# Abstract Title Page

*Not included in page count.*

## Title:

Causal Moderation Analysis Using Propensity Score Methods

## Authors and Affiliations:

Nianbo Dong
Peabody Research Institute
Vanderbilt University

nianbo.dong@vanderbilt.edu

# Abstract Body

## Background / Context:

Randomized experiments are often used to estimate the *overall* causal effects of an intervention (Boruch, 1997). In addition to the *overall* effects, policy makers and researchers have increased interest in exploring the "black box" of the interventions, that is, the intervention mechanism, such as: (1) the mediator, through which, the intervention works to improve the outcome, and (2) the moderator, by which, the intervention works differently (i.e., treatment effect heterogeneity[1]). The mediation and moderation analysis (Baron & Kenny, 1986) have been widely applied in educational research. Researchers have paid particular attention to the causal mediation analysis (e.g., Imai, Keele, & Tingley, 2010; Raudenbush, 2011). However, fewer studies examined the causal moderation analysis.

The conventional moderation analysis is "within a correlational analysis framework" as stated by Baron and Kenny (1986):

> *In general terms, a moderator is a qualitative (e.g., sex, race, class) or quantitative (e.g., level of reward) variable that affects the direction and/or strength of the relation between an independent or predictor variable and a dependent or criterion variable.*
>
> *Specifically within a correlational analysis framework, a moderator is a third variable that affects the zero-order correlation between two other variables. (p. 1174)*

The moderator effect can be represented by an interaction between the treatment variable (predictor) and moderator in the statistical models (Baron & Kenny, 1986), such as ANOVA (analysis of variance), ANCOVA (analysis or covariance), or ordinary least square (OLS) regression for single level analysis, and HLM (hierarchical linear model) for multilevel analysis (Bauer & Curran, 2005; Raudenbush & Bryk, 2002).

The ideal situation for making causal inference of moderator effect would be that both the treatment variable and moderator can be manipulated under double randomization, such that neither the treatment variable nor moderator is correlated with any covariates (measured or unmeasured). Hence, the moderator effects can be unbiasedly estimated, and will not depend on the model specification. Otherwise, when the moderator is correlated with covariates and the statistical model is misspecified (e.g., lack of higher order of the covariates), the coefficients of the moderator and the interaction will be estimated with bias.

Recently, researchers realized this issue and applied propensity score methods (Rosenbaum & Rubin, 1983) for subgroup analysis (see Hill, Brooks-Gunn, & Waldfogel, 2003; Lochman, Boxmeyer, Powell, Roth, & Windle, 2006; Peck, 2003; Schochet & Burghardt, 2007). For example, when the moderator is a binary variable, the basic practice is to match participants between the treatment and control groups within each of two levels of moderator. These separate matches may make baseline equivalent between the treatment and control groups for each level of moderator, hence it may produce unbiased estimates of treatment effects at that level of moderator. However, without making efforts to make baseline equivalent among all four groups, it may not produce unbiased estimate of the interaction effect (i.e., whether the difference in treatment effects between two levels of moderator is solely due to the moderator).

Other researchers, for example, Imai & van Dyk (2004), have generalized propensity score

[1] Treatment effect heterogeneity has multiple forms. Moderator effect is analogous to the *fixed* block effect in randomized block designs (or multisite experiments). There could be *random* block effects in randomized block designs.

methods to study the main effects of bivariate treatment variables (two continuous treatment variables) using subclassification. Dong (2011) examined the various propensity score applications (e.g., stratification, Imai & van Dyk, 2004; inverse of propensity score weighting, Imbens, 2000, and matching, etc.) in analyzing the main and interaction effects of two binary factors through Monte Carlo simulation. Note that although Imai & van Dyk (2004) or Dong (2011) did not explicitly claim that those propensity score methods were for moderation analysis, we could still apply these methods to make causal inference of moderator effect. Furthermore, Dong (2011) examined the effects of two categorical factors, however, the moderator could be a categorical or continuous variable.

