Abstract Title Page

Not included in page count.

Title:

The Effects of Head Start on Children’s Kindergarten Retention, Reading and Math Achievement in Fall Kindergarten — an Application of Propensity Score Method and Sensitivity Analysis

Author(s):

Nianbo Dong, University of Pennsylvania

Abstract Body

Limit 5 pages single spaced.

Background/context:

Description of prior research and/or its intellectual context and/or its policy context.

The Head Start program is a large-scale educational and care program for economically disadvantaged preschool children and their families. In particular, it focuses on helping children develop the early reading and math skills. Since its launch in 1965, the mixed effects of Head Start have been debated. One of the main reasons is that there are few studies using randomized experiment design and the most are observational studies. In an observational study, the effect estimates of Head Start can be biased due to the observable and unobservable factors which affect the selection into the Head Start program and the outcomes of participants.

In order to obtain causal inference in observational studies, researchers have applied propensity score and instrumental variable (IV) methods to examine the effects of Head Start (e.g., Magnuson, Ruhm, & Waldfogel, 2007; Zhai, 2007). One of the assumptions for propensity score and IV methods to lead to unbiased estimates of treatment effects from an observational study is that there are no unmeasured confounders for the assignment of treatment and for the assignment of IV variable, respectively (Rosenbaum, 2002b). Hence, it is desirable to learn how the inference about treatment effects would be altered by hidden biases of various magnitudes in an observational study using propensity score or IV method.

Purpose/objective/research question/focus of study:

Description of what the research focused on and why.

Using data from the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), this paper applied optimal propensity score matching method to evaluate the effects of Head Start on children’s kindergarten retention, reading and math achievement in fall kindergarten comparing with center-based care. Both parametric and nonparametric methods are used for impact analyses. Sensitivity analysis is conduced to assess the influence of hidden biases of various magnitudes.

Setting:

Specific description of where the research took place.

The data come from children’s family, kindergarten, and schools.

Population/Participants/Subjects:

Description of participants in the study: who (or what) how many, key features (or characteristics).

Data are from the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K). ECLS-K uses national representative sample. Data include information about children's cognitive skills, children's parents and school environments and so on, from pre-school to grade 5 (1998-99 cohorts). The numbers of children attending the Head Start program and center-based care are more than 1,300 and 6,400, respectively.

Intervention/Program/Practice:

Specific description of the intervention, including what it was, how it was administered, and its duration.

Head Start program is an educational and care program for disadvantaged preschool children and their families. In particular, it focuses on helping children develop the early reading and math skills. Comparison group is center-based care.

Research Design:

Description of research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).

This is a secondary data analysis. The optimal propensity score matching algorithm is used to match data. Sensitivity analysis is conducted to examine the effects of hidden variables.

Data Collection and Analysis:

Description of plan for collecting and analyzing data, including description of data.

Data is public used ECLS-K data. The optimal propensity score matching method is used to evaluate the effect of Head Start comparing with center-based care[1].

The optimal propensity score matching algorithm is better than greedy matching (the nearest neighbor matching) in that it minimizes the overall distance between propensity scores (Rosenbaum, 2002b). Data are matched using SAS PROC ASSIGN (Coca-Perraillon, 2007; Ming & Rosenbaum, 2001).

Backward logistic regression is used to estimate propensity score[2]. Covariate balance is tested for unmatched data and matched data respectively (Table 1 & 2, Figure 1 & 2). After

[1] The propensity score is the probability that a person with observed covariates receives the treatment rather than the control (Rosenbaum & Rubin, 1983). The optimal propensity score matching is to form the matched pairs to minimize the overall absolute propensity score differences between the matched pairs (Rosenbaum, 2002b). Propensity score method is an effective tool for adjusting for the measured confounders (overt biases).

