SSM - Population Health 19 (2022) 101223

Contents lists available at ScienceDirect

SSM - Population Health

journal homepage: www.elsevier.com/locate/ssmph

Assessment of the predictive power of a causal variable: An application to the Head Start impact study

Sun Yeop Lee [a], Rockli Kim [b,c], Justin Rodgers [d,**], S.V. Subramanian [d,e,*]

[a] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[b] Division of Health Policy and Management, College of Health Science, Korea University, Seoul, South Korea
[c] Interdisciplinary Program in Precision Public Health, Department of Public Health Sciences, Graduate School of Korea University, Seoul, South Korea
[d] Harvard Center for Population & Development Studies, Cambridge, MA, USA
[e] Department of Social and Behavioral Sciences, Harvard T.H. Chan School of Public Health, Boston, MA, USA

A B S T R A C T

In a study attempting to estimate a causal effect of a causal variable, an assessment of the predictive power of the causal variable can shed light on the heterogeneity around its average effect. Using data from the Head Start Impact Study, a randomized controlled trial of the Head Start, a nation-wide early childhood education program in the United States, we provide a parallel comparison between measures of average effect and predictive power of the Head Start on five cognitive outcomes. We observed that one year of the Head Start increased scores for all five outcomes, with effect sizes ranging from 0.12 to 0.19 standard deviations. Percent variation explained by the Head Start ranged from 0.56 to 1.62%. For binary versions of the outcomes, the overall pattern remained; the Head Start on average improved the outcomes by meaningful magnitudes. In contrast, in a fully adjusted model, the Head Start only improved area under the curve (AUC) by less than 1% and its influence on the variance of predicted probabilities was negligible. The Head-Start-only model only achieved AUC ranging from 50.22 to 55.24%. Negligible predictive power despite the significant average effect suggests that the heterogeneity in effects may be large. The average effect estimates may not generalize well to different populations or different Head Start program settings. Assessment of the predictive power of a causal variable in randomized data should be a routine practice as it can provide helpful information on the causal effect and especially its heterogeneity.

1. Introduction

The Head Start, one of the largest and the only federally funded early childhood education program in the United States, aims to enhance school readiness of children of low-income families by providing educational, health, and social services (Head Start & Early Head Start, 2020). In 2002, the Head Start Impact Study (HSIS), a randomized controlled trial (RCT) of the Head Start, was designed to test the causal effect of the Head Start on children's developmental outcomes, including cognition, social-emotional measures, health, and parenting (Puma et al., 2010). Official reports of the HSIS reported one- or two-year positive causal effects of the Head Start, especially for cognitive outcomes, although they faded away in a few years (Puma et al., 2010, 2012). Follow-up studies that conducted subgroup analyses additionally found that the effects were larger and more pronounced for certain subgroups, such as children with low cognitive abilities or Spanish as primary language (Bitler et al., 2014), or children who would have received home-based care had they not enrolled in the Head Start (Feller et al., 2016; Zhai et al., 2014). Existing literature on the heterogeneity of the Head Start effect is extensive, but most focused on average effect, taking a mean-centric approach (Lee et al., 2021). Little attention has been directed to an assessment of the predictive power measured by various metrics such as percent variation explained or area under the curve (AUC).

Predictive power measures the capacity of a variable to correctly identify or estimate outcomes in independent data and is generally assessed during predictive model development. In a study attempting to estimate a causal effect of a causal variable (e.g., a random assignment to the Head Start), an assessment of the predictive power of the causal variable can shed light on the heterogeneity around its average effect. In observational settings, exposures (i.e., birthweight) with well-established average associations (or effects) have often been observed to have low predictive power, suggesting that they are not necessarily suitable for individual classification (e.g., medical screening, eligibility criteria for social programs and policies) (Kim et al., 2018; Merlo et al., 2017; Swaminathan et al., 2020); a strong effect on average does not necessarily mean strong predictive power (Pepe et al., 2004; Varga et al., 2020; Wald et al., 1999). In an RCT setting, an evaluation of the

* Corresponding author. Harvard Center for Population & Development Studies, 9 Bow Street, Cambridge, MA, 02138, USA.
** Corresponding author.
E-mail addresses: sun.yp.lee@gmail.com (S.Y. Lee), jrogers@hsph.harvard.edu (J. Rodgers), svsubram@hsph.harvard.edu (S.V. Subramanian).

https://doi.org/10.1016/j.ssmph.2022.101223
Received 9 May 2022; Received in revised form 17 August 2022; Accepted 30 August 2022
Available online 6 September 2022

predictive power is rarely considered and is not a regular analytic practice, despite the complementary nature of assessments of average effect and predictive power (Merlo et al., 2017; Shmueli, 2010). As individual causal effects are expected to be heterogeneous (Kravitz et al., 2004; Plewis, 2002), the magnitude of this heterogeneity should be examined. It has implications when deciding whether to scale up, terminate, or tailor to specific populations (Cintron et al., 2022; Subramanian et al., 2018).