**Purpose / Objective / Research Question / Focus of Study:**

This paper is based on previous studies in applying propensity score methods to study multiple treatment variables (e.g., Dong, 2011; Imai & van Dyk, 2004) to examine the causal moderator effect. The propensity score methods will be demonstrated in a case study to examine the causal moderator effect, where the moderators are categorical and continuous variables.

**Significance / Novelty of study:**

Moderation analysis is an important approach to examining the treatment effect heterogeneity in intervention studies. It is essential to assure that the treatment heterogeneity is solely due to the moderator. This paper proposes good propensity score applications in causal moderator analysis. The procedure of applying propensity score methods, as illustrated in a case study, will provide suggestions to researchers in conducting causal moderator analysis, such as analysis of the effects of implementation fidelity on treatment effects.

**Research Design:**

We first lay out the causal framework for moderation analysis. We then review and identify applicable propensity score methods. Finally, we use a case study to demonstrate how to use the proposed propensity score methods to conduct causal moderation analysis.

***Causal Framework for Moderation Analysis***

Using the counterfactual model (Holland, 1986; Rosenbaum, 2002; Rubin, 1974), we present the potential outcomes for individuals under the situation that both treatment and moderator are binary variables in Table 1. This is analogous to a $2 \times 2$ factorial experimental design (Shadish, Cook, & Campbell, 2002, p.264). The main treatment effect is the marginal mean difference, Y(1,.) - Y(0,.); the main effect of moderator is the marginal mean difference, Y(.,1) - Y(.,0). The moderator effect (i.e., the difference in treatment effect between two levels of moderator) is [Y(1,1) - Y(0,1)] – [Y(1,0) - Y(0,0)].

Alternatively, this counterfactual model can be illustrated in Figure 1. Note that this path diagram looks similar with the one that Raudenbush (2011) used to illustrate the potential outcome for *mediation* analysis (Case 2: Treatment -> Mediating Treatment -> Y, p. 22), but they are different. In this diagram there is no causal relationship between Treatment and Moderator (i.e., there is no an arrow from Treatment to Moderator), and it is Treatment and Moderator together that cause the outcome (Y). The moderator provides two paths through which the intervention causes different outcomes.  On the contrary, in Raudenbush's (2011) diagram, there is a causal link (arrow) from Treatment to Mediator, which finally causes outcome. Although the mechanism is different between mediation and moderation, we can use the same counterfactual model to make causal inference.

### *Applicable Propensity Score Methods for Causal Moderation Analysis*

Dong (2011) reviews the propensity score methods used for analyzing two treatment variables. Through Monte Carlo simulation, Dong (2011) suggested three good propensity score approaches to reducing bias and mean square error (MSE) of parameter estimates in analyzing two binary factors: (1) inverse of propensity score weighting based on one multinomial propensity score model (Imbens, 2000), (2) subclassification (Imai & van Dyk, 2004) and (3) factorial matching based on two binary propensity score models. We focus on the first two approaches because the third approach is relatively complicated and the sample retained for final analysis is much smaller (Dong, 2011).

The first approach, is to convert $2\times2$ design to $4\times1$ design, i.e., a design having one new treatment variable with four levels (e.g., Y(0,0) as group 1, Y(0,1) as group 2, Y(1,0) as group 3, and Y(1,1) as group 4). Then we use the inverse of the *generalized propensity score* as weight to estimate the effects of multi-valued treatments (Imbens, 2000). The weight is $1/r(T_i = t, X)$, where $T_i$ denotes the treatment that subject *i* actually received, and $r(T_i = t, X)$ denotes the *generalized propensity score* which is the conditional probability of receiving particular treatment *t* given pre-treatment covariate *X*.