[2] The logistic regression model Wald chi-square = 2106.73, p – value < 0.0001. The Hosmer-Lemeshow goodness fit test p – value = 0.21, which indicates plausibility of the model. Overall accuracy for identifying presence and absence of assignment to treatment group (Head Start) as based on conjoint maximum sensitivity and specificity levels, corresponding to the area under the receive operating characteristic (ROC) curve is 81.0%. The covariates in

matching, covariates are much more balanced, however, there are three slightly unbalanced covariates. Thus, covariance adjustment is used in analyses. Impact analyses using Ordinary Least Square regression (OLS) by regressing outcome difference between treated and control child in each pair on covariate difference between the covariates in each pair. In addition, in order to obtain more robust results, nonparametric estimate (Hodges Lehmann estimate with covariance adjustment) on matched data are conducted since there are some outliers for the outcome measures of reading and math, and the outcome measures are not normally distributed (Figure 3 – 6) [3]. Impact estimates using OLS to analyze original unmatched data were provided for comparison. Furthermore, sensitivity intervals of impact estimate with covariance adjustment based on Wilcoxon singed rank statistic are provided to illustrate the effect of hidden biases of various magnitudes (Rosenbaum, 2002a, 2002b) [4]. The formula of the 95% Sensitivity Intervals is given in Appendix C.

Findings/Results:

Description of main findings with specific details.

Table 3 presents the impact estimates of reading and math of Head Start comparing with center-based care. Using OLS to analyze the matched data (optimal propensity score matching), children in Head Start program are estimated to perform about 2.36 and 2.14 points IRT score lower than peers in center-based care in reading and math in fall kindergarten, respectively (*p* – value < 0.0001). The corresponding standardized effect sizes are -0.32 and -0.31 for reading and

the final propensity score estimate model include: Black, Hispanic, other race, birth weight, disability status, number of sibling, mother working status during pregnancy to kindergarten, mother marriage status in birth, SES, poverty level, location (rural or not), SES*poverty, and black* poverty.

[3] *P* – values of Shapiro-Wilk, Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling test are all less than 0.01.

[4] A sensitivity analysis is a specific statement about the magnitude of hidden bias that would need to be present to explain the associations actually observed in a particular study (Rosenbaum, 2002a, 2002b). Suppose there is an unobserved covariate u that determines the probability of a child participating in Head Start (receiving treatment) in addition to the observed covariates $\mathbf{x}$. The sensitivity analysis model has two parts, a logit form linking treatment assignment S_j to the covariates ($\mathbf{x}_j, u_j$) and a constraint on u_j, namely

$$\log\left(\frac{\pi_j}{1-\pi_j}\right) = \kappa(\mathbf{x}_j) + \gamma u_j \text{ with } 0 \leq u_j \leq 1$$

where π_j is the probability that child *j* receives the treatment (participating in Head Start). $\kappa(\cdot)$ is an unknown function and γ is an unknown parameter. $\Gamma = e^{\gamma}$ is the odds ratio of *j* receives the treatment (participating in Head Start) compared to *k* receiving treatment. If $\Gamma = 1$, i.e., $\gamma = 0$, then study is free of hidden bias. For $\Gamma > 1$ there is hidden bias. Γ is a measure of the degree of departure from a study that is free of hidden bias.

Consider the additive treatment effect model $Y_i^{(1)} = Y_i^{(0)} + \beta$. A 95% *sensitivity interval* for impact estimate (β) with sensitivity parameter Γ is a random interval that in at least 95% of studies will contain the true β assuming that the true γ, call it γ_0 satisfies $\exp(\gamma_0) \leq \Gamma$ (Rosenbaum, 2002b).

math respectively[5]. Using the same matched data, the nonparametric estimate (Hodges Lehmann estimate) with covariate adjustment gives us smaller results (-1.92 for reading and math both). Furthermore, using unmatched data, OLS gives us slightly different estimate in math (-2.07), but much more different estimate in reading (-2.57). The standardized effect size for reading and math are -0.24 and -0.22 respectively[6].

Table 4 presents the impact estimates of kindergarten retention of Head Start comparing with center-based care. Using logistic regression to analyze unmatched data and matched data gives us almost identical results. The odd ratio estimates of Kindergarten retention for children in Head Start comparing peers in center-based care are 1.04 with 95% confidence interval (0.65, 1.69) and 1.05 with 95% confidence interval (0.71, 1.57), respectively for using matched data and using unmatched data. Neither estimate is statistically significant. It suggests that there is no different impact on kindergarten retention between Head Start and center-based care.

Table 5 presents the 95% sensitivity intervals of impact estimates of reading based on Wilcoxon signed rank test with covariance adjustment. If the study is free of hidden bias (Γ = 1), the 95% Hodges Lehmann estimate on reading based on Wilcoxon signed rank test is (-2.51, -1.34). When Γ = 1.43, the 95% sensitivity interval would include 0, which suggests the impact is not statistically significant. In other word, if Γ = 1.43, matched children differ by a factor of 1.43 times in their odds of participating in Head Start due to differences in the unobserved covariate. In this case, there is no statistical difference in reading achievement between children in Head Start and in center-based care. This indicates that the impact estimate is sensitive to hidden bias.