Therefore, using the HSIS data, we evaluated the Head Start in terms of both average effect (e.g., effect size, odds ratio (OR)) and the predictive power (e.g., percent variation explained, AUC, variance of predicted probabilities). Specifically, we estimated average effects on and predictive power for five outcomes (i.e., Peabody Picture Vocabulary Test (PPVT), Letter-Word Identification, Applied Problem, Spelling, and Pre-Academic Skill) after one year of the Head Start. Since the HSIS has already been extensively analyzed in terms of average effects, here we only focused on outcomes with previous reports of statistically significant one-year average effects (Lee, Rodgers, et al., 2022). Therefore, the estimates for average effects are reproductions of previous results. Predictive power of a known causal variable is the inferential target of interest in this study.

2. Methods

2.1. Data

The HSIS was designed to evaluate the effectiveness of the Head Start on children's cognitive, behavioral, social-emotional, and health outcomes (Puma et al., 2010). The Head Start programs provide early childhood education and social services to children (i.e., aged <1–5 years) and their families through local non-profit and for-profit community agencies such as centers and schools. The HSIS implemented a multi-stage sampling procedure for recruiting participants. The sampling procedure first categorized the initial 1715 programs into 161 geographic clusters and 25 strata based on region, state-level childcare policy, race/ethnicity, and urbanicity. Next, one cluster was randomly selected from each stratum, excluding programs that were closed, merged, or saturated and grouping those with small sample sizes. These programs were then stratified by type and local contextual characteristics. Then, three programs were randomly chosen from each stratum. Lastly, centers were randomly chosen from the final set of programs. This multi-stage sampling procedure resulted in a final sample of 4442 children at 378 centers within 84 programs. This sample was then followed over time from 2002 to 2008, with measurements collected at baseline and annual to biennial follow-ups on a host of developmental and related measures. Detailed explanations of the sampling procedure and other study protocols are available in the official HSIS reports (Puma et al., 2010, 2011).

2.2. Treatment

The Head Start intervention included educational, health, nutritional, and social services with the goal of improving school readiness and child development. All Head Start centers must adhere to the Head Start Performance Standards, which are federally regulated to ensure the comprehensiveness and quality of the services provided by the centers (Puma et al., 2010). Thus, the treatment is a mixture of various services with the nation-wide, pre-specified standards.

The assignment to the Head Start was randomized within each Head Start center, offering the assigned children to participate in the Head Start. To potentially benefit as many children as possible with the program, the randomization was intentionally designed to yield a higher proportion of children in the treatment group. The treatment of interest specifically represents the assignment to one year of the Head Start in the baseline year. Like any RCT, noncompliance to the treatment/control assignment occurred; 12% of the control group enrolled in the Head Start, and 19% of the treatment group did not enroll in the Head Start.

2.3. Outcomes

Children were assessed on a multitude of developmental outcomes over the course of the HSIS, commencing during children's preschool years (ages 3–4 years) in the baseline year of 2002, with follow-ups in Spring 2003, Spring 2004, Spring 2005, Spring 2006, and Spring 2008 (3rd grade). In this study, we focused on outcomes with previous reports of statistically significant average effects after one year of the Head Start: PPVT, Letter-Word Identification, Applied Problem, Spelling, and Pre-Academic Skill. Outcomes without average effects would not have meaningful predictive power. Predictive power of the Head Start for the outcomes with average effects is of our interest.

Cognitive outcomes were measured by one-on-one child assessments for 45–60 min. PPVT measures receptive vocabulary in standard English (Cronbach's $\alpha = 0.62$–0.84). Letter-Word Identification measures the ability to identify letters and words from a picture or isolated letters and words ($\alpha = 0.82$–0.94). Spelling measures the ability to correctly spell spoken words ($\alpha = 0.70$–0.94). Applied Problem measures an ability to analyze and solve math problems ($\alpha = 0.85$–0.90). Pre-Academic Skill is a composite measure of Letter-Word Identification, Applied Problems, and Spelling ($\alpha = 0.67$–0.85).

2.4. Covariates

While the HSIS was an RCT, the HSIS official reports recommended covariate adjustment to enhance statistical precision and adjust for any systematic bias at baseline (Puma et al., 2010, 2011). Therefore, for the average effect assessment, we adjusted for children's sociodemographic variables and study-related variables. Sociodemographic variables included gender (male, female), race/ethnicity (White/other, Black, Hispanic), primary language at baseline (English, Spanish), special needs (yes, no), primary caregiver's age (continuous), teen mom at birth (yes, no), living with a single parent (yes, no), recent immigrant parents (yes, no), parents' marital status (not married, married, separated/divorced/widowed), parental education level (less than high school, high school graduates, beyond high school), urbanicity (urban, rural), household risk (low, moderate, high). Study-related variables included age cohort (age 3, age 4) and baseline measures of the outcomes in this study.