The second approach is to use two independent binary logistic regression models to estimate the propensity scores for Treatment and Moderator, respectively. The estimated two propensity score functions are used to subclassify the data into several subclasses. Figure 2 illustrates 3 × 3 subclassification based on two propensity score functions (Imai & van Dyk, 2004). Data are subclassified into three subclasses (lower third, middle third, and upper third) based on each of two propensity score functions, respectively. Each cell of the 3 × 3 table represents a subclass based on two propensity score functions jointly. The overall treatment effect point estimate and its standard error estimate are weighted average across 9 subclasses (Imben & Rubin, 2009). In addition to 3 × 3 subclassification, Imai & van Dyk's (2004) also presented simulation results for 2 × 2 and 4 × 4 subclassification. In general, more subclasses produce less bias.

When the moderator is a continuous variable, the first approach is not applicable, however, the second approach still works. The Gaussian linear regression model can be used to estimate the propensity score of the continuous moderator (Imai & van Dyk, 2004). The other procedures are same (e.g., subclassification based on two propensity score functions).

### *Usefulness / Applicability of Method*

We demonstrate the applications of these two approaches in a case study whose purpose is to estimate the average effects of Head Start program as compared with other center-based care on child reading achievement, and examine if the effects differ by child care quality.

#### *Sample and Measures*

The dataset used for this case study is the Early Childhood Longitudinal Study – Birth cohort (ECLS-B). ECLS-B began as a nationally representative sample of 14,000 children born in 2001 in the United States randomly selected. ECLS-B collects information on a rich array of individual-, household-, teacher-, and child care- level measures. We use the Item Response Theory (IRT) scale score for reading skill measured at kindergarten as outcome variable. We focus on observed global classroom/care quality. Head Start and other center-based care were assessed using the Early Childhood Environmental Rating Scale – Revised (ECERS-R) (Harms, Clifford, & Cryer, 1998). The ECLS-B administered 37 of the 43 items included in the original ECERS-R, covering six subscales: (1) furnishing and display, (2) personal care routines, (3) listening and talking, (4) learning activities, (5) interaction, and (6) program structure. Each of the 37 items was rated on 7-point Likert scales. We use the overall ECERS-R score in our analysis. The covariates used in this study include the mental IRT scale score measured at two years old, reading and math IRT scales measured at pre-kindergarten, gender, race (white, black, Hispanic, and other race), birth weight, age at outcome assessment, English speaking at home, disability status, special education status, health problem, family's SES, mother education, income, poverty level, welfare receipt, home violence, household structure, and one dummy variable indicating the year the child was in kindergarten (2007 vs. 2006).

*Analytic Procedure*

For the purpose of simplification, in this demonstration we use listwise deletion to handle missing data[2]. The original ECERS-R overall score is a continuous variable. We dichotomize it with a median split to indicate high quality and low quality. Both the continuous score and the dichotomous score are used for demonstration. We first estimate the overall effect of Head Start as compared with other center-based care using conventional OLS regression and propensity score methods without including care quality measure and its interaction term with care type. When child care quality is treated as a binary variable, the following approaches are used to examine the moderation effect of care quality on the effect of Head Start:

1. Conventional OLS regression analysis with interaction of care quality (low quality vs. high quality) and care type (Head Start vs. other center-based care).
2. Inverse of propensity score weighting analysis based on one multinomial propensity score model (Imbens, 2000).
3. Subclassification based on two binary propensity score models (Imai & van Dyk, 2004).

When care quality is analyzed as a continuous variable, we use Approach 1 with modification that care quality is a median-centered continuous variable, and Approach 3 with modification that subclassification is based on one binary propensity score model and one Gaussian linear regression model.

