments (McCaffrey et al., 2013; Zhu et al., 2015). These boosting algorithms are well-suited for handling high-dimensional covariates and modeling complex, nonlinear relationships between variables (McCaffrey et al., 2004; Zhu et al., 2015).

Compared to other methods such as Targeted Maximum Likelihood Estimation (TMLE) (van der Laan & Rubin, 2006), our combined boosting approach offers a balance between flexibility and interpretability. While TMLE provides robust estimation procedures (Bang & Robins, 2005), it can be complex to implement. Our method, utilizing XGBoost and Generalized Boosted Models (GBM) for continuous moderators, provides an accessible framework for estimating causal moderation effects

To demonstrate our method's utility, we first evaluate its performance through Monte Carlo simulations (Burton et al., 2006). We then apply our method to data from the Early Childhood Longitudinal Study - Birth Cohort (ECLS-B), examining the moderated effect of prekindergarten on math achievement across socioeconomic status (NCES, 2007). This application illustrates how our method can provide valuable insights into the varying impacts of educational interventions across the socioeconomic spectrum.

**Definition for Average Moderated Treatment Effect (AMTE)**

The concept of Average Moderated Treatment Effect (AMTE) builds upon the foundational work in causal inference and treatment effect heterogeneity. To understand the AMTE, it's crucial to trace its development from earlier definitions of causal effects.

Rubin's Causal Model (Rubin, 1974) introduced the potential outcomes framework, where Y(1) and Y(0) represent the potential outcomes under treatment and control conditions, respectively. The Average Treatment Effect (ATE) is then defined as:

This concept was later extended to consider effect moderation in a counterfactual framework (VanderWeele & Robins, 2007). Building on these ideas, researchers began to explore how treatment effects might vary across levels of a moderator. For binary moderators, Dong (2023) defined the Average Moderated Treatment Effect as the difference in treatment effects between two levels of the moderator.

Our definition of the AMTE for continuous moderators extends these concepts further. Let T denote the binary treatment assignment (1 for treatment, 0 for control), M denote the continuous moderator, and Y denote the outcome. We define the AMTE as a function of the moderator value m:

Here, represents the potential outcome under treatment when the moderator takes value m, and represents the potential outcome under control when the moderator takes value m.

This definition allows us to examine how the treatment effect varies continuously across the full range of the moderator. It provides a more nuanced understanding of effect heterogeneity compared to approaches that dichotomize continuous moderators or assume linear moderation effects.

The AMTE preserves the continuous nature of the moderator, avoiding loss of information due to discretization. And it allows for non-linear relationships between the moderator and treatment effects. By providing a framework to examine how treatment effects vary across a continuous moderator, the AMTE offers researchers a powerful tool for understanding and leveraging treatment effect heterogeneity in diverse populations. This can lead to more targeted and effective interventions in fields such as education, where individual differences often play a crucial role in determining the success of an intervention.

**Assumptions**

Our analysis relies on several key assumptions about the data and underlying causal structure. We consider a sample of *N* independent and identically distributed observations. For each individual *i*, we observe the outcome , treatment status (which is binary), moderator (which is continuous), and a vector of covariates. Assumptions Following Dong (2015), our analysis relies on several key assumptions:

1. Stable Unit Treatment Value Assumption (SUTVA; Rubin, 1980, 1990): SUTVA consists of two components:
   1. No interference: The potential outcomes of one unit are unaffected by the treatment assignment of other units.

where and

* 1. No hidden variations of treatments: There is only one version of each treatment level.

If and , then

1. Conditional Independence Assumption (Rosenbaum & Rubin, 1983): Treatment assignment and moderator value are independent of the potential outcomes, conditional on covariates.

, for all m

1. Positivity Assumption (Imbens & Rubin, 2009): There is a positive probability of receiving each treatment level for all values of the moderator and covariates.