2.5. Statistical analyses

We employed two distinct analytic approaches: 1) an average effect assessment and 2) a predictive power assessment. All analyses were performed in R (version 4.1.1) (R Core Team, 2020).

Three-level multilevel linear regressions were specified in order to account for the complex sampling design (level-3: program; level-2: center; level-1: child). For average effect, we estimated the fixed effect parameter estimate of the Head Start, adjusting for covariates and random effects. The average effect was presented as both a raw score and Cohen's *d* effect size (Cohen, 1992). For predictive power, we estimated child-level variances of an outcome in a full model with the Head Start (i.e., Model 1) and a full model without the Head Start (i.e., Model 2). Model 1 was specified as,

$$Y_{ijk} = \beta_0 + \beta_1 T_{ijk} + \beta X'_{ijk} + \left(v_{0k} + u_{0jk} + e_{0ijk}\right)$$

$$[v_{0k}] \sim N\left(0, \sigma^2_{v_0}\right) \rightarrow [u_{0jk}] \sim N\left(0, \sigma^2_{u_0}\right) \rightarrow [e_{0ijk}] \sim N\left(0, \sigma^2_{e_0}\right)$$

where Y_{ijk} is an outcome variable for child *i* in center *j* in program *k*, X'_{ijk} is a vector of child-level covariates, and T_{ijk} is an indicator variable for the treatment group (i.e., the Head Start). Total variance is partitioned into the program-level ($\sigma^2_{v_0}$), the center-level ($\sigma^2_{u_0}$), the child-level ($\sigma^2_{e_0}$).

Model 2 is equivalent to Model 1 excluding the treatment group indicator variable. We report the percent change between the child-level variances of the two models as percent variation explained by the Head Start.

Binary versions of the outcomes were also utilized so that our analyses could be performed in binary outcome scenarios. The outcomes were dichotomized at the 25th, 50th, and 75th percentiles and were indicators for scoring above the varying thresholds. Three-level multilevel logistic regressions with (i.e., Model 3) and without the Head Start (i.e., Model 4) were estimated to assess the average effects and predictive power improvements measured by AUC contributed to the Head Start. Model 3 was specified as,

$$logit(Y_{ijk}) = \beta_0 + \beta_1 T_{ijk} + \beta X'_{ijk} + (v_{0k} + u_{0jk} + e_{0ijk})$$

$$[v_{0k}] \sim N(0, \sigma^2_{v_0}) \rightarrow [u_{0jk}] \sim N(0, \sigma^2_{u_0}) \rightarrow [e_{0ijk}] \sim B(1, Y_{ijk})$$

Additionally, variances of predicted probabilities from Model 3 and 4 were compared. In general, a better-calibrated model would have more extreme predictions, hence the greater variance. To visualize, predicted probabilities for the 50th percentile threshold dichotomized outcomes were compared side by side in histograms.

Lastly, a logistic regression with only the Head Start as an independent variable (i.e., Model 5) was run to estimate independent predictive power (i.e., AUC) of the Head Start.

3. Results

At baseline, there was a total sample size of 4442 children participating in the study, of which 2646 were assigned to the Head Start group and 1796 were assigned to the control group (Table 1). A slightly higher proportion of the participants were Hispanics/others (36.0%) than White (33.7%) and Black (30.3%). About a quarter (25.7%) used Spanish as a primary language. Approximately half (50.4%) of children lived with a single parent, 38.0% had mothers who did not graduate from high school, and approximately one-fifth (19.2%) were recent immigrants. Out of the 4442 children, 81–82% of them composed the final analytic sample depending on the availability of data on the outcomes and covariates (Table 2).

In a series of multilevel linear regressions adjusting for the selected covariates and random effects, one year of Head Start increased scores for PPVT ($\beta[95\%\ CI] = 5.66[4.05, 7.26]; d = 0.14$), Letter-Word Identification ($\beta[95\%\ CI] = 5.17[3.78, 6.55]; d = 0.19$), Applied Problem ($\beta[95\%\ CI] = 3.38[1.93, 4.84]; d = 0.12$), Spelling ($\beta[95\%\ CI] = 3.02[1.72, 4.31]; d = 0.12$), and Pre-Academic Skill ($\beta[95\%\ CI] = 3.82[2.81, 4.83]; d = 0.17$) (Table 2). Percent variation explained by the Head Start for PPVT, Letter-Word Identification, Applied Problem, Spelling, and Pre-Academic Skill were 1.34%, 1.57%, 0.60%, 0.56%, and 1.62%, respectively.