*Findings / Results*

We first checked covariate balance by examining the standardized mean difference among comparison groups in all analyses. Because our main interest is in analyzing two factors, we present the results of covariate balance checking for moderator analysis. Tables 2 and 3 present covariate balance checking results among child care type by care quality groups before applying propensity score methods and applying the inverse of propensity score weighting in outcome analysis. Stuart (2007) suggests that the standardized mean difference should ideally be less than 0.25, and a value greater than 0.50 is "particularly problematic". In Table 2, nine out of 22 covariates have the maximum standardized mean difference among four groups larger than 0.50. After weighted by the inverse of propensity score, 20 out of 22 covariates have the maximum standardized mean difference among four groups smaller than 0.25, and the other two covariates have the maximum standardized mean difference among four groups smaller than 0.50. This suggests that the inverse of propensity score weighting can greatly reduce selection bias. Covariance balance was also examined for subclassification method and the covariates are much

[2] The results may contain bias if data are not missing completely at random (Allison, 2001).

more balanced than without using any propensity score methods.

Table 4 presents the overall effect of Head Start as compared with other center-based care on children’s kindergarten reading achievement. Basically, statistically significantly negative effects are found to associate with Head Start in both the conventional OLS and propensity score analyses (inverse of propensity score, 5 and 7 subclasses of propensity scores).

Table 5 presents the moderator effect of a binary care quality measure. Neither the conventional OLS nor propensity score analyses shows statistically significant difference on the average effect of Head Start as compared with other center-based care between high and low quality care.

Table 6 presents the moderator effect of a continuous care quality measure. Neither the conventional OLS nor propensity score analyses shows statistically significant difference on the slope of care quality measure for Head Start as compared with other center-based care.

In sum, the reading achievement at kindergarten for children in Head Start was significantly lower than their peers in other center-based care. The effect of Head Start as compare with other center-based care on kindergarten reading did not differ by care quality.

**Conclusions:**

In moderation analysis, it is important to eliminate selection bias for sample in different treatment and moderator groups. This paper demonstrates approaches to conducting causal moderation analysis using propensity score methods.

# Appendices

*Not included in page count.*

## Appendix A. References

Allison, P. D. (2001). *Missing Data.* Thousand Oaks, CA: Sage Publications.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182.

Bauer, D.J., & Curran, P.J. (2005). Probing interactions in fixed and multilevel regression: Inferential and graphical techniques. *Multivariate Behavioral Research*, 40, 373-400.

Boruch, R.F. (1997). *Randomized Experiments for Planning and Evaluation: A Practical Guide*. Thousand Oaks, CA: Sage.

Dong, N. (2011). *Using Propensity Score Methods to Approximate Factorial Experimental Designs*. Paper presented at the Fall 2011 Annual Conference of the Society for Research on Educational Effectiveness (SREE), Washington, DC. Available at http://www.sree.org/conferences/2011f/program/downloads/abstracts/323.pdf

Harms, T., Clifford, R. M., & Cryer, D. (1998). *Early Childhood Environment Rating Scale-Revised*. New York, NY: Teachers College Press.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945-60.

Hill, J. L., Brooks-Gunn, J. & Waldfogel, J. (2003). Sustained Effects of High Participation in an Early Intervention for Low-Birth-Weight Premature Infants. *Developmental Psychology.* 39(4):730–44.

Imai, K., Keele, L. & Tingley, D. (2010). A General Approach to Causal Mediation Analysis. *Psychological Methods,* Vol. 15, No. 4, 309–334.

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.

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. (2009). *Causal Inference in Statistics, and in the Social and Biomedical Sciences*. Cambridge Univ. Press, New York. Forthcoming.

Lochman, J. E., Boxmeyer, C. L., Powell, N. P., Roth, D., & Windle, M. (2006). Masked intervention effects: Analytic methods for addressing low dosage of intervention. *New Directions for Evaluation, 110,* 19-32.

Lu, B., Greevy, R., Xu, X., & Beck, C. (2011). Optimal Nonbipartite Matching and Its Statistical Applications. *The American Statistician, 65* (1). 21-29.

Lu, B., & Rosenbaum, P. (2004). Optimal Pair Matching With Two Control Groups, *Journal of Computational and Graphical Statistics*, 13, 422–434.