Table 6 presents the 95% sensitivity intervals of impact estimates of math based on Wilcoxon signed rank test with covariance adjustment. Similarly, if the study is free of hidden bias (Γ = 1), the 95% sensitivity interval of impact estimate on math based on Wilcoxon signed rank test is (-2.48, -1.37). When Γ = 1.42, the 95% sensitivity interval would include 0, which suggests the impact is not statistically significant. In other word, if Γ = 1.42, matched children differ by a factor of 1.42 times in their odds of participating in Head Start due to differences in the unobserved covariate. In this case, there is no statistical difference in math achievement between children in Head Start and in center-based care. This also indicates that the impact estimate is sensitive to hidden bias.

[5] The pooled standard deviation for reading and math are 10.64 and 9.08 respectively.

[6] The pooled standard deviation for reading and math are 7.35 and 6.81 respectively.

Conclusions:

Description of conclusions and recommendations of author(s) based on findings and over study. (To support the theme of 2009 conference, authors are asked to describe how their conclusions and recommendations might inform one or more of the above noted decisions—curriculum, teaching and teaching quality, school organization, and education policy.)

Children in Head Start program tend to perform statistically significantly worse than peers in center-based care in reading and math in fall kindergarten. However, the sensitivity intervals of impact estimate based on Wilcoxon signed rank test with covariance adjustment indicate that the conclusion above is sensitive to hidden bias. When $\Gamma > 1.43$, i.e., matched children differ by a factor of above 1.43 times in their odds of participating in Head Start due to differences in the unobserved covariate, the impact on reading and math between Head Start and center-based care will have no statistically significant difference. Besides, there is no statistically difference of the effect on kindergarten retention between Head Start and center-based care.

Optimal propensity score matching method and OLS regression on unmatched data method gave us almost identical results on estimates of math and kindergarten retention, but a little different of reading. In observational study, in theory propensity score method can perform better than direct OLS regression in terms of adjusting for measured confounders (overt biases). However, the improvement is limited in this case. This conclusion is consistent with Bloom, Michalopoulos, & Hill (2005).

Furthermore, since there are some outliers in the outcome measures of reading and math, and the outcome measures are not normally distributed, the distribution-free Hodges-Lehmann estimate based on the Wilcoxon signed rank test is more robust.

Appendixes

Not included in page count.

Appendix A. References

References are to be in APA format. (See APA style examples at the end of the document.)

Bloom, H. S., Michalopoulos, C., & Hill, C. J. (2005) Using Experiments to Assess Nonexperimental Comparison-Group Methods for Measuring Program Effects. In H. S. Bloom (Eds.), *Learning More from Social Experiment* (pp. 173- 235). New York: Russell Sage Foundation.

Coca-Perraillon, M. (2007). Local and Global Optimal Propensity Score Matching. *SAS Global Forum*.

Magnuson, K. A, Ruhm, C., & Waldfogel, J. (2007). Does Prekindergarten Improve School Preparation and Performance. Economics of Education Review, 26(1), 33-51.

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

Rosenbaum, P. R. (2002a). Covariance Adjustment in Randomized Experiments and Observational Studies. *Statistical Science. 17* (3): 286 – 327.

Rosenbaum, P. R. (2002b). *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, 1, 41–55.

Zhai, F. (2007). The Effects of Head Start Participation on Child Health From Kindergarten to the 5th Grade. *Extended Abstract for PAA 2007 Annual Conference Submission*. Retrieved May 27th, 2008, from http://paa2007.princeton.edu/abstractViewer.aspx?submissionId=71643

Appendix B. Tables and Figures

Not included in page count.

Table 1

Covariate balance test in unmatched data

Variable	Control (Center-Based Care)	Treated (Head Start)	Standardized Difference[a]
Male (%)	51.0	47.3	-7.3
Black (%)	11.0	33.8	57.0*
Hispanic (%)	12.1	22.4	27.6*
Other Race (%)	9.2	15.7	19.8*
Poverty (%)	10.0	52.0	101.9*
Disabled (%)	14.5	18.2	10.1*
Mother worked during pregnancy and kindergarten (%)	74.1	60.3	-29.7*
Mother was married in birth (%)	81.9	45.8	-81.2*
Located in rural (%)	15.8	33.9	42.8*
Age (month)	65.7	65.6	-4.2
Number of sibling	1.3	1.9	40.0*
Birth weight (pound)	7.5	7.1	-25.7*
SES	0.3	-0.6	-127.4*

Source: ECLS-K.