When being run at varying thresholds for dichotomizing outcomes, a series of multilevel logistic regressions adjusting for the covariates and random effects had an overall pattern of the Head Start increasing the odds of scoring high on cognitive outcomes with odds ratios ranging from 1.16 to 1.66 for PPVT, 1.47 to 1.77 for Letter-Word Identification, 1.07 to 1.34 for Applied Problems, 1.24 to 1.47 for Spelling, and 1.54 to 1.58 for Pre-Academic Skill (Table 3). In contrast, the Head Start did not meaningfully contribute to improvement in AUC with the difference ranging from 0.01 to 0.18% for PPVT, 0.13–0.49% for Letter-Word Identification, 0.00–0.12% for Applied Problems, 0.04–0.22% for Spelling, and 0.14–0.23% for Pre-Academic Skill. Moreover, the variance of predicted probabilities was negligibly affected by the addition of the Head Start in the model (Table 3; Fig. 1). In a logistic regression that only included the Head Start, the AUC for discriminating children with high cognitive scores ranged from 50.22 to 53.16% for PPVT, 53.98–55.24% for Letter-Word Identification, 50.66–52.02% for Applied Problem, 52.11–53.02% for Spelling, and 53.22–53.46% for Pre-Academic Skill.

Table 1
Sample characteristics at baseline by the treatment and control groups.

		Overall	Control	Head Start	Missing
N		4442	1796	2646	
Age cohort (%)	3	2449 (55.1)	985 (54.8)	1464 (55.3)	0
	4	1993 (44.9)	811 (45.2)	1182 (44.7)	
Gender (%)	male	2239 (50.4)	912 (50.8)	1327 (50.2)	0
Race/ethnicity (%)	White	1496 (33.7)	623 (34.7)	873 (33.0)	0
	Black	1348 (30.3)	536 (29.8)	812 (30.7)	
	Hispanic & others	1598 (36.0)	637 (35.5)	961 (36.3)	
Primary language (%)	English	3301 (74.3)	1345 (74.9)	1956 (73.9)	0
	Spanish	1141 (25.7)	451 (25.1)	690 (26.1)	
Parental education (%)	more	1274 (28.7)	505 (28.1)	769 (29.1)	0
	high school	1481 (33.3)	592 (33.0)	889 (33.6)	
	less	1687 (38.0)	699 (38.9)	988 (37.3)	
Single parent (%)		2239 (50.4)	907 (50.5)	1332 (50.3)	0
Recent immigrant (%)		855 (19.2)	337 (18.8)	518 (19.6)	0
Marital status (%)	married	1972 (44.4)	806 (44.9)	1166 (44.1)	0.1
	separated & divorced & widowed	724 (16.3)	290 (16.1)	434 (16.4)	
	never	1742 (39.2)	699 (38.9)	1043 (39.4)	
Special needs (%)		570 (12.8)	204 (11.4)	366 (13.8)	0
Teen mom (%)		752 (16.9)	330 (18.4)	422 (15.9)	0
Urban (%)		3746 (84.3)	1513 (84.2)	2233 (84.4)	0
Household risk (%)	low	3383 (76.2)	1399 (77.9)	1984 (75.0)	0
	moderate	741 (16.7)	277 (15.4)	464 (17.5)	
	high	318 (7.2)	120 (6.7)	198 (7.5)	
Caregiver's age (mean (SD))		28.91 (7.34)	28.65 (7.06)	29.08 (7.52)	0

4. Discussion

Using the HSIS data, we present findings in parallel for two distinct evaluation metrics: average effect and predictive power. Across the outcomes with meaningfully sized average effects after one year of the Head Start, we found that the predictive power was consistently small, regardless of whether the outcomes were continuous or binary. The negligible predictive power was also observed consistently across varying binary thresholds.

Average effect assessments on continuous outcomes were reproductions of previous studies. After one year of the Head Start, those who were assigned to the Head Start scored higher on a range of cognitive outcomes than those who were assigned to the control group. Across the continuous outcomes, the effect size was less than 0.2 in Cohen's *d*. While the effect size of 0.2 is considered as "small" based on Cohen's simple typology (Cohen, 1992), it can be considered "large" if

Table 2
Measures of average effect and predictive power on continuous outcomes.

		Average effect	Predictive power		
	Sample size (follow-up rate)	Model 1 regression coefficient (95% CI); Cohen's *d*	Model 1 child-level variance	Model 2 child-level variance	Percent variation explained by Head Start
Outcome					
PPVT	3621 (82%)	5.66 (4.05, 7.26); 0.14	557.74	565.29	1.34
Letter-Word Identification	3627 (82%)	5.17 (3.78, 6.55); 0.19	416.47	423.13	1.57
Applied Problem	3601 (81%)	3.38 (1.93, 4.84); 0.12	454.99	457.72	0.60
Spelling	3635 (82%)	3.02 (1.72, 4.31); 0.12	365.38	367.43	0.56
Pre-Academic Skill	3594 (81%)	3.82 (2.81, 4.83); 0.17	218.65	222.24	1.62

Note. CI, confidence intervals., PPVT, Peabody Picture Vocabulary Test.
Model 1: a full model with the Head Start.
Model 2: a full model without the Head Start.

