MULTIVARIATE BEHAVIORAL RESEARCH
2023, VOL. 58, NO. 2, 221–240

Identifying and Estimating Causal Moderation for Treated and Targeted Subgroups

Nianbo Dong[a], Benjamin Kelcey[b], and Jessaca Spybrook[c]

[a]University of North Carolina at Chapel Hill; [b]University of Cincinnati; [c]Western Michigan University

ABSTRACT

Extant literature on moderation effects narrowly focuses on the average moderated treatment effect across the entire sample (AMTE). Missing is the average moderated treatment effect on the treated (AMTT) and other targeted subgroups (AMTS). Much like the average treatment effect on the treated (ATT) for main effects, the AMTS changes the target of inferences from the entire sample to targeted subgroups. Relative to the AMTE, the AMTS is identified under weaker assumptions and often captures more policy-relevant effects. We present a theoretical framework that introduces the AMTS under the potential outcomes framework and delineates the assumptions for causal identification. We then propose a generalized propensity score method as a tool to estimate the AMTS using weights derived with Bayes Theorem. We illustrate the results and differences among the estimands using data from the Early Childhood Longitudinal Study. We conclude with suggestions for future research.

KEYWORDS

Average moderated treatment effect across the entire sample (AMTE); average moderated treatment effect on the treated (AMTT); average moderated treatment effect on targeted subgroups (AMTS); generalized propensity scores; potential outcome

Introduction

A critical consideration in making causal inferences from a sample is the a priori specification of the target population and definition of the causal parameter of interest (e.g., Ahern, 2018; Hernán, 2018). Causal inference researchers have repeatedly distinguished among different types of effects based on different samples and inferential targets. For example, prior research has distinguished among several different types of main effects of a treatment including the average effect of the treatment on the treated (ATT), the average treatment effect on the untreated (ATU), and the average treatment effect (ATE) (Imai et al., 2008; Imbens, 2004; Kurth et al., 2006; McCaffrey et al., 2004; Ridgeway et al., 2021).

Based on the potential outcomes framework (Neyman, 1923/1990; Rubin, 1974), the ATE contrasts the potential outcomes (Y) for those in the treated and untreated conditions: $E[Y(1) - Y(0)]$ with $E[]$ as the expectation operator, $Y(0)$ as the potential outcome under the untreated condition, and $Y(1)$ as the potential outcome under the treatment condition. The ATT also contrasts the expected outcomes across conditions but does so conditional upon receipt of the treatment ($Z = 1$): $E[Y(1) - Y(0)|Z = 1]$. Conceptually, this estimand captures the average treatment effect for those who were materially exposed to the treatment condition. The untreated counterpart of this estimand, the average treatment effect of the treatment on the untreated (ATU), contrasts the potential outcomes across conditions conditional upon receipt of a control or comparative condition ($Z = 0$) and is defined as $E[Y(1) - Y(0)|Z = 0]$. ATU represents the average treatment effect for those in the untreated group should they receive the treatment.

The scope of inference for treatment effects has also expanded to complements of the main treatment effect. For example, researchers and policy makers are increasingly interested in differential (moderated) treatment effects associated with dissimilar subgroups based on pretreatment variables (Aiken & West, 1991; Baron & Kenny, 1986; Frazier et al., 2004; Kraemer et al., 2001, 2008).

A principal finding in this literature suggests that effects and inferences may critically diverge in different samples when assumptions are violated and/or when individuals in treatment conditions

systematically differ (e.g., Bun & Harrison, 2019; Dong, 2012, 2015). Although literature has thoroughly documented these considerations for main effects (e.g., Dong et al., 2020; Mayer et al., 2016; Yang et al., 2021), the moderation effects counterpart to this literature has largely focused only on the average moderated treatment effect (AMTE) for the entire sample under the potential outcomes framework (e.g., Bansak, 2018; Dong, 2012, 2015; Dong & Kelcey, 2020; Egami & Imai, 2019). There is little to no research on the average moderator effects on subsamples; that is, the analogous ATT/ATU version of the average moderated treatment effect on targeted subgroups (AMTS) (e.g., the average moderated treatment effect on the treated subgroup) has not been studied and well defined.