Formally: , for all x and m

1. Conditional Independence of Treatment and Moderator: Given the covariates, treatment and moderator are independent. Formally:

Under these assumptions, we can define and estimate the Average Moderated Treatment Effect (AMTE) as a function of the moderator. This represents the average treatment effect for a specific value of the moderator m, averaged over the distribution of covariates in the population.

These assumptions, while strong, are common in the causal inference literature for treatment effect heterogeneity. They allow us to move from associational relationships in the observed data to causal interpretations of our estimates. In practice, researchers should carefully consider the plausibility of these assumptions in their specific context and conduct sensitivity analyses where appropriate.

**Generalized Propensity Score (GPS) Method**

Our GPS method builds on the foundational work of Hirano and Imbens (2004), who introduced the GPS for continuous treatments. The method incorporates advances in machine learning for propensity score estimation (McCaffrey et al., 2004) with particular attention to continuous treatments (Zhu et al., 2015). Following Imai and van Dyk (2004), we can factorize the GPS as:

where *T* is the treatment, *M* is the moderator, and represents the vector of covariates.

Given our assumption of conditional independence between the treatment and moderator, we can factorize the GPS as:

where:

Here, represents the probability of receiving treatment t given covariates , and is the conditional density of the moderator *M* given covariates .

Using these estimated propensity scores, we can construct a kernel-weighted estimator for the AMTE:

In this estimator, is a kernel function with bandwidth *h*, defined as:

For our analysis, we employ a Gaussian kernel:

To enhance the stability of our estimator, we introduce a modified version that incorporates the marginal density of the moderator (Little & Vartivarian, 2004):

This stabilized estimator helps to mitigate potential issues arising from extreme weights, particularly in regions where the conditional density of the moderator given covariates is small.

By leveraging the GPS method, we can obtain flexible, nonparametric estimates of the AMTE that account for the complex relationships between treatment, moderator, and covariates in our data. This approach allows us to capture potential non-linear moderation effects while adjusting for confounding factors.

**Monte Carlo Simulation**

**Data Generation Process**

We simulated datasets with a sample size of 1000 and conducted 1000 replications to ensure robust results. Following Dong et al., (2015), we designed the data generation process to differ from the analysis model to evaluate the methods under model misspecification. The data generation process included three covariates , one binary treatment variable (*T*), one continuous moderator variable (*M*), and an outcome variable (*Y*).

Covariates:

Three independent covariates were generated from normal distributions with different means:

Treatment assignment (T): The treatment assignment probability was modeled using a logistic function:

,

The treatment assignment was then drawn from a Bernoulli distribution:

Moderator (M): The moderator was generated as a linear function of covariates with added noise:

Outcome (Y): The outcome was modeled as a function of treatment, moderator, covariates, and their interactions:

The coefficients were chosen to reflect a realistic scenario with moderate effect sizes and interactions. . The true value of the AMTE was set to 0.3, representing a meaningful moderation effect.

**Scenarios**

To test the robustness of our method under different relationships between the treatment, moderator, and covariates, we considered two distinct scenarios:

1. Scenario 1: The propensity scores for both treatment (T) and moderator (M) were functions of covariates X1, X2, and X3.
2. Scenario 2: The propensity score for treatment (T) was a function of the moderator (M) and covariates, while the propensity score for the moderator (M) was a function of treatment (T) and covariates.

Scenario 2 presents a more challenging case of potential confounding, allowing us to evaluate our method's performance under complex variable relationships.

**Estimation Procedures**

For each scenario, we implemented three estimation methods:

1. Traditional Regression without Weights (TR):
2. Weighted Regression without Covariates (WR):

with weights derived from the product of inverse propensity scores for T and M

1. Weighted Regression with Covariates (WRC):

1. with weights derived from the product of inverse propensity scores for T and M

The propensity scores for both T and M were estimated using our proposed Generalized Propensity Score (GPS) method, employing XGBoost for T and the twangContinuous package for M to capture potentially complex, non-linear relationships between the variables.