Table 3
Measures of average effect and predictive power on binary outcomes (i.e., "high" scores).

	Average effect	Predictive power				
	Model 3	Model 3	Model 4	Model 4 → Model 3		Model 5
	OR (95% CI)	AUC (%)	AUC (%)	Improvement in AUC contributed to Head Start (%)	Change in the variance of predicted probability	AUC (%)
Outcome						
PPVT						
>0.25	1.66 (1.36, 2.02)	88.58	88.40	0.18	0.018	53.16
>0.50	1.42 (1.18, 1.71)	89.90	89.78	0.13	0.006	51.63
>0.75	1.16 (0.92, 1.45)	92.54	92.53	0.01	0.001	50.22
Letter-Word Identification						
>0.25	1.62 (1.38, 1.90)	80.44	79.95	0.49	0.033	54.56
>0.50	1.77 (1.50, 2.09)	83.72	83.06	0.65	0.034	55.24
>0.75	1.47 (1.20, 1.80)	87.05	86.92	0.13	0.010	53.98
Applied Problem						
>0.25	1.34 (1.11, 1.60)	85.62	85.50	0.12	0.008	52.02
>0.50	1.28 (1.08, 1.52)	85.87	85.76	0.12	0.005	51.56
>0.75	1.07 (0.86, 1.34)	90.14	90.14	0.00	0.000	50.66
Spelling						
>0.25	1.47 (1.24, 1.74)	84.56	84.34	0.22	0.013	53.02
>0.50	1.30 (1.09, 1.54)	85.49	85.36	0.13	0.005	52.12
>0.75	1.24 (1.01, 1.53)	88.42	88.38	0.04	0.004	52.11
Pre-Academic Skill						
>0.25	1.58 (1.30, 1.90)	86.42	86.21	0.22	0.017	53.22
>0.50	1.56 (1.30, 1.87)	88.21	87.98	0.23	0.011	53.23
>0.75	1.54 (1.23, 1.92)	91.33	91.20	0.14	0.009	53.46

Note. OR, odds ratio. CI, confidence intervals. AUC, area under the curve.
Model 3: a full model with the Head Start.
Model 4: a full model without the Head Start.
Model 5: a simple model with only the Head Start.

the cost of the Head Start is taken into account in a cost-effectiveness framework (Harris, 2009; Ludwig & Phillips, 2008). In the binary versions of the outcomes, the Head Start had consistent, positive effects across varying thresholds of dichotomization. For PPVT and Applied Problem, the Head Start effect was larger when the children were grouped by lower thresholds, which was in agreement with previous findings in which those with lower baseline cognitive ability benefited more from the Head Start (Bitler et al., 2014; Lee, Rodgers, et al., 2022). To clarify, estimating average effect of the Head Start in this study was not to question conclusions from previous studies (Bitler et al., 2014; Chor, 2018; Feller et al., 2016; Lipscomb et al., 2013; McCoy et al., 2016; Miller et al., 2016; Zhai et al., 2014), but to provide a parallel comparison to the predictive power assessment.

In contrast to the meaningfully sized effect on average, the predictive power assessment showed a weak ability of the Head Start to predict outcomes after one year of the study. In continuous outcomes, the Head Start explained far less than 2% of the total between-child variance for each outcome. While there are no standard guidelines, one report has suggested using Cohen's typology in which 2% is considered "small" (Cohen, 1992; Lorah, 2018). In binary scenarios, the results are clearer. Regardless of the thresholds, the Head Start did not have any meaningful improvement in AUC with less than 1%. In addition, AUC of the Head Start alone was nearly identical to 50%, meaning that the Head Start predicts the outcomes almost at random. Furthermore, the predicted probabilities were only negligibly affected in terms of their variances. The distributions of the predicted probabilities were almost unchanged when the Head Start was considered, reflecting its limited influence on the predictive power.

Our findings collectively showed that the Head Start had some average effects of significance but negligible predictive power, suggesting that the heterogeneity in individual effects may be large. Indeed, many studies have explored heterogeneous effects (Lee et al., 2021), and several studies found that large amounts of the heterogeneity was unexplained by measured variables in the HSIS data (Ding et al., 2016, 2019; Lee, Rodgers, et al., 2022). Such heterogeneity may be due to the heterogeneous populations included in the HSIS or heterogeneous implementations of the Head Start programs across the United States. Either way, the large heterogeneity in effects may mean that the average effect estimates are not generalizable to different populations or different Head Start program settings.