The purpose of this article is to develop the average moderated treatment effect on targeted subgroups (AMTS) based on the potential outcomes framework (Neyman, 1923/1990; Rubin, 1974), delineate identification assumptions, and to develop an estimator. The remainder of the paper is organized as follows. First, we introduce a motivating example that focuses on the main effect of preschool on the academic achievement for all children and the differential effect of preschool on the academic achievement for children with different home language background. Second, we review the ATE, ATT, and ATU, and discuss their assumptions for causal inference. Third, we present a theoretical framework that introduces the AMTE and AMTS definitions for a binary moderator under the potential outcomes framework and delineates the assumptions for causal identification of the AMTE and AMTS. Then we propose the generalized propensity score method as a tool to estimate the AMTE and AMTS using the weights derived based on Bayes Theorem. Fourth, we demonstrate the application of our proposed definitions and estimation methods to the motivating example. Finally, we discuss our findings and conclude with some suggestions of future directions of research.

Motivating example

Our example focuses on the main and differential effect of preschool on the academic achievement of children with different home language background. The early childhood care and education (ECCE) programs, such as center-based programs like preschool, pre-Kindergarten, and Head Start seek to close the achievement gap at school entry. Some studies indicate positive main effects of ECCE on student's academic achievement (e.g., Magnuson et al., 2007) while several studies indicate mixed effects of ECCE on student's academic achievement (e.g., Barnett, 2011; Lipsey et al., 2013). Further, Lipsey et al. (2013) suggested the effects of ECCE may differ for certain subgroups and found that that non-native English speaking children experienced greater benefit in terms of academic achievement from the Tennessee voluntary prekindergarten program than the native English speaking children during the pre-k year but less benefit in kindergarten and the first grade. Given the mixed findings, the policy questions in this example include: (1) Is there a main effect of preschool (treated) compared to parental care (untreated) on the academic achievement for all children? (2) Is there a differential (moderated) effect of preschool on the academic achievement for children with different home language background (moderator: speaking English at home or not)? In this example, both the treatment and moderator variables are dichotomous.

Review of ATE, ATT, and ATU

To illustrate the differences among ATE, ATT, and ATU, we consider the motivating example for which we would like to evaluate the main effect of a dichotomous treatment (i.e., preschool vs. parental care). Assume the potential outcomes can be expressed as a linear function such that:

$$Y_i(Z) = \beta_0 + \beta_1 Z_i + \beta_2 X_i + \beta_3 X_i Z_i + e_i, e_i \sim N\left(0, \sigma^2\right), \tag{1}$$

where $Y_i(Z)$ is the potential outcome for subject i receiving treatment Z. Z_i represents the treatment status: 1 for the treated condition (preschool), and 0 for the untreated condition (parental care). X_i is a baseline moderating covariate for home language background: 1 for speaking English at home, and 0 for not speaking English at home. The coefficient, β_1, is the treatment effect of preschool when $X_i = 0$ (not speaking English at home), and β_3 is the moderated treatment effect that depends on the value of the covariate (moderator), X_i. Under this simple example, the ATE is estimated as

$$\begin{aligned} ATE &= E\left[Y(1) - Y(0)\right] \\ &= E[(\beta_0 + \beta_1 + \beta_2 X_i + \beta_3 X_i + e_i) - (\beta_0 + \beta_2 X_i + e_i)] \\ &= E[\beta_1 + \beta_3 X_i] = \beta_1 + \beta_3 E(X_i). \end{aligned} \tag{2}$$

Conceptually, the ATE summarizes the average effect for the entire sample by taking the (unconditional) expectation of the moderating covariate over

treated and untreated conditions. Similarly, the ATT and ATU in this example are estimated as

$$\begin{aligned} ATT &= E[Y(1) - Y(0)|Z = 1] = E[((\beta_0 + \beta_1 + \beta_2 X_i + \beta_3 X_i + e_i) \\ &\quad - (\beta_0 + \beta_2 X_i + e_i))|Z = 1] = E[(\beta_1 + \beta_3 X_i)|Z = 1] \\ &= \beta_1 + \beta_3 E(X_i|Z = 1)\Big). \end{aligned} \tag{3}$$

$$\begin{aligned} ATU &= E[Y(1) - Y(0)|Z = 0] = E[((\beta_0 + \beta_1 + \beta_2 X_i + \beta_3 X_i + e_i) \\ &\quad - (\beta_0 + \beta_2 X_i + e_i))|Z = 0] = E[(\beta_1 + \beta_3 X_i)|Z = 0] \\ &= \beta_1 + \beta_3 E(X_i|Z = 0)\Big). \end{aligned} \tag{4}$$

In contrast to the ATE, the ATT (or ATU) describes the average effect for only those that took up the treatment (or untreated) by taking the expectation of the moderating covariate conditional upon treatment status.