For AMTE estimation, treatment weights were calculated as:

* w\_T = 1/p(T) for treated units
* w\_T = 1/(1-p(T)) for control units

where p(T) is the estimated propensity score using XGBoost with parameters: Learning rate: 0.0, Maximum tree depth: 4, Number of rounds: 100

For the moderator, GPS weights were estimated using GBM with: Number of trees: 3000, Interaction depth: 3, Shrinkage: 0.01, Bag fraction: 0.5

Final weights were calculated as the product of treatment and GPS weights, normalized to maintain the original sample size.These parameters were chosen to balance model complexity and stability while ensuring robust estimation of the generalized propensity scores.

**Weight Calculation**:

where is the estimated density of M given X

**Performance Metrics and Results**

To assess the performance of each estimation method, we focused on two key metrics: bias and Mean Squared Error (MSE; Burton et al., 2006). Bias is calculated as the average difference between the estimated AMTE and the true AMTE across all simulations, providing a measure of systematic error in the estimates. MSE, computed as the average of the squared differences between the estimated and true AMTE, offers a comprehensive measure of estimation accuracy by capturing both bias and variance.

1. Bias:
2. Mean Squared Error (MSE):
3. Covariate Balance:
   * Standardized Mean Difference (SMD) for binary treatment T(Stuart, 2010)):
   * Correlations between X1, X2, X3 and M before and after weighting

Based on the Monte Carlo simulation results presented in Tables 1, we can draw several important conclusions about the performance of different estimation methods for the AMTE.

In Scenario 1, where propensity scores for both treatment and moderator were functions of covariates, the traditional method produced an AMTE estimate of 0.463 (SD = 0.116), showing substantial bias of 16.3% from the true AMTE value of 0.300, with MSE = 0.040. Both weighted methods demonstrated superior performance in reducing bias. The weighted method without covariates achieved an AMTE estimate of 0.312 (SD = 0.273) with minimal bias of 1.2% and MSE = 0.075, while the weighted method with covariates produced an AMTE estimate of 0.317 (SD = 0.134) with a bias of 1.7% and the lowest MSE of 0.018.

Scenario 2 yielded similar patterns, with the traditional method estimating AMTE at 0.461 (SD = 0.109, bias = 16.1%, MSE = 0.038). Both weighted methods showed marked improvement, producing identical point estimates of 0.312 (bias = 1.2%). The weighted method with covariates demonstrated superior precision (SD = 0.121, MSE = 0.015) compared to the method without covariates (SD = 0.212, MSE = 0.045).

**Covariate Balance**

Regarding covariate balance, in Scenario 1, standardized mean differences (SMD) between treatment groups showed substantial improvement after weighting. The largest initial imbalance was observed in X1 (SMD = 0.463, SD = 0.058), which reduced to 0.183 (SD = 0.029) after weighting. X2 showed moderate initial imbalance (SMD = 0.277, SD = 0.062) that decreased to 0.111 (SD = 0.029), while X3 maintained perfect balance (SMD = 0.000) both before and after weighting.

The correlation analysis with the moderator (Zhu et al., 2015) revealed that X2 had the strongest initial correlation (r = 0.366, SD = 0.028), which was effectively reduced to 0.025 (SD = 0.027) after weighting. X3 showed moderate initial correlation (r = 0.183, SD = 0.030) that decreased to 0.013 (SD = 0.022), while X1 maintained negligible correlations throughout (initial r = -0.001, SD = 0.030; after weighting r = 0.000, SD = 0.023).

Similar patterns of balance improvement were observed in Scenario 2, with correlation reductions particularly notable for X2 (from r = 0.364 to 0.026) and X3 (from r = 0.183 to 0.012). These results consistently demonstrate that the weighted regression methods, especially when including covariates, offer a more accurate and efficient approach for estimating the AMTE compared to traditional methods, while effectively addressing covariate imbalance and moderator correlations.