Ming, K. & Rosenbaum P.R. (2001). A Note on Optimal Matching with Variable Controls Using the Assignment Algorithm. *Journal of Computational and Graphical Statistics, 10* (3), 455-463.

Peck, L. R. (2003). Subgroup analysis in social experiments: Measuring program impacts based on post-treatment choice. *American Journal of Evaluation. 24* (2): 157-187.

Raudenbush, S. W. (2011). *Modeling Mediation: Causes, Markers, and Mechanisms*. Opening

Address at the Spring 2011 Society for Research on Educational Effectiveness (SREE) Conference, Washington, DC. Retrieved from the website: http://www.sree.org/conferences/2011/program/downloads/slides/raudenbush.pdf

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Rosenbaum, P. R. (2002). *Observational Studies*, 2nd ed. New York: Springer-Verlag.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41-55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*(387), 516-524.

Rubin, D. B. (1974). Estimating the causal effects oftreatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 66, 688-701.

Schochet, P. Z. & Burghardt, J. (2007). Using Propensity Scoring to Estimate Program-Related Subgroup Impacts in Experimental Program Evaluations. *Evaluation Review*, 31 (2), 95 – 120.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.

## Appendix B. Tables and Figures

*Not included in page count.*

Table 1: Potential Outcomes

|  |  | **Moderator** |  | Marginal Mean |
|---|---|---|---|---|
|  |  | 0 | 1 |  |
| **Treatment** | 0 | Y(0,0) | Y(0,1) | Y(0,.) |
|  | 1 | Y(1,0) | Y(1,1) | Y(1,.) |
| Marginal Mean |  | Y(.,0) | Y(.,1) |  |

Table 2: Covariate Balance Check among Child Care Type by Care Quality Groups before Applying Propensity Score Methods

| Variables | Other Center-Based Care[a] | | Head Start[a] | | Maximum Difference[b] |
|---|---|---|---|---|---|
| | Low ECERS | High ECERS | Low ECERS | High ECERS | |
| Reading at prekindergarten | 27.37 | 29.85 | 20.86 | 22.49 | 0.88 |
| | (9.96) | (12.76) | (7.53) | (8.16) | |
| Math at prekindegarten | 31.35 | 32.86 | 25.20 | 26.59 | 0.80 |
| | (9.75) | (10.49) | (8.46) | (8.30) | |
| Mental ability at 2 years old | 127.53 | 128.79 | 125.55 | 123.86 | 0.47 |
| | (10.38) | (11.64) | (8.66) | (10.07) | |
| Birth weight | 2969.64 | 3094.14 | 2996.35 | 2888.05 | 0.25 |
| | (846.14) | (833.14) | (765.66) | (832.67) | |
| Age (month) | 65.69 | 65.46 | 65.18 | 65.80 | 0.17 |
| | (3.70) | (3.69) | (3.53) | (3.36) | |
| Mother's education | 14.37 | 14.77 | 12.17 | 12.24 | 1.09 |
| | (2.53) | (2.71) | (1.82) | (1.89) | |
| Mom age at birth (year) | 28.11 | 29.73 | 24.78 | 24.43 | 0.86 |
| | (6.49) | (6.36) | (5.63) | (5.44) | |
| Income (thousand) | 67.87 | 82.72 | 22.54 | 27.20 | 1.26 |
| | (54.1) | (60.14) | (17.03) | (26.16) | |
| SES | 0.16 | 0.34 | -0.65 | -0.60 | 1.31 |
| | (0.81) | (0.82) | (0.61) | (0.65) | |
| Health problem | 0.26 | 0.26 | 0.24 | 0.22 | 0.08 |
| | (0.54) | (0.56) | (0.57) | (0.46) | |
| Violence | 0.07 | 0.07 | 0.10 | 0.09 | 0.09 |
| | (0.33) | (0.32) | (0.37) | (0.37) | |
| Disability | 0.82 | 1.19 | 1.17 | 1.42 | 0.36 |
| | (1.63) | (1.85) | (1.39) | (1.51) | |
| Kindergarten in 2007 (%) | 18 | 18 | 22 | 13 | 0.24 |
| | (39) | (39) | (42) | (34) | |
| Girl (%) | 52 | 49 | 55 | 49 | 0.13 |
| | (5) | (5) | (5) | (5) | |
| Black (%) | 18 | 15 | 5(5) | 29 | 0.84 |
| | (38) | (36) | | (45) | |
| Hispanic (%) | 13 | 14 | 19 | 27 | 0.39 |
| | (33) | (35) | (39) | (45) | |
| Other Race (%) | 22 | 19 | 11 | 14 | 0.28 |
| | (42) | (39) | (32) | (35) | |
| Single parent (%) | 26 | 15 | 43 | 36 | 0.65 |
| | (44) | (36) | (5) | (48) | |
| Welfare receipt (%) | 7 | 7 | 18 | 22 | 0.50 |
| | (25) | (26) | (39) | (42) | |
| NonEnglish (%) | 15 | 16 | 17 | 21 | 0.18 |
| | (35) | (37) | (38) | (41) | |
| Special education (%) | 7 | 6 | 8 | 7 | 0.05 |
| | (25) | (24) | (27) | (26) | |
| Poverty below185% (%) | 42 | 32 | 86 | 83 | 1.20 |
| | (49) | (47) | (35) | (38) | |
| Sample size[c] | 350 | 250 | 150 | 200 | |