Prior research has demonstrated these connections by showing that the ATE is the weighted average of ATT and ATU (Abadie & Imbens, 2008). More specifically, the $ATE = \frac{n_t}{n_t+n_c} ATT + \frac{n_c}{n_t+n_c} ATU$, where n_t, and n_c are sample sizes for the treatment and control groups. In a randomized trial with full treatment compliance the ATT, ATU, and ATE are all equal in expectation because the treatment and control samples and their covariate distributions are similar due to random assignment, i.e., $E(X_i Z = 1) = E(X_i Z = 0) = E(X_i)$. However, in a randomized trial with treatment noncompliance[1] or a nonrandomized study, the three estimands may differ because the samples in the treatment and comparison groups may be systematically different due to treatment noncompliance or self-selection, e.g., $\mathrm{E}(X_i Z = 1) \neq E(X_i Z = 0)$. As a result, when a treatment effect is moderated by a covariate ($\beta_3 \neq 0$), the treatment effects diverge across the different samples.

In nonrandomized studies, the distinctions among the ATE, ATT, and ATU are useful from both theoretical and practical standpoints. Theoretically, for example, the adoption of the ATT can be used to partially relax identification assumptions that undergird much of the causal inference framework for the ATE (See Moreno-Serra, 2007 for a review). Under the potential outcomes framework (Neyman, 1923/1990; Rubin, 1974), identification of the ATE requires two key assumptions (in addition to other assumptions):

$$\{Y(0), Y(1)\} \perp Z|\boldsymbol{X}, \tag{5}$$

$$0 < Pr(Z = 1|\boldsymbol{X}) < 1, \tag{6}$$

where $\boldsymbol{X}$ is a vector of covariates, and $Pr(Z = 1|\boldsymbol{X})$ is the probability of being in the treatment group conditional on the covariates. The first assumption (Eq. 5) is often known as unconfoundedness or ignorable treatment assignment (Rosenbaum & Rubin, 1983) and is commonly referred to as selection on observed variables. This assumption requires the set of potential outcomes be independent of the treatment assignment conditional upon the observed covariates. The second assumption (Eq. 6) is often referred to as common support or overlap and requires that the probability of receiving treatment for each level of the covariates is between zero and one (i.e., no one receives treatment or control with certainty). Rosenbaum and Rubin (1983) referred to the combination of the first and second assumptions as "strong ignorability".

Both assumptions can be weakened when taking up the ATT and ATU. In particular, identification of the ATT only requires relaxed versions of the original assumptions (in addition to other assumptions):

$$Y(0) \perp Z|\boldsymbol{X}, \tag{7}$$

$$Pr(Z = 1|\boldsymbol{X}) < 1. \tag{8}$$

Similarly, identification of the ATU requires the assumptions:

$$Y(1) \perp Z|\boldsymbol{X}, \tag{9}$$

$$0 < Pr(Z = 1\boldsymbol{X}). \tag{10}$$

Under the ATT (or ATU), the first assumption (expression 7 or 9) is known as weak unconfoundedness. This assumption is a weaker version of its ATE counterpart assumption (i.e., Eq. 5). For example, for the ATT, the moments of the distribution of $Y(1)$ for the treated are directly measurable and the assumption only requires that the potential outcome under the control condition is independent of the treatment assignment given the observed variables. In parallel, for the ATU, the moments of the distribution of $Y(0)$ for the untreated are directly measurable and the assumption only requires that the potential outcome under the treatment condition is independent of the treatment assignment given the observed variables. Similarly, the second ATT (or ATU) assumption (Eq. 8 or 10) captures what is commonly referred to as weak overlap or common support because it requires only that the probability of receiving treatment for each level of the covariates is less than one, i.e., no one receives treatment with certainty, (or more than 0, i.e., no one receives control with certainty).

There is also practical purchase in differentiating among the ATE, ATT, and ATU. Research projects take up a broad range of foci that leverage different designs and necessitate different targets of inference for summarizing treatment effects. For example, the students who speak English at home may go to

[1]See Angrist et al. (1996) and Sagarin et al. (2014) for more discussion about treatment noncompliance.

preschool at a much higher rate than their counterparts who do not speak English at home. In such settings, researchers may have different interest in the ATT, ATU, and ATE because the different samples represent different policy targets; e.g., the ATU captures the effect of preschool on the sample with more immigrant children who do not speak English at home, a policy-relevant segment of the population.