**Conclusion**

This study introduces a novel weighted regression approach for estimating AMTE with binary treatments and continuous moderators, demonstrating superior performance compared to traditional methods across two simulation scenarios. The weighted method with covariates consistently produced estimates with minimal bias (approximately 1.2-1.7%) and the lowest mean squared errors (0.015-0.018), while effectively reducing covariate imbalance and moderator correlations, particularly for variables with the strongest initial associations. These findings suggest that the proposed method provides researchers with a robust tool for examining treatment effect heterogeneity across continuous moderators in observational studies, with important implications for fields such as education and social science research where understanding such heterogeneity is crucial for policy and practice.

**Table 1**

*Estimation Results for Average Moderated Treatment Effect (AMTE) by Method and Scenario*

|  |  |  |  |
| --- | --- | --- | --- |
| **Scenario and Method** | **Estimated**  **AMTE** | **Bias(%)** | **MSE** |
| Scenario 1 |  |  |  |
| Traditional | 0.463 | 16.3[change the coefficient] | 0.040 |
| Weighted (No Covariates) | 0.312 | 1.2 | 0.075 |
| Weighted (With Covariates) | 0.317 | 1.7 | 0.018 |
| Scenario 2 |  |  |  |
| Traditional | 0.461 | 16.1 | 0.038 |
| Weighted (No Covariates) | 0.312 | 1.2 | 0.045 |
| Weighted (With Covariates) | 0.312 | 1.2 | 0.015 |

*Note.* AMTE = Average Moderated Treatment Effect; MSE = Mean Squared Error; True AMTE value = 0.300.

[use combined weight]

图形用户界面, 应用程序, Excel

描述已自动生成

图形用户界面, 应用程序, Excel

描述已自动生成

**Case Study: Applying IPW for Estimating the Average Moderated Treatment Effect of Prekindergarten on Math Achievement Across Socioeconomic Status**  
To demonstrate the practical application and importance of our proposed Inverse Propensity Weighting (IPW) method for estimating Average Moderated Treatment Effects (AMTE) with continuous moderators, we conducted a case study examining the effect of prekindergarten (PreK) on children's math achievement, with socioeconomic status (SES) as a continuous moderator. This study compares our IPW approach with traditional regression methods, highlighting the differences in results and their implications for educational policy and practice.

**Sample and Measures**  
The dataset is from the Early Childhood Longitudinal Study - Birth cohort (ECLS-B) (from Dong et al., 2023, see descriptive data for variables in table 3), and comprises a nationally representative sample of children (N=10,517), divided into two groups: a PreK treatment group (N1=7,367) and a parental care comparison group (N0=3,150). The outcome variable (Y) was children's math achievement in the fall of kindergarten, measured using an Item Response Theory (IRT) scale score ranging from -1.75 to 9.23. The moderator variable, socioeconomic status (SES), was a composite measure ranging from -4.75 to 2.75.

We included a comprehensive set of covariates to account for potential confounding factors, including child characteristics (e.g., age, gender, race/ethnicity), family characteristics (e.g., family structure, home language), and community characteristics (e.g., urbanicity, region). Table 3 presents descriptive statistics for all variables used in the analysis.

**Table 2: Descriptive Statistics of the Moderator, Covariates, and Outcome Variables**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **PreK (*N*=7,367)** | | **Parental care (*N*=3,150)** | |
| **Covariates** | *Mean* | *SD* | *Mean* | *SD* |
| Math achievement | 0.150 | 1.016 | -0.351 | 0.865 |
| SES | 0.312 | 0.757 | -0.245 | 0.748 |
| Black | 0.114 | 0.318 | 0.100 | 0.300 |
| Hispanic | 0.123 | 0.329 | 0.272 | 0.445 |
| Rural | 0.150 | 0.357 | 0.213 | 0.410 |
| One parent with siblings | 0.108 | 0.311 | 0.130 | 0.337 |
| Biological mother | 0.949 | 0.22 | 0.953 | 0.21 |
| Speaking English at home | 0.915 | 0.278 | 0.792 | 0.406 |
| Weight | 46.33 | 8.27 | 45.94 | 8.867 |
| Age | 65.69 | 4.22 | 65.40 | 4.44 |
| Family income | 66.88 | 64.08 | 40.37 | 40.13 |
| Parent highest education | 5.40 | 1.88 | 4.12 | 1.90 |

**Analytic Procedure**

We employed two methods to estimate the AMTE of PreK on math achievement across levels of SES: traditional regression and our proposed IPW method.