Note. [a]Formulas for calculating standardized difference for continuous variables and dichotomous variables are $d = \frac{100(\bar{x}_T - \bar{x}_C)}{\sqrt{\frac{s_T^2 + s_C^2}{2}}}$ and $d = \frac{100(\hat{p}_T - \hat{p}_C)}{\sqrt{\frac{\hat{p}_T(1-\hat{p}_T) + \hat{p}_C(1-\hat{p}_C)}{2}}}$, respectively (Rosenbaum, 2002b). Sample sizes for control and treated group are 6,021 and 1,169 respectively.

*The absolute value of standardized difference larger than 10 is considered as unbalanced (Rosenbaum, 2002b).

Table 2
Covariate balance test in matched data

Variable	Control (Center-Based Care)	Treated (Head Start)	Standardized Difference[a]
Male (%)	50.1	47.7	-4.8
Black (%)	29.6	32.9	7.3
Hispanic (%)	24.6	24.1	-1.2
Other Race (%)	11.2	13.1	6.0
Poverty (%)	43.4	50.4	14.0*
Disabled (%)	18.6	17.6	-2.5
Mother worked during pregnancy and kindergarten (%)	66.7	62.7	-8.5
Mother was married in birth (%)	50.0	46.7	-6.6
Located in rural (%)	27.0	31.5	9.9
Age (month)	65.1	65.5	9.2
Number of sibling	1.7	1.8	10.3*
Birth weight (pound)	7.2	7.2	-5.5
SES	-0.5	-0.6	-10.4*

Source: ECLS-K.
Note. [a]Formulas for calculating standardized difference for continuous variables and dichotomous variables are $d = \frac{100(\bar{x}_T - \bar{x}_C)}{\sqrt{\frac{s_T^2 + s_C^2}{2}}}$ and $d = \frac{100(\hat{p}_T - \hat{p}_C)}{\sqrt{\frac{\hat{p}_T(1-\hat{p}_T) + \hat{p}_C(1-\hat{p}_C)}{2}}}$, respectively (Rosenbaum, 2002b). Sample sizes for control and treated group are 6,021 and 1,169 respectively.
*The absolute value of standardized difference larger than 10 is considered as unbalanced (Rosenbaum, 2002b).

Table 3

Impact estimates of reading and math of Head Start comparing with center-based care

Estimate Method		Reading[a]	Math[b]
Using original data (unmatched) [c]	Ordinary Least Square Regression (OLS) [e]	-2.57 †	-2.07 †
Optimal Propensity Score Matching [d]	OLS (covariance adjustment) [f]	-2.36 †	-2.14 †
	Hodges Lehmann estimate (covariance adjustment) [f]	-1.92 †	-1.92 †

Source: ECLS-K.
Note. [a]Reading IRT scale score in fall kindergarten. [b]Math IRT scale score in fall kindergarten. [c] *N* = 6,743 and 6,961 for Reading and Math, respectively. [d] *N* = 1,700 and 2,018 for reading and math, respectively. [e] Controlling for age, gender, black, Hispanic, other race, disability status, number of sibling, birth weight, location (rural), SES, mother marriage status in child birth, mother working status during pregnancy to kindergarten, mother education, Black*SES, SES*SES, and SES*SES* poverty status (insignificant variables are not included in the estimate model). [f] Controlling for age, gender, black, Hispanic, disability status, number of sibling, location (rural), SES, mother marriage status in child birth, mother working status during pregnancy to kindergarten, poverty status, and SES*poverty status (insignificant variables are not included in the estimate model).
⋆*p* < .05. ⋆⋆ *p* < .01. ⋆⋆⋆*p* < .001. † *p* < .0001.