Theoretical framework

Just as the distinction among ATE, ATT, and ATU can be used to understand requisite assumptions and probe a diverse set of research purposes when detailing the main effect, the distinction between the AMTE and AMTS can be useful for relaxing the assumptions necessary for causal inference while aligning research goals, policy, and estimands. For instance, when we investigate whether the effect of preschool was moderated by students' home language background, we can distinguish between the AMTE and several versions of the AMTS. The AMTE describes the average moderated treatment effect across the entire sample; that is, it describes the extent to which treatment effects varied as a function of students' home language background for the entire sample regardless of their selected treatment status. In contrast, the AMTS decomposes this overall summary into the average moderated treatment effect for targeted subgroups while diminishing identification assumptions as outlined above.

As an example, consider a conceptual counterpart of the ATT, the average moderated treatment effect on the treated (AMTT). The AMTT can be used to capture the moderation effects owing to the home language background for those who selected into preschool. This estimand is conceptually analogous to the ATT in that it describes the moderation effect for only the portion of the sample that received treatment. Alternatively, more fine-grained distinctions can also be made using the AMTS—for instance, we can narrowly describe the moderation effects for just those students who were exposed to preschool and also spoke English at home. As we detail below, when appropriate, the shift has theoretical and practical advantages that parallel the differences between ATE and the ATT/ATU described above.

Potential outcomes framework for causal moderation analysis

When a potentially manipulable pretreatment covariate moderates a treatment effect, the potential outcomes for participant i depend on both the treatment status (Z) and moderator value (R). In the case of a dichotomous treatment and a dichotomous moderator, we can define the potential outcomes for participants with reference to their statuses on these variables as

		Moderator (R)	
		0	1
Treatment (Z)	0	$Y(0,0)$	$Y(0,1)$
	1	$Y(1,0)$	$Y(1,1)$

Figure 1. Potential outcomes $Y\ (Z = z, R = r)$

$$\text{PotentialOutcome}: = Y_i(Z = z, R = r), \tag{11}$$

where Y_i is the potential response for individual i when the treatment (Z) is set to z ($Z = 0$ for untreated, e.g., parental care, or 1 for treated, e.g., preschool) and the moderator (R) is set to r ($R = 0$ for the reference moderator subgroup, e.g., not speaking English at home, or $R = 1$ for the moderator subgroup, e.g., speaking English at home) (Dong, 2012, 2015; Dong & Kelcey, 2020). The potential outcomes for the causal moderated treatment effect are presented in Figure 1. Under this definition, each individual has four potential outcomes: (a) $Y(0,0)$, (b) $Y(0,1)$, (c) $Y(1,0)$, and (d) $Y(1,1)$.

Similar to definitions for the main effect that distinguish the ATE and the average treatment effect on subsamples (e.g., ATT for the treated sample) (Imai et al., 2008; Imbens, 2004; Kurth et al., 2006; McCaffrey et al., 2004), we differentiate between two general types of moderation effects: (a) the average moderated treatment effect (AMTE) that pertains to the entire sample and (b) the average moderated treatment effects on targeted subgroups (AMTS) that pertain to selected subgroups.

The AMTE can be defined using the contrasts among four potential outcomes:

$$\begin{aligned} AMTE &= E[Y(1,1) - Y(0,1)] - E[Y(1,0) - Y(0,0)] \\ &= E[Y(1,1) - Y(0,0)] - E[Y(1,0) - Y(0,0)] - E[Y(0,1) - Y(0,0)] \\ &= E[Y(1,1) - Y(1,0)] - E[Y(0,1) - Y(0,0)]. \end{aligned} \tag{12}$$

The AMTE is the difference in the average treatment effects between the moderator subgroup $R = 1$

Table 1. Summary of formulas for various AMTS estimands.