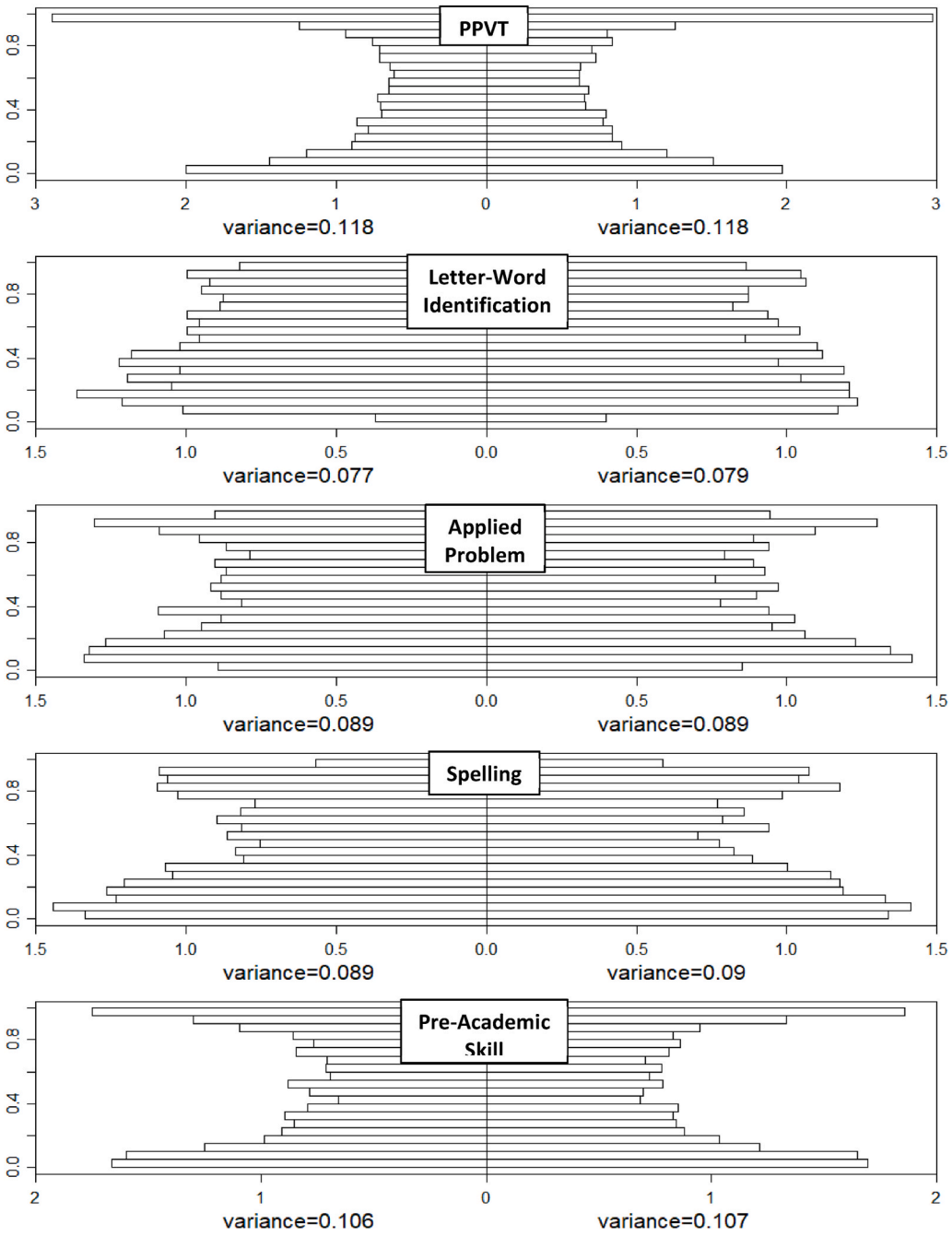

Fig. 1. Comparison of predicted probabilities with and without the Head Start. Predicted probabilities for "high" scores from Model 3 (with the Head Start; right) and Model 4 (without the Head Start; left) are shown. Probabilities are on the x-axis, and densities are the y-axis.

The present study has some limitations. First, measurement errors in the outcomes may inflate estimates of variance, erroneously decreasing the predictive power. However, such an error would not explain the level of predictive power we observed in this study. Second, we did not adjust for noncompliance of the treatment status. Previous works has already thoroughly investigated the Head Start effects adjusting for noncompliance. We did not have an aim at replicating this, and our intent-to-treat estimates are still unbiased and can achieve our study objective of comparing predictive power to average effect. If noncompliance was adjusted, the average effect and predictive power may have been slightly higher, but based on previous simulations on the relationships between average effect and predictive power (Pepe et al., 2004; Wald et al., 1999), our conclusion would not change.