Table 4

Impact estimates of kindergarten retention of Head Start comparing with center-based care

Estimate Method		Kindergarten Retention[a]
Using original data (unmatched)[b]	Logistic Regression[d]	1.05 (0.71, 1.57)
Optimal Propensity Score Matching[c]	Logistic Regression[d]	1.04 (0.65, 1.69)

Source: ECLS-K.
Note. [a]Parameters entered are odds ratio. 95% confidence interval in parenthesis.
[b] N = 6,889. [c] N = 2,078. [d]Controlling for age, gender, black, disability status, number of sibling, birth weight, location (rural), and mother marriage status in child birth (insignificant variables are not included in the estimate model).

Table 5

Sensitivity intervals of impact estimate of reading of Head Start comparing with center-based care based on Wilcoxon signed rank test[a]

Sensitivity parameter (Γ)	95% Sensitivity interval
1	(-2.51, -1.34)
1.4	(-3.81, -0.07)
1.43	(-3.90, 0.01)
1.5	(-4.09, 0.19)
2	(-5.24, 1.28)

Note. [a]Controlling for age, black, Hispanic, disability status, number of sibling, location (rural), SES, mother marriage status in child birth, mother working status during pregnancy to kindergarten, poverty status, and SES* poverty status (insignificant variables are not included in the estimate model).

Table 6

Sensitivity intervals of impact estimate of math of Head Start comparing with center-based care based on Wilcoxon signed rank test[a]

Sensitivity parameter (Γ)	95% Sensitivity interval
1	(-2.48, -1.37)
1.4	(-3.84, -0.04)
1.42	(-3.89, 0.02)
1.5	(-4.12, 0.23)
2	(-5.30, 1.37)

Note. [a] Controlling for age, gender, black, Hispanic, disability status, number of sibling, location (rural), SES, mother marriage status in child birth, and mother working status during pregnancy to kindergarten (insignificant variables are not included in the estimate model).

Figure 1. Overlap checking on propensity score (unmatched data)

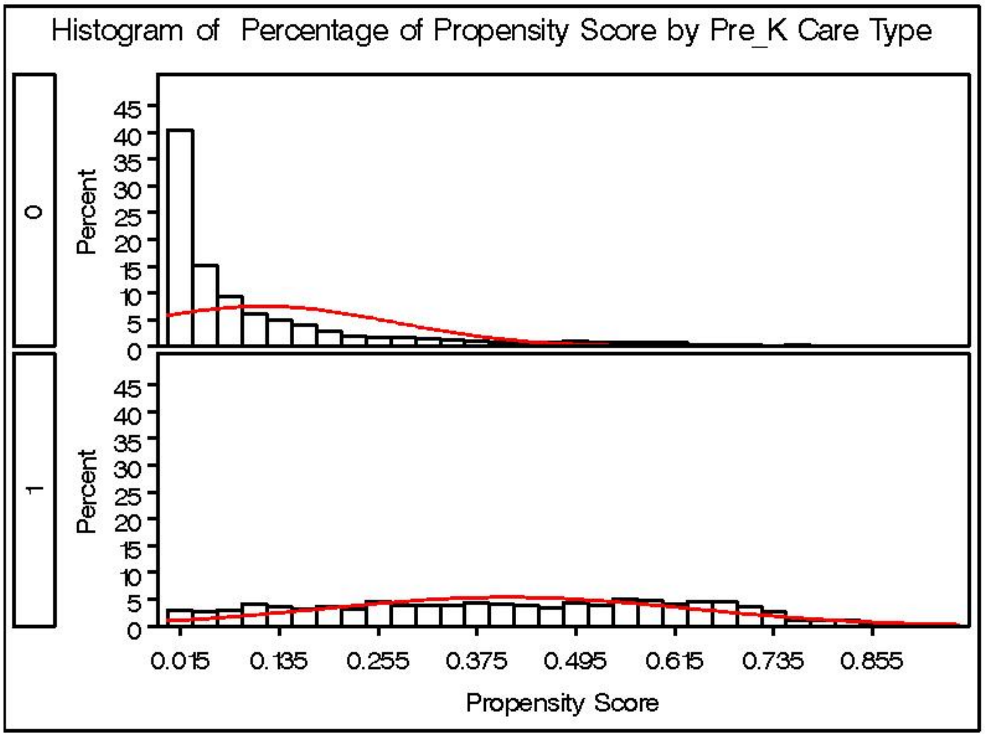


Figure 2. Overlap checking on propensity score (matched data)