Estimand	Formula
$AMTS_{(Z=0,R=0)}$	$= E[Y(1,1) - Y(0,1)\|Z=0,\ R=0] - E[Y(1,0) - Y(0,0)\|Z=0,\ R=0]$
$AMTS_{(Z=0,R=1)}$	$= E[Y(1,1) - Y(0,1)\|Z=0,\ R=1] - E[Y(1,0) - Y(0,0)\|Z=0,\ R=1]$
$AMTS_{(Z=1,R=0)}$	$= E[Y(1,1) - Y(0,1)\|Z=1,\ R=0] - E[Y(1,0) - Y(0,0)\|Z=1,\ R=0]$
$AMTS_{(Z=1,R=1)}$	$= E[Y(1,1) - Y(0,1)\|Z=1,\ R=1] - E[Y(1,0) - Y(0,0)\|Z=1,\ R=1]$
$AMTS_{(Z=1)}$	$= E[Y(1,1) - Y(0,1)\|Z=1] - E[Y(1,0) - Y(0,0)\|Z=1]$
$AMTS_{(Z=0)}$	$= E[Y(1,1) - Y(0,1)\|Z=0] - E[Y(1,0) - Y(0,0)\|Z=0]$
$AMTS_{(R=1)}$	$= E[Y(1,1) - Y(0,1)\|R=1] - E[Y(1,0) - Y(0,0)\|R=1]$
$AMTS_{(R=0)}$	$= E[Y(1,1) - Y(0,1)\|R=0] - E[Y(1,0) - Y(0,0)\|R=0]$

(i.e., $E[Y(1,1) - Y(0,1)]$) and the reference moderator subgroup $R=0$ (i.e., $E[Y(1,0) - Y(0,0)]$), for the entire sample. The AMTE measures the additional effect of both treatment and moderator beyond the average effects of treatment ($E[Y(1,0) - Y(0,0)]$) and moderator ($E[Y(0,1) - Y(0,0)]$) in the total effect of treatment and moderator ($E[Y(1,1) - Y(0,0)]$).

Alternatively, the AMTE can be regarded as the difference in the average moderator subgroup differences (gaps) between the treated group $Z=1$ (i.e., $E[Y(1,1) - Y(1,0)]$) and the untreated group $Z=0$ (i.e., $E[Y(0,1) - Y(0,0)]$), for the entire sample. More conceptually, this AMTE definition aligns with the interaction effect for a factorial design where two concurrent treatments exist by Hong (2015), the average marginal interaction effect (AMIE) by Egami and Imai (2019), and the average treatment moderation effect (ATME) by Bansak (2018).

AMTS

In contrast, the average moderated treatment effect on targeted subgroups (AMTS) focuses on the difference among the potential outcomes for a specific subgroup ($Z=z$ and/or $R=r$), and it can be defined as:

$$AMTS_{(Z=z,R=r)} = E[Y(1,1) - Y(0,1)|Z=z, R=r] - E\big[Y(1,0) - Y(0,0)|Z=z, R=r\big], \tag{13}$$

where $z=0$ for untreated or 1 for treated, and $r=0$ for the reference moderator subgroup or 1 for the moderator subgroup.

The AMTS for targeted subgroups are summarized in Table 1. For example, the $AMTS_{(Z=0,R=0)}$ is the average treatment effect difference between the students who spoke English at home ($r=1$) and the students who did not speak English at home ($r=0$) for those that had similar characteristics with the students who were in parental care ($Z=0$) and did not speak English at home ($R=0$).

In addition, we can leverage the AMTS to describe effect differences for subsamples that solely condition upon the treatment status. The $AMTS_{(Z=1)}$ describes the expected treatment effect difference between students who spoke and did not speak English at home for those that were exposed to the treatment (preschool), i.e., the average moderated treatment effect on the treated (AMTT). Similarly, the $AMTS_{(Z=0)}$ describes the expected treatment effect difference between students who spoke and did not speak English at home for those that were exposed to the untreated condition (parental care), i.e., the average moderated treatment effect on the untreated (AMTU).

We can also detail similar distinctions for moderator-based subsamples. For example, the $AMTS_{(R=1)}$ describes the expected treatment effect difference between students who spoke and did not speak English at home for those that had similar characteristics with students who spoke English at home. Similarly, the $AMTS_{(R=0)}$ the expected treatment effect difference between students who spoke and did not speak English at home for those that had similar characteristics with students who did not speak English at home.

If both treatment and moderator are randomly assigned, the AMTS will be equal across all four treatment-by-moderator subgroups and the other subgroups defined solely upon the treatment or moderator variable, and the AMTS is equal to the AMTE. However, if either the treatment or the moderator is not randomly assigned, the AMTS may differ across subgroups because the subsamples and thus the covariate distributions across subgroups may be different.