[a] Entries are the means and standard deviation (in parentheses).

[b] Entries are the maximum standardized mean differences among four groups, which were calculated based on the pooled standard deviations across all four groups.

[c] Sample size is rounded to the nearest 50 per NECS regulation.

Table 3: Covariate Balance Check among Child Care Type by Care Quality Groups Weighted by the Inverse of Propensity Score

| Variables | Other Center-Based Care[a] | | Head Start[a] | | Maximum Difference[b] |
|---|---|---|---|---|---|
| | Low ECERS | High ECERS | Low ECERS | High ECERS | |
| Reading at prekindergarten | 25.82 | 25.59 | 23.94 | 25.55 | 0.09 |
| | (16.94) | (20.84) | (22.13) | (21.4) | |
| Math at prekindegarten | 29.60 | 29.44 | 27.75 | 29.37 | 0.09 |
| | (16.79) | (19.11) | (24.00) | (20.88) | |
| Mental ability at 2 years old | 126.27 | 126.52 | 124.76 | 125.44 | 0.09 |
| | (18.75) | (21.41) | (21.41) | (20.26) | |
| Birth weight | 2982.38 | 3014.68 | 2964.93 | 2876.49 | 0.09 |
| | (1437.14) | (1573.4) | (1935.03) | (1738.71) | |
| Age (month) | 65.42 | 65.43 | 65.28 | 65.69 | 0.06 |
| | (6.34) | (6.98) | (8.46) | (6.79) | |
| Mother's education | 13.64 | 13.56 | 12.61 | 13.15 | 0.22 |
| | (4.40) | (5.12) | (4.37) | (4.60) | |
| Mom age at birth (year) | 27.14 | 26.90 | 25.28 | 26.42 | 0.15 |
| | (11.18) | (12.76) | (12.94) | (13.41) | |
| Income (thousand) | 55.24 | 54.49 | 31.69 | 46.87 | 0.26 |
| | (88.69) | (101.5) | (54.51) | (98.1) | |
| SES | -0.10 | -0.12 | -0.45 | -0.23 | 0.23 |
| | (1.42) | (1.65) | (1.44) | (1.55) | |
| Health problem | 0.24 | 0.25 | 0.23 | 0.21 | 0.03 |
| | (0.85) | (1.03) | (1.32) | (0.93) | |
| Violence | 0.09 | 0.09 | 0.09 | 0.08 | 0.01 |
| | (0.6) | (0.67) | (0.79) | (0.72) | |
| Disability | 1.31 | 1.08 | 1.22 | 1.30 | 0.06 |
| | (3.92) | (3.42) | (3.73) | (3.37) | |
| Kindergarten in 2007 (%) | 19 | 16 | 23 | 17 | 0.16 |
| | (66) | (70) | (98) | (78) | |
| Girl (%) | 51 | 51 | 49 | 52 | 0.05 |
| | (84) | (95) | (117) | (104) | |
| Black (%) | 24 | 26 | 29 | 26 | 0.13 |
| | (72) | (83) | (1.07) | (92) | |
| Hispanic (%) | 17 | 18 | 21 | 21 | 0.10 |
| | (64) | (73) | (96) | (85) | |
| Other Race (%) | 18 | 19 | 15 | 14 | 0.12 |
| | (65) | (75) | (84) | (73) | |
| Single parent (%) | 27 | 28 | 31 | 36 | 0.19 |
| | (75) | (85) | (109) | (99) | |
| Welfare receipt (%) | 13 | 12 | 19 | 15 | 0.20 |
| | (56) | (62) | (92) | (73) | |
| NonEnglish (%) | 16 | 18 | 16 | 17 | 0.07 |
| | (62) | (73) | (85) | (78) | |
| Special education (%) | 9 | 7 | 6 | 9 | 0.11 |
| | (48) | (48) | (57) | (58) | |
| Poverty below185% (%) | 56 | 57 | 75 | 62 | 0.38 |
| | (84) | (94) | (102) | (101) | |
| Sample size[c] | 350 | 250 | 150 | 200 | |