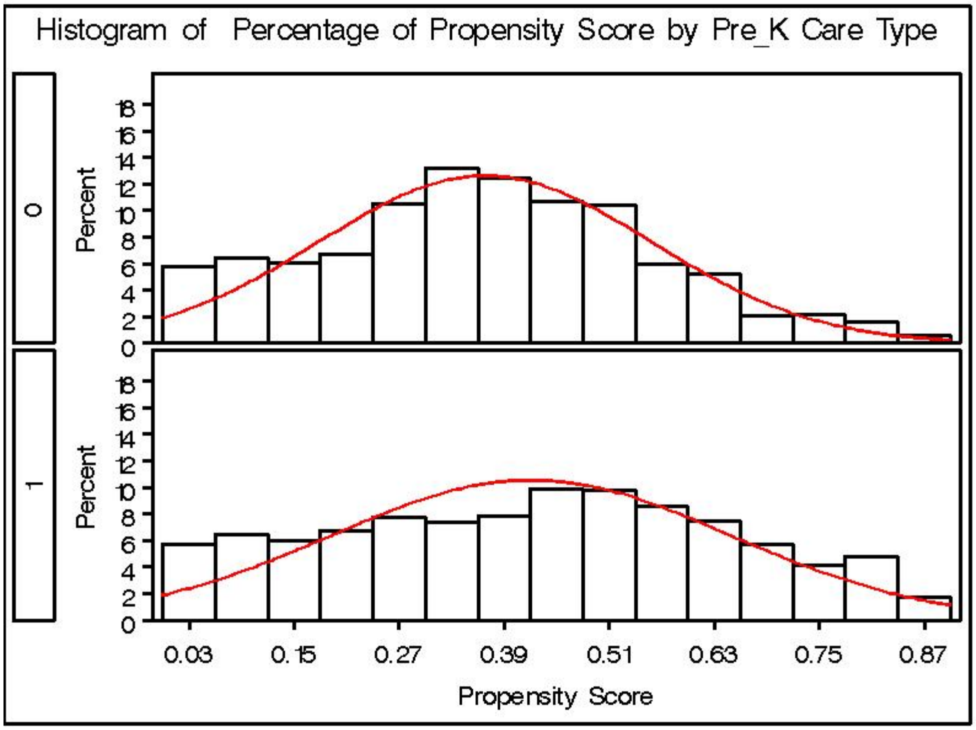


Figure 3. Boxplot of fall kindergarten reading IRT score by child care type

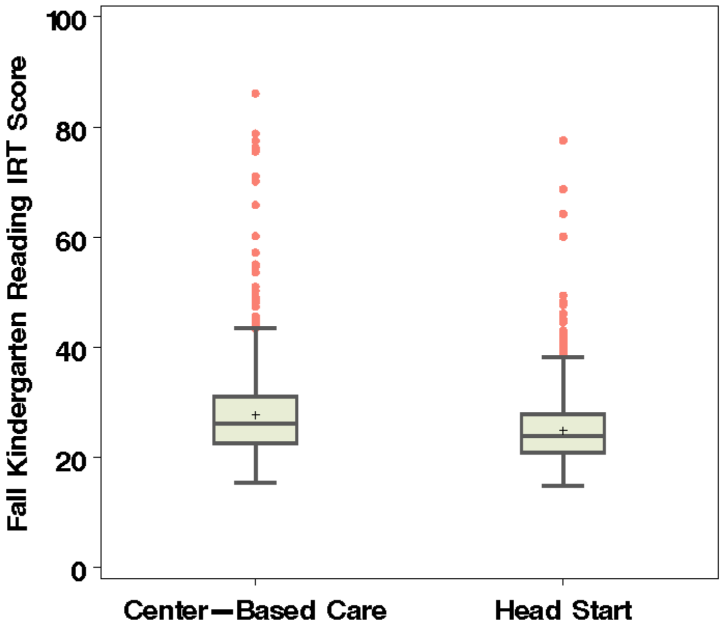


Figure 4. Boxplot of fall kindergarten math IRT score by child care type