In general, the AMTE equals the weighted sum of AMTS across four subgroups with weights based on the proportion of total individuals in each subgroup, that is,

$$\text{AMTE} = \frac{n_{(Z=0,R=0)}}{N} AMTS_{(Z=0,R=0)} + \frac{n_{(Z=0,R=1)}}{N} \text{AMTS}_{(Z=0,R=1)} + \frac{n_{(Z=1,R=0)}}{N} \text{AMTS}_{(Z=1,R=0)} + \frac{n_{(Z=1,R=1)}}{N} \text{AMTS}_{(Z=1,R=1)}, \tag{14}$$

where $n_{(Z=0,R=0)}$, $n_{(Z=0,R=1)}$, $n_{(Z=1,R=0)}$, and $n_{(Z=1,R=1)}$ are sample sizes for four treatment-by-moderator subgroups, and $N = n_{(Z=0,R=0)} + n_{(Z=0,R=1)} + n_{(Z=1,R=0)} + n_{(Z=1,R=1)}$. The $AMTS_{(Z=1)}$, $AMTS_{(Z=0)}$,

Table 2. Summary of formulas for calculation of AMTS for the subgroup that is solely based on treatment or moderator.

Estimand	Formula
$AMTS_{(Z=1)}$	$= \frac{n_{(Z=1,R=0)}}{N_{(Z=1)}} AMTS_{(Z=1,R=0)} + \frac{n_{(Z=1,R=1)}}{N_{(Z=1)}} AMTS_{(Z=1,R=1)}$
$AMTS_{(Z=0)}$	$= \frac{n_{(Z=0,R=0)}}{N_{(Z=0)}} AMTS_{(Z=0,R=0)} + \frac{n_{(Z=0,R=1)}}{N_{(Z=0)}} AMTS_{(Z=0,R=1)}$
$AMTS_{(R=1)}$	$= \frac{n_{(Z=0,R=1)}}{N_{(R=1)}} AMTS_{(Z=0,R=1)} + \frac{n_{(Z=1,R=1)}}{N_{(R=1)}} AMTS_{(Z=1,R=1)}$
$AMTS_{(R=0)}$	$= \frac{n_{(Z=0,R=0)}}{N_{(R=0)}} AMTS_{(Z=0,R=0)} + \frac{n_{(Z=1,R=0)}}{N_{(R=0)}} AMTS_{(Z=1,R=0)}$

Note: $n_{(Z=0,R=0)}$, $n_{(Z=0,R=1)}$, $n_{(Z=1,R=0)}$, and $n_{(Z=1,R=1)}$ are sample sizes for four treatment-by-moderator subgroups. $N_{(Z=1)} = n_{(Z=1,R=0)} + n_{(Z=1,R=1)}$, $N_{(Z=0)} = n_{(Z=0,R=0)} + n_{(Z=0,R=1)}$, $N_{(R=1)} = n_{(Z=0,R=1)} + n_{(Z=1,R=1)}$, and $N_{(R=0)} = n_{(Z=0,R=0)} + n_{(Z=1,R=0)}$.

$AMTS_{(R=1)}$, and $AMTS_{(R=0)}$ follow a similar pattern and are shown in Table 2.

Assumptions for AMTE and AMTS

The assumptions for the causal AMTE are analogous to the assumptions for the factorial design with two concurrent treatments (e.g., Egami & Imai, 2019) in that the moderator is potentially manipulable[2]:

1. The stable unit treatment and moderator value assumption (SUTMVA). The potential outcome for one unit should be unaffected by the particular assignment of treatments or moderators to the other units and there is only one version of the treatment and the moderator. This assumption extends the single treatment variable version of SUTVA (Rubin, 1980) to the two-variable version (Egami & Imai, 2019). That is, the extension applies equally to treatment assignments and moderator values in that the effects are only identified when there is no influence of one student's treatment or moderator value on the potential outcomes of another student. The extension also applies to the intersections or combinations of the treatment and moderator. That is, the potential outcomes of a student must also be independent of the treatment by moderator values of another student. Applied to our preschool example, this assumption is violated when, for example, the proportion of the students who spoke English at home and selected into the treatment condition influences the potential outcomes of students. This can arise, for instance, when an immigrant student who did not speak English at home becomes disheartened or discouraged by the dominance of the nonimmigrant students in preschool such that it alters his/her potential outcomes.
2. Ignorability of the treatment and moderator given covariates (Egami & Imai, 2019). The assignment mechanism for the treatment and moderator does not depend on potential outcomes given observable covariates. That is, $\{Y(0,0), Y(0,1), Y(1,0), Y(1,1)\} \perp (Z,R)|\boldsymbol{X}$, where $\boldsymbol{X}$ is a vector of covariates. This assumption requires that the potential outcomes given covariates are independent of the treatment and moderator status. Put differently, there are no variables that confound the relationships between the outcome, treatment and moderator. In a randomized experiment, this assumption automatically holds for the treatments, but not necessarily for the moderators. In nonrandomized studies, this necessitates that both the treatment and moderator assignment is independent of the potential outcomes conditional upon observed covariates. Applied to our preschool example, when the assignment of treatment (preschool) is random, if the home language status is not randomly assigned, this assumption can be violated when other covariates (e.g., socio-economic status (SES)) that are correlated with the home language status and affect the potential outcome are not appropriately accounted for (Dong, 2015).
3. Independence of the treatment and moderator. The treatment and moderator are independent given covariates: $Z \perp R|\boldsymbol{X}$. This assumption holds in all randomized studies because of the random assignment of treatment. In nonrandomized studies, however, this necessitates that, for example, treatment assignment is independent of the moderator conditional upon observed covariates. Applied to our preschool example, when the assignment of treatment (preschool) is not random this assumption can be violated when, for example, higher SES students who speak English at home tend to go to preschool.
4. Treatment-by-moderator common support: $0 < Pr(Z,R|\boldsymbol{X}) < 1$. The assumption requires the overlap of the sample among the treatment-by-moderator subgroups, i.e., the probability of an individual in either of the four groups should be between 0 and 1. This assumption may not automatically hold in randomized experiments where treatment is randomized because the moderator may not be randomized, and it is necessary for both randomized and nonrandomized studies. In