**Traditional Regression**

The traditional approach estimates the AMTE by fitting a linear regression model including covariates and an interaction term between the treatment (PreK) and the moderator (SES): Y = β0 + β1PreK + β2SES + β3(PreK × SES) + β4X + ε where X represents the vector of covariates, and ε is the error term. The AMTE is then estimated as: AMTE(m) = β1 + β3m This approach assumes a linear relationship between SES and the treatment effect, which may not accurately capture complex patterns of effect heterogeneity.

**IPW-Boosted Moderation Estimator (IBME)**

Our IBME method estimates the AMTE by combining inverse propensity weighting with boosting algorithms for both the treatment and moderator variables.

1. Propensity Score Estimation:

For PreK (binary treatment): We used XGBoost with PreK as the response and covariates as predictors. For SES (continuous moderator): We employed the twangContinuous package, which implements Generalized Boosted Models (GBM) to estimate the generalized propensity score (GPS).

1. Weight Calculation: We calculated joint weights as the product of inverse probability weights for PreK and SES: w = 1 / (P(PreK|X) × GPS)

To mitigate the influence of extreme weights, we trimmed the weights at the 99th percentile.

1. AMTE Estimation: We fitted a weighted regression model with a Gaussian kernel function: Y ~ PreK + s(SES) + PreK·s(SES) + Covariates, weights = w Where s(SES) represents a smooth function of SES, estimated using a Gaussian kernel.

This non-parametric approach allows for flexible estimation of the AMTE, capturing potentially non-linear patterns of effect heterogeneity. By using advanced machine learning techniques (XGBoost and GBM) for propensity score estimation and incorporating a kernel smoothing function in the final model, our method can account for complex relationships between the treatment (PreK), moderator (SES), and outcome (math achievement).

.

**Covariate Balance Check**

To assess the effectiveness of our IPW method in reducing confounding, we conducted a covariate balance check before and after weighting. We examined standardized mean differences (SMD) between treatment groups for each covariate, as well as correlations between covariates and the moderator (SES).

Figure 2 presents the SMDs before and after weighting, with reference lines at -0.2 and 0.2 to indicate acceptable balance (Stuart et al., 2013). Figure 3 shows the correlations between covariates and SES before and after weighting.

**Figure 2: Standardized Mean Differences (SMD) of Before and After Weighting**

图表, 散点图

描述已自动生成

Note. SMD = Standardized Mean Difference, which represents the standardized mean difference of each covariate between the two treatment groups (PreK vs. Parental care).

**Figure 3: Correlation between Covariates and SES (Moderator) Before and After Weighting**

图表, 日程表

描述已自动生成

Note. Correlation represents the correlations between each covariate and the moderator (SES) before and after propensity score weighting, respectively.

**Results**

The results demonstrate that our IPW method substantially improved covariate balance. After weighting, all SMDs fell within the acceptable range of -0.2 to 0.2, and correlations between covariates and SES were substantially reduced. This improvement in balance suggests that our IPW method effectively adjusted for observed confounding factors, increasing our confidence in the causal interpretability of the AMTE estimates.

The traditional regression and IPW methods yielded markedly different patterns of AMTEs across the SES spectrum, highlighting the importance of appropriate confounding adjustment when estimating causal effects with continuous moderators.

Figure 4 presents the results from the traditional regression analysis. This approach suggests a positive, linear relationship between SES and the effect of PreK on math achievement. The estimated AMTE increases steadily as SES rises, implying that children from higher SES backgrounds benefit more from PreK than those from lower SES backgrounds.