Using the RCT data of the Head Start, we found that while the Head Start had significant effects on cognitive outcomes on average, it did not have an ability to predict the outcomes well. This indicates that the heterogeneity in the individual effects across children is quite large. While the Head Start has been labeled as cost-effective and indeed has positive effects on cognitive outcomes on average, the magnitude of heterogeneity in individual effects should also be assessed. Furthermore, tailoring the program to specific subgroups or settings may be important in the case of the Head Start. Assessment of the predictive power of a causal variable in randomized data should be a routine practice as it can provide helpful information on the causal effect and especially its heterogeneity. Beyond the HSIS data, the predictive power of treatments, interventions, and programs that are well-known to be effective on average should be assessed to gain a better understanding of their effects.

Funding

This work was supported by the Robert Wood Johnson Foundation (grant ID: 75602).

Author contributions

RK and SVS conceptualized and designed the study. SYL contributed to the conceptualization of the study, led interpretation of the data, conducted the final analyses, and wrote the manuscript. JR contributed to the initial analyses as well as to writing the first draft of the manuscript. All authors approved of the final draft.

Ethical statement

The HSIS data are not collected specifically for this study and no one on the study team has access to identifiers linked to the data. These activities do not meet the regulatory definition of human subject research. As such, an Institutional Review Board (IRB) review is not required. The Harvard Longwood Campus IRB allows researchers to self-determine when their research does not meet the requirements for IRB oversight via guidance online regarding when an IRB application is required using an IRB Decision Tool.

Declaration of competing interest

The authors have no competing interests.

Data availability

The authors do not have permission to share data.

References

Bitler, M. P., Hoynes, H. W., & Domina, T. (2014). Experimental evidence on distributional effects of head start (No. 20434; NBER working paper). *National Bureau of Economic Research, Inc*. https://www.nber.org/papers/w20434.pdf.

Chor, E. (2018). Multigenerational Head Start participation: An unexpected marker of progress. *Child Development, 89*(1), 264–279. https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/epdf/10.1111/cdev.12673.

Cintron, D. W., Adler, N. E., Gottlieb, L. M., Hagan, E., Tan, M. L., Vlahov, D., Glymour, M. M., & Matthay, E. C. (2022). Heterogeneous treatment effects in social policy studies: An assessment of contemporary articles in the health and social sciences. *Annals of Epidemiology*. https://doi.org/10.1016/J.ANNEPIDEM.2022.04.009

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Ding, P., Feller, A., & Miratrix, L. (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B, 78*(3), 655–671. http://wileyonlinelibrary.com/journal/rss-datasets.

Ding, P., Feller, A., & Miratrix, L. (2019). Decomposing treatment effect variation. *Journal of the American Statistical Association, 114*(525), 304–317. https://doi.org/10.1080/01621459.2017.1407322

Feller, A., Grindal, T., Miratrix, L., & Page, L. (2016). Compared to what? Variation in the impacts of early childhood education by alternative care type. *Annals of Applied Statistics, 10*(3), 1245–1285. https://doi.org/10.1214/16-aoas910

Harris, D. N. (2009). Toward policy-relevant benchmarks for interpreting effect sizes: Combining effects with costs. *Educational Evaluation and Policy Analysis, 31*(1), 3–29. https://doi.org/10.3102/0162373708327524

Head Start & Early Head Start. (2020). First five years fund. https://www.ffyf.org/issues/head-start-early-head-start/?mc_cid=4c8abeeea8&mc_eid=e63ec363fd.

Kim, R., Kawachi, I., Coull, B. A., & Subramanian, S. V. (2018). Contribution of socioeconomic factors to the variation in body-mass index in 58 low-income and middle-income countries: An econometric analysis of multilevel data. *Lancet Global Health, 6*(7), e777–e786. https://doi.org/10.1016/S2214-109X(18)30232-8

Kravitz, R. L., Duan, N., & Braslow, J. (2004). Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *The Milbank Quarterly, 82*(4), 661–687. https://doi.org/10.1111/j.0887-378X.2004.00327.x

Lee, S. Y., Kim, R., Rodgers, J., & Subramanian, S. V. (2021). Treatment effect heterogeneity in the head start impact study: A systematic review of study characteristics and findings. *SSM - Population Health, 16*(100916). https://doi.org/10.1016/j.ssmph.2021.100916

Lee, S. Y., Kim, R., Rodgers, J., & Subramanian, S. V. (2022). Assessment of heterogeneous Head Start treatment effects on cognitive and social-emotional outcomes. *Scientific Reports, 12*(1), 1–11. https://doi.org/10.1038/s41598-022-10192-1