[2]Rubin and others have argued that a causal effect cannot be defined without at least a clear hypothetical manipulation (e.g., Rubin, 1986, 2010). To claim a causal moderator effect, the moderator needs to be potentially manipulable to mimic some hypothetical factorial experiments.

our preschool example, this assumption requires that each student has a nonzero probability to be in all four treatment (preschool)-by-moderator (home language background) subgroups.

Similar to the contrast between the assumptions for the ATT/ATU and ATE, the assumptions buttressing strong ignorability (2 and 4) can be weakened for the AMTS because the potential outcomes for the targeted inference group are directly measurable and only assumptions about the potential outcomes under the comparison subgroups are needed for estimating the counterfactual.

5. The assignment mechanism for the treatment and moderator that are not for the targeted inference group do not depend on potential outcomes given observable covariates. That is, $\{Y(0,1), Y(1,0), Y(1,1)\} \perp (Z,R)|\boldsymbol{X}$ for $AMTS_{(Z=0,R=0)}$; $\{Y(0,0), Y(1,0), Y(1,1)\} \perp (Z,R)|\boldsymbol{X}$ for $AMTS_{(Z=0,R=1)}$; $\{Y(0,0), Y(0,1), Y(1,1)\} \perp (Z,R)|\boldsymbol{X}$ for $AMTS_{(Z=1,R=0)}$; $\{Y(0,0), Y(0,1), Y(1,0)\} \perp (Z,R)|\boldsymbol{X}$ for $AMTS_{(Z=1,R=1)}$.
6. The probability of being the targeted inference subgroup for an individual in the other three subgroups should be between 0 and 1.

Estimation of AMTS using the generalized propensity score

A common approach to estimating causal effects under the potential outcomes framework is the use of propensity scores (Rosenbaum & Rubin, 1983). We draw on this approach to estimate AMTE and AMTS. When the treatment variable is dichotomous, the propensity score is the probability of being in the treatment group given the covariates (Rosenbaum & Rubin, 1983). Imbens (2000) extended it to treatments with multiple categories, i.e., the generalized propensity score. The generalized propensity score is the conditional probability of receiving treatment z given pretreatment covariate $\boldsymbol{X}$, i.e., $\pi = Pr(Z = z|\boldsymbol{X})$. The inverse of the generalized propensity score as a weight can be used to estimate the causal effects of multi-valued treatments (Imbens, 2000). Dong (2015) applied the generalized propensity score method to estimate the AMTE by collapsing two dimensions (treatment and moderator) to one dimension (a variable with multiple categories). Dong's (2015) simulation demonstrated good performance of the generalized propensity score in estimating the effects of two variables on one outcome. We extend Dong's (2015) work to apply the generalized propensity score method to estimate the AMTS. We use Bayes Theorem to derive the weights based on the generalized propensity score to estimate the AMTS and AMTE. The procedure follows.

(1) We first convert the two dimensions (treatment by moderator, 2×2) of design to one dimension of design with 4 categories by creating a new independent variable, S, where $S = 1$ if $Z = 0$ and $R = 0$, $S = 2$ if $Z = 0$ and $R = 1$, $S = 3$ if $Z = 1$ and $R = 0$, and $S = 4$ if $Z = 1$ and $R = 1$. This step converts the estimation of the effects of two predictors to the estimation of the effect of one predictor with four values.