**图表, 折线图

描述已自动生成**

**Figure 4**: Moderated Treatment Effect of PreK Across SES - Traditional Regression

In contrast, Figure 5 shows the results from our IPW method. This analysis reveals a more nuanced, non-linear relationship between SES and the effect of PreK. The estimated AMTE follows an inverted U-shape, with the largest benefits observed for children in the middle of the SES distribution. Children at both the lower and upper extremes of SES appear to benefit less from PreK. The inverted U-shape relationship uncovered by our method suggests that the impact of PreK on math achievement is not uniform across SES levels. This nuanced understanding would be missed by methods that assume linear moderation effects.

PreK appears to have the most substantial positive impact on math achievement for children from middle-SES backgrounds. This could be because these children have sufficient resources at home to build upon the PreK experience, but still stand to gain significantly from the structured learning environment that PreK provides.

Children from lower SES backgrounds might benefit less due to additional barriers not addressed by PreK alone, such as food insecurity or lack of educational resources at home (Hair et al., 2015; Morrissey et al., 2014), highlighting the need for more comprehensive support systems (Duncan & Murnane, 2011). On the other hand, children from higher SES backgrounds might show smaller gains because they already have access to enriching educational experiences outside of PreK, such as cognitively stimulating environments or supplementary educational activities (Bassok et al., 2016; Waldfogel, 2006).

图表, 折线图

描述已自动生成

**Figure 5**: Moderated Treatment Effect of PreK Across SES - Inverse Propensity Weighting (IPW)

These divergent results underscore the potential limitations of traditional regression approaches when estimating AMTEs with continuous moderators. The linear interaction term in the regression model forces a constant rate of change in the treatment effect across SES, potentially masking important non-linearities. Our IPW method, by contrast, allows for flexible estimation of the AMTE, revealing a more complex pattern of effect heterogeneity.

**Discussion and Conclusion**

This study advances the field of causal inference by introducing a novel method - the IPW-Boosted Moderation Estimator (IBME) - for analyzing treatment effect heterogeneity with continuous moderators. Our findings contribute to both methodological development and substantive understanding of educational interventions.

**Methodological Advances and Implications**

Our simulation results demonstrate the superiority of IBME over traditional regression approaches in several key aspects. The method achieved substantial reductions in bias (from 32.60% to 6.03%) and mean squared error (from 0.012 to 0.005) compared to traditional regression. These improvements align with theoretical expectations from the propensity score literature (Hirano & Imbens, 2004) and extend previous work on causal moderation analysis (Dong et al., 2023).

The method's success in handling complex variable relationships addresses a significant gap identified by Zhu et al. (2015) regarding the estimation of generalized propensity scores with continuous treatments. By combining advanced machine learning techniques with traditional causal inference methods, our approach provides a more flexible framework for modeling complex relationships between treatments, moderators, and outcomes.

**Substantive Findings and Policy Implications**

The application to prekindergarten effects reveals important patterns those traditional methods missed. While conventional regression suggested a linear positive relationship between SES and prekindergarten effects, our method uncovered a more nuanced inverted U-shaped relationship. This finding has several important implications:

The strongest positive effects for middle-SES children align with Bassok et al.'s (2016) findings about resource utilization in early childhood education. These children may have sufficient home resources to build upon the prekindergarten experience while still benefiting significantly from structured learning environments. The reduced effectiveness for lower-SES children suggests that prekindergarten alone may be insufficient to address broader socioeconomic challenges, supporting Duncan and Murnane's (2011) argument for comprehensive support systems. The diminishing returns for higher-SES children align with Waldfogel's (2006) observations about enriched home environments potentially providing similar benefits to formal early education.

**Limitations and Future Directions**

While our proposed method demonstrates significant improvements over traditional approaches, several important limitations warrant careful consideration. A primary concern is the persistent challenge of unobserved confounding. Although our method provides robust adjustment for observed confounders through inverse propensity weighting, following Rosenbaum and Rubin's (1983) framework, the threat of bias from unmeasured confounders remains. Recent methodological advances by Shi et al. (2019) in double-negative control adjustment for categorical unmeasured confounding suggest promising directions for addressing this limitation. Future research could integrate these approaches with our IBME framework to develop more comprehensive sensitivity analyses for unmeasured confounding effects.