Lee, S. Y., Rodgers, J., Kim, R., & Subramanian, S. V. (2022). Distributional effects on children's cognitive and social-emotional outcomes in the head start impact study: A quantile regression approach. *SSM - Population Health, 18*, Article 101108. https://doi.org/10.1016/J.SSMPH.2022.101108

Lipscomb, S. T., Pratt, M. E., Schmitt, S. A., Pears, K. C., & Kim, H. K. (2013). School readiness in children living in non-parental care: Impacts of Head Start. *Journal of Applied Developmental Psychology, 34*, 28–37. https://doi.org/10.1016/j.appdev.2012.09.001

Lorah, J. (2018). Effect size measures for multilevel models: Definition, interpretation, and TIMSS example. *Large-Scale Assessments in Education, 6*(1), 1–11. https://doi.org/10.1186/S40536-018-0061-2/TABLES/1

Ludwig, J., & Phillips, D. A. (2008). Long-term effects of head start on low-income children. *Annals of the New York Academy of Sciences, 1136*, 257–268. https://doi.org/10.1196/annals.1425.005

McCoy, D. C., Morris, P. A., Connors, M. C., Gomez, C. J., & Yoshikawa, H. (2016). Differential effectiveness of Head Start in urban and rural communities. *Journal of Applied Developmental Psychology, 43*, 29–42. https://doi.org/10.1016/j.appdev.2015.12.007

Merlo, J., Mulinari, S., Wemrell, M., Subramanian, S. V., & Hedblad, B. (2017). The tyranny of the averages and the indiscriminate use of risk factors in public health: The case of coronary heart disease. *SSM - Population Health, 3*, 684–698. https://doi.org/10.1016/j.ssmph.2017.08.005

Miller, E. B., Farkas, G., & Duncan, G. J. (2016). Does Head Start differentially benefit children with risks targeted by the program's service model? *Early Childhood Research Quarterly, 34*, 1–12.

Pepe, M. S., Janes, H., Longton, G., Leisenring, W., & Newcomb, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology, 159*(9), 882–890. https://doi.org/10.1093/aje/kwh101

Plewis, I. (2002). Modelling impact heterogeneity. *Journal of the Royal Statistical Society - Series A: Statistics in Society, 165*(1), 31–38. https://doi.org/10.1111/1467-985X.0asp1

Puma, M., Bell, S., Cook, R., Heid, C., Shapiro, G., Broene, P., Jenkins, F., Fletcher, P., Quinn, L., Friedman, J., Ciarico, J., Rohacek, M., Adams, G., & Spier, E. (2010). *Head Start impact study final report*.

Puma, M., Bell, S., Cook, R., Heid, C., Shapiro, G., Broene, P., Jenkins, F., Fletcher, P., Quinn, L., Friedman, J., Ciarico, J., Rohacek, M., Adams, G., & Spier, E. (2011). Head start impact study technical report. https://www.acf.hhs.gov/sites/default/files/opre/hs_impact_study_tech_rpt.pdf.

Puma, M., Cook, R., Heid, C., Broene, P., Jenkins, F., Mashburn, A., & Downer, J. (2012). Third grade follow-up to the head start impact study final report. http://www.acf.hhs.gov/programs/opre.

R Core Team. (2020). *R: A language and environment for statistical computing*. (4.0.3). *R Foundation for Statistical Computing*. https://www.r-project.org/.

Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289–310. https://doi.org/10.1214/10-STS330

Subramanian, S. V., Kim, R., & Christakis, N. A. (2018). The "average" treatment effect: A construct ripe for retirement. A commentary on deaton and cartwright. *Social Science & Medicine, 210*, 77–82. https://doi.org/10.1016/j.socscimed.2018.04.027

Swaminathan, A., Kim, R., & Subramanian, S. V. (2020). Association does not imply prediction: The accuracy of birthweight in predicting child mortality and anthropometric failure. *Annals of Epidemiology, 50*, 7–14. https://doi.org/10.1016/j.annepidem.2020.08.001

Varga, T. V., Niss, K., Estampador, A. C., Collin, C. B., & Moseley, P. L. (2020). Association is not prediction: A landscape of confused reporting in diabetes – a systematic review. *Diabetes Research and Clinical Practice, 170*, Article 108497. https://doi.org/10.1016/j.diabres.2020.108497

Wald, N. J., Hackshaw, A. K., & Frost, C. D. (1999). When can a risk factor be used as a worthwhile screening test? *BMJ, 319*(7224), 1562. https://doi.org/10.1136/bmj.319.7224.1562

Zhai, F., Brooks-Gunn, J., & Waldfogel, J. (2014). Head Start's impact is contingent on alternative type of care in comparison group. *Developmental Psychology, 50*(12), 2572–2586. https://doi.org/10.1037/a0038205.NS-