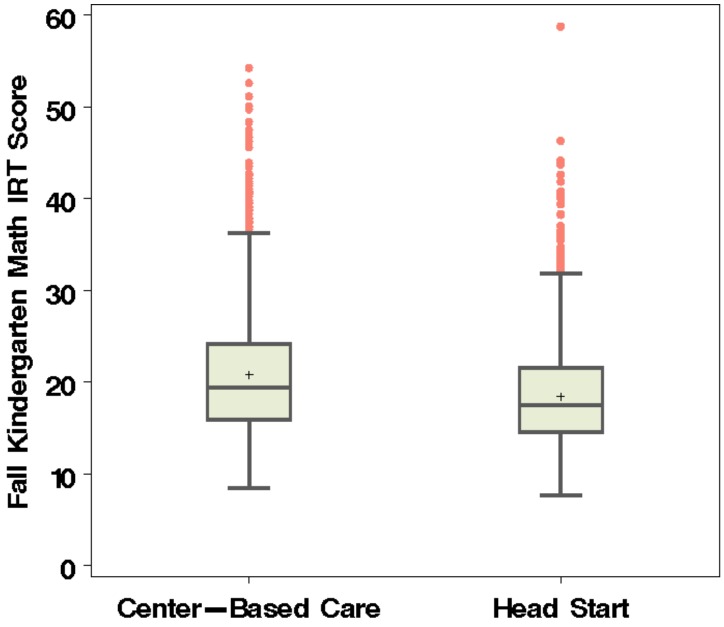


Figure 5. Q-Q Plot of fall kindergarten reading IRT score by child care type

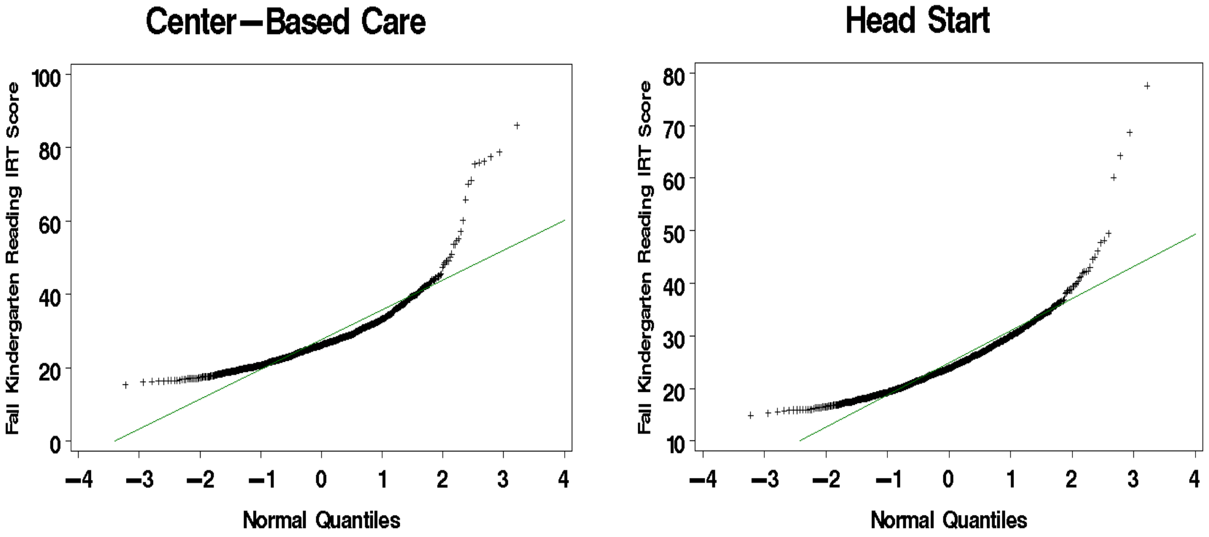
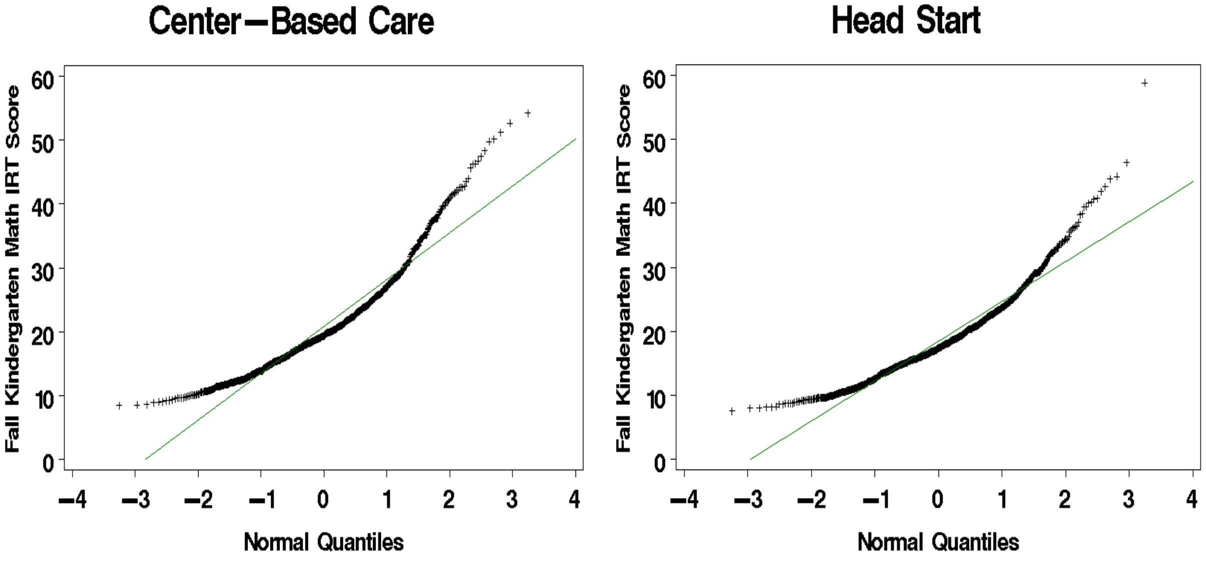