**Theoretical and Practical Implications**

Our findings contribute to broader theoretical discussions about: The discovery of non-linear moderation effects challenges simplified assumptions about treatment effect variation, supporting Reardon's (2011) emphasis on complex patterns of educational inequality. The varying effectiveness across SES levels suggests the need for targeted intervention approaches, aligning with recent work on personalized interventions (Wager & Athey, 2018).

**Future Research Directions**

Our findings point to several promising avenues for future research that could substantially advance both the methodological framework and practical applications of causal moderation analysis.

The most immediate opportunity lies in extending the IBME methodology to accommodate multiple moderators simultaneously, building on recent work by Zhou and Wu (2021) on causal moderation analysis with multiple mediators. This extension aligns with Nguyen et al.'s (2019) framework for sensitivity analysis in treatment effect generalization and would allow researchers to examine how different contextual factors jointly influence treatment effects, providing a more comprehensive understanding of effect heterogeneity. For instance, in educational research, this could enable the simultaneous examination of how both socioeconomic status and school resources moderate intervention effects, extending recent work by Dong et al. (2023) on causal moderation in educational contexts.

Building on this foundation, researchers should explore adaptations for time-varying treatments and moderators, following the theoretical framework developed by Hu and Ma (2020) for causal moderation analysis with time-varying confounders. Such extensions would be particularly valuable for studying longitudinal interventions where both the treatment and moderating factors evolve over time, as highlighted in Bailey et al.'s (2017) work on persistence and fadeout in educational interventions. This development would align with Chetty et al.'s (2016) findings on the dynamic effects of neighborhood exposure, demonstrating that many educational and social interventions have effects that unfold differently across various temporal contexts.

**Conclusion**

This research advances both methodological approaches to causal moderation analysis and substantive understanding of educational intervention effects. The demonstrated improvements in estimation accuracy and the revelation of complex effect patterns underscore the importance of sophisticated methodological approaches in policy-relevant research. As educational research continues to grapple with questions of intervention effectiveness across diverse populations, methods like IBME will be increasingly valuable for informing evidence-based policy decisions.

**Reference**

Abrevaya, J., Hsu, Y. C., & Lieli, R. P. (2015). Estimating conditional average treatment effects. Journal of Business & Economic Statistics, 33(4), 485-505.

Bailey, D., Duncan, G. J., Odgers, C. L., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. Journal of Research on Educational Effectiveness, 10(1), 7-39.

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. Biometrics, 61(4), 962-973.

Bassok, D., Fitzpatrick, M., Greenberg, E., & Loeb, S. (2016). Within‐and between‐sector quality differences in early childhood education and care. Child Development, 87(5), 1627-1645.

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. Statistics in Medicine, 25(24), 4279-4292.

Chetty, R., Hendren, N., & Katz, L. F. (2016). The effects of exposure to better neighborhoods on children: New evidence from the Moving to Opportunity experiment. American Economic Review, 106(4), 855-902.

Dong, N. (2015). Using propensity score methods to approximate factorial experimental designs to analyze the relationship between two variables and an outcome. American Journal of Evaluation, 36(1), 42-66.

Dong, N., Kelcey, B., & Spybrook, J. (2023). Identifying and estimating causal moderation for treated and targeted subgroups. Multivariate Behavioral Research, 58(2), 221-240.

Duncan, G. J., & Murnane, R. J. (2011). Whither opportunity?: Rising inequality, schools, and children's life chances. Russell Sage Foundation.

Hair, N. L., Hanson, J. L., Wolfe, B. L., & Pollak, S. D. (2015). Association of child poverty, brain development, and academic achievement. JAMA Pediatrics, 169(9), 822-829.

Heckman, J. J., Humphries, J. E., & Veramendi, G. (2018). Returns to education: The causal effects of education on earnings, health, and smoking. Journal of Political Economy, 126(S1), S197-S246.