[a] Entries are the means and standard deviation (in parentheses) weighted by the inverse of propensity score.

[b] Entries are the maximum standardized mean differences among four groups, which were calculated based on the pooled standard deviations across all four groups.

[c] Sample size is rounded to the nearest 50 per NECS regulation.

Table 4: The Overall Effect of Head Start as Compared with Other Center-based Care on Kindergarten Reading

| Analysis | Estimate | Standard Error | *p*-value | N[a] |
|---|---|---|---|---|
| Conventional OLS | -2.20 | 0.79 | 0.006 | 950 |
| Inverse of Propensity Score Weighting | -2.54 | 0.64 | <.0001 | 950 |
| 5 Subclasses[b] | -3.17 | 0.93 | 0.001 | 750 |
| 7 Subclasses[c] | -2.67 | 0.93 | 0.004 | 650 |

Source: ECLS-B
Note: The covariates included in the propensity score models and adjusted in the outcome analysis include: the mental IRT scale score measured at two years old, reading and math IRT scales measured at pre-kindergarten, gender, race (white, black, Hispanic, and other race), birth weight, age at outcome assessment, English speaking at home, disability status, special education status, health problem, family's SES, mother education, income, poverty level, welfare receipt, home violence, household structure, and one dummy variable indicating the year the child was in kindergarten (2007 vs. 2006).

[a] Sample size was rounded to the nearest 50 per NCES regulations.

[b] The number of the final subclasses used for weighted impact estimate was four. One subclass was excluded from analysis because it contained less than five children in Head Start.

[c] The number of the final subclasses used for weighted impact estimate was five. Two subclasses were excluded from analysis because each of them contained less than seven children in Head Start.

Table 5: The Moderator Effect of Binary Care Quality Measure on the Effect of Head Start as Compared with Other Center-based Care on Kindergarten Reading

| Analysis | Estimate | Standard Error | *p*-value | N[a] |
|---|---|---|---|---|
| Conventional OLS | -0.13 | 1.40 | 0.92 | 950 |
| Inverse of Propensity Score Weighting | -1.26 | 1.30 | 0.33 | 950 |
| Subclassification[b] | -1.53 | 1.93 | 0.43 | 600 |

Source: ECLS-B
Note: The covariates included in the propensity score models and adjusted in the outcome analysis include: the mental IRT scale score measured at two years old, reading and math IRT scales measured at pre-kindergarten, gender, race (white, black, Hispanic, and other race), birth weight, age at outcome assessment, English speaking at home, disability status, special education status, health problem, family's SES, mother education, income, poverty level, welfare receipt, home violence, household structure, and one dummy variable indicating the year the child was in kindergarten (2007 vs. 2006).