Figure 6. Q-Q Plot of fall kindergarten math IRT score by child care type



Appendix C. The Formula of the 95% Sensitivity Intervals

Consider matched pair study units 1, 2, …, *N* are divided into matched pairs (1,2),…, (*N*-1, *N*), one of whom receives treatment and the other control, Wilcoxon signed rank statistic is given by: $T = t(\mathbf{S}, \mathbf{y}) = \sum_{j=1}^{N/2} d_j \sum_{k=1}^{2} c_{jk} S_{2j-2+k}$, where y = (y_1,…, y_N) is the vector of observed outcomes, and $\mathbf{S} = (S_1, ..., S_N)$ denote the treatment assignment. c_{jk} is a binary indicator, c_{jk} = 1 or 0, and c_{jk} are the functions of y. d_j is the rank of $| y_{2j-1} - y_{2j} |$.

Using the normal approximation, the endpoints of the 95% sensitivity intervals are given by Rosenbaum (2002b):

$$\inf \left\{ \beta : \frac{\mathrm{T}_{\beta,obs} - \mathrm{E}(\mathrm{T}^{+}_{\beta})}{\sqrt{\mathrm{var}(\mathrm{T}^{+}_{\beta})}} \le 1.96 \right\} \text{ and } \sup \left\{ \beta : \frac{\mathrm{T}_{\beta,obs} - \mathrm{E}(\mathrm{T}^{-}_{\beta})}{\sqrt{\mathrm{var}(\mathrm{T}^{-}_{\beta})}} \ge 1.96 \right\}$$

where $\mathrm{T}_{\beta,obs}$ is Wilcoxon signed rank statistic. T^{+}_{β} is defined to be the sum of $N/2$ random variables where *j*th random variable takes the value d_j (the rank of $| y_{2j-1} - y_{2j} |$) with probability p^{+}_{j} (p^{+}_{j}=0 if $y_{2j-1} = y_{2j}$; $p^{+}_{j} = \Gamma / (1+\Gamma)$ if $y_{2j-1} \ne y_{2j}$) and take the value 0 with probability $1 - p^{+}_{j}$. T^{-}_{β} is defined similarly with T^{+}_{β} with p^{-}_{j} (p^{-}_{j}=0 if $y_{2j-1} = y_{2j}$; p^{-}_{j}=$1/(1+\Gamma)$ if $y_{2j-1} \ne y_{2j}$) in place of p^{+}_{j}. $E(T^{+}_{\beta}) = \sum_{j=1}^{N/2} d_j p^{+}_{j}$ and $Var(T^{+}_{\beta}) = \sum_{j=1}^{N/2} d^{2}_{j} p^{+}_{j} (1 - p^{+}_{j})$. For T^{-}_{β}, the expectation and variance are given by the same formulas with p^{-}_{j} in place of p^{+}_{j}. As the number of pairs $N/2$ increases, the distributions of T^{+}_{β} and T^{-}_{β} are approximated by Normal distributions.