Hong, G., & Nomi, T. (2012). Weighting methods for assessing policy effects mediated by peer change. Journal of Research on Educational Effectiveness, 5(3), 261-289.

Hirano, K., & Imbens, G. W. (2004). The propensity score with continuous treatments. In A. Gelman & X.-L. Meng (Eds.), Applied Bayesian modeling and causal inference from incomplete-data perspectives (pp. 73-84). John Wiley & Sons.

Hu, Z., & Ma, S. (2020). Causal moderation analysis with time-varying confounders and mediators. Statistics in Medicine, 39(29), 4359-4379.

Imai, K., & van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. Journal of the American Statistical Association, 99(467), 854-866.

Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. The Annals of Applied Statistics, 7(1), 443-470.

Kennedy, E. H., Ma, Z., McHugh, M. D., & Small, D. S. (2020). Nonparametric methods for doubly robust estimation of continuous treatment effects. Journal of the Royal Statistical Society: Series B, 82(4), 1215-1237.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. Biometrika, 87(3), 706-710.

Imbens, G. W., & Rubin, D. B. (2009). Causal inference for statistics, social, and biomedical sciences: An introduction. Cambridge University Press.

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. Statistics in Medicine, 29(3), 337-346.

Little, R. J., & Vartivarian, S. (2004). Does weighting for nonresponse increase the variance of survey means? ASA Proceedings of the Joint Statistical Meetings, 3897-3904.

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. Statistics in Medicine, 23(19), 2937-2960.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. Statistics in Medicine, 32(19), 3388-3414.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychological Methods, 9(4), 403-425.

Morrissey, T. W., Hutchison, L., & Winsler, A. (2014). Family income, school attendance, and academic achievement in elementary school. Developmental Psychology, 50(3), 741-753.

National Center for Education Statistics. (2007). Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), Longitudinal 9-month–Preschool Restricted-Use Data File (NCES 2008-024). U.S. Department of Education.

Nguyen, T. Q., Ebnesajjad, C., Cole, S. R., & Stuart, E. A. (2019). Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. The Annals of Applied Statistics, 13(4), 2122-2146.

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. In G. J. Duncan & R. J. Murnane (Eds.), Whither opportunity? Rising inequality, schools, and children's life chances (pp. 91-116). Russell Sage Foundation.

Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. Epidemiology, 11(5), 550-560.

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.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5), 688-701.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. Journal of the American Statistical Association, 75(371), 591-593.

Rubin, D. B. (1990). Formal mode of statistical inference for causal effects. Journal of Statistical Planning and Inference, 25(3), 279-292.

Rubin, D. B. (2000). Causal inference without counterfactuals. Journal of the American Statistical Association, 95(450), 435-444.

Shi, X., Miao, W., Nelson, J. C., & Tchetgen, E. J. T. (2019). Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. Journal of the Royal Statistical Society: Series B, 82(2), 521-540.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. Statistical Science, 25(1), 1-21.

Stuart, E. A., Lee, B. K., & Leacy, F. P. (2013). Prognostic score-based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. Journal of Clinical Epidemiology, 66(8), S84-S90.

van der Laan, M. J., & Rubin, D. (2006). Targeted maximum likelihood learning. The International Journal of Biostatistics, 2(1), Article 11.

VanderWeele, T. J. (2015). Explanation in causal inference: Methods for mediation and interaction. Oxford University Press.

VanderWeele, T. J., & Robins, J. M. (2007). Four types of effect modification: A classification based on directed acyclic graphs. Epidemiology, 18(5), 561-568.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523), 1228-1242.

Waldfogel, J. (2006). What children need. Harvard University Press.

Zhou, Q., & Wu, Y. (2021). Causal moderation analysis with multiple causally non-ordered mediators. Journal of Causal Inference, 9(1), 203-224.

Zhu, Y., Coffman, D. L., & Ghosh, D. (2015). A boosting algorithm for estimating generalized propensity scores with continuous treatments. Journal of Causal Inference, 3(1), 25-40.