[a] Sample size was rounded to the nearest 50 per NCES regulations.

[b] The number of the original subclasses was nine while the final subclasses used for weighted impact estimate was six. Three subclasses were excluded from analysis because at least one of four groups (cells) contained less than five children.

Table 6: The Moderator Effect of Continuous Care Quality Measure on the Effect of Head Start as Compared with Other Center-based Care on Kindergarten Reading

| Analysis | Estimate | Standard Error | *p*-value | N[a] |
|---|---|---|---|---|
| Conventional OLS | -0.19 | 0.72 | 0.79 | 950 |
| Subclassification[b] | -0.32 | 0.98 | 0.74 | 600 |

Source: ECLS-B
Note: The covariates included in the propensity score models and adjusted in the outcome analysis include: the mental IRT scale score measured at two years old, reading and math IRT scales measured at pre-kindergarten, gender, race (white, black, Hispanic, and other race), birth weight, age at outcome assessment, English speaking at home, disability status, special education status, health problem, family's SES, mother education, income, poverty level, welfare receipt, home violence, household structure, and one dummy variable indicating the year the child was in kindergarten (2007 vs. 2006).

[a] Sample size was rounded to the nearest 50 per NCES regulations.

[b] The number of the original subclasses was nine while the number of the final subclasses used for weighted impact estimate was six. Three subclasses were excluded from analysis because each of them contained less than five children in Head Start.

Figure 1: Potential Outcomes and Paths

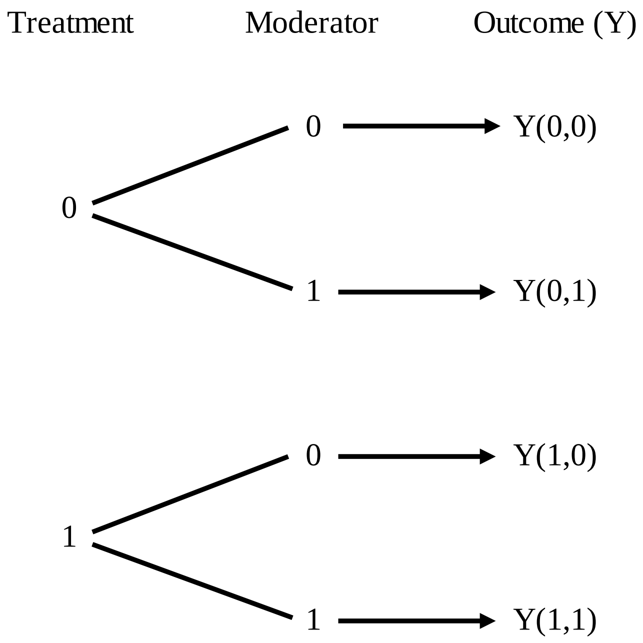

Figure 2. 3 × 3 Subclassification Based on Two Propensity Score Functions

| | | Propensity function for Moderator | | |
|---|---|---|---|---|
| | | Lower third | Middle third | Upper third |
| Propensity function for Treatment | Upper third | Subclass I | Subclass II | Subclass III |
| | Middle third | Subclass IV | Subclass V | Subclass VI |
| | Lower third | Subclass VII | Subclass VIII | Subclass XI |

Note: Adapted from Figure 4 by Imai & van Dyk (2004, p.861). Data are subclassified into three subclasses (lower third, middle third, and upper third) based on each of two propensity score functions, respectively. Each cell of the 3 × 3 table represents a subclass based on two propensity score functions jointly.