(2) We then estimate generalized propensity scores (Imbens, 2000). For instance, we can use multinomial logistic regression, random forests, or boosted regression (Cham & West, 2016; McCaffrey et al., 2004) to estimate the generalized propensity scores for individual i of being in a certain category/subgroup given covariates ($\boldsymbol{X}$): $\pi_i(s) = \Pr(S_i = s|\boldsymbol{X}_i)$, where $s = 1$, 2, 3, or 4. Note that although the coefficients of the covariates may vary depending on which reference outcome subgroup is used, the probability of being in a certain subgroup will not change with the reference subgroup (Long, 1997). Each individual has four generalized propensity scores, among which, one is the probability of being in the actual/observed subgroup and the other three are the probabilities of potentially being in the other subgroups. We also assess the overlap of the generalized propensity scores across subgroups.

(3) We use different propensity score methods for estimating the AMTE and AMTS. We elaborate on potential methods below.

(3a) We use the inverse probability of treatment weighting (IPTW) to estimate the AMTE (e.g., Dong, 2015). The weights are $w_i(s) = \frac{1}{\widehat{\pi_i(s)}}$, where $\widehat{\pi_i(s)}$ is the estimated generalized propensity score of being in the actual/observed subgroup, s. Note that if an individual has a propensity score close to 0 or 1 when the treatment variable is binary, the resulting IPTW-ATE weight can be very large. Further, the resulting IPTW-ATE estimator has a large variance and is not approximately normally distributed (Robins et al., 2000). To overcome this limitation, Robins et al. (2000) proposed the stabilized IPTW-ATE weighting for the binary treatment variable by taking the proportion of individuals in the treated group into account of the weight. Although the stabilized IPTW-ATE weighting approach has demonstrated appropriate estimation of the variance of main effect and appropriate type I error rates (Xu et al., 2010), it should be used with caution, e.g., researchers should conduct appropriate covariate balance diagnosis (see Austin & Stuart, 2015 for a detailed review).

Below we extend the stabilized IPTW-ATE weighting for the binary treatment variable to the stabilized IPTW-AMTE weighting. Recall that the key to applying propensity score methods is to make the distribution of the features of the sample in the comparison groups resemble the distribution of the features of the sample of interest for inference group (e.g., Lenis et al., 2019; Ridgeway et al., 2015). For the AMTE estimation, we are interested in the entire sample for making inferences. Hence, we need to make the feature distribution of the sample in each of our groups resemble the feature distribution of the entire sample. That is, we want to find the weights $w_{AMTE}(\boldsymbol{X}|S=s)$ for individuals in the actual/observed Subgroup s, where $s=1$, 2, 3, or 4, such that

$$f(\boldsymbol{X}) = w_{AMTE}(\boldsymbol{X}|S=s)f(\boldsymbol{X}|S=s), \tag{15}$$

where $f(\boldsymbol{X})$ is the marginal density of the covariates ($\boldsymbol{X}$) for the entire sample, and $f(\boldsymbol{X}|S=s)$ is the marginal density of the covariates for Subgroup s, and $s=1$, 2, 3, or 4.

Rearranging and applying Bayes Theorem we find

$$\begin{aligned} w_{AMTE}(\boldsymbol{X}|S=s) &= \frac{f(\boldsymbol{X})}{f(\boldsymbol{X}|S=s)} \\ &= \frac{f(\boldsymbol{X})}{f(\boldsymbol{X})f(S=s|\boldsymbol{X})/f(S=s)} \\ &= \frac{f(S=s)}{f(S=s|\boldsymbol{X})} = \frac{n_s}{N}\left(\frac{1}{f(S=s|\boldsymbol{X})}\right), \end{aligned} \tag{16}$$

where $N = n_1 + n_2 + n_3 + n_4$ and $f(S=s) = \frac{n_s}{N}$ is the proportion of the sample size for Group s in the total sample. Note that $f(S=s|\boldsymbol{X})$ is the generalized propensity score ($\pi(s)$) for individuals in Group s. Hence, we can use the weight below to estimate the AMTE:

$$w_{AMTE}(\boldsymbol{X}|S=s) = \frac{n_s}{N}\left(\frac{1}{\widehat{\pi_i(s)}}\right). \tag{17}$$

For instance, the AMTE weight for the students who did not go to preschool and did not speak English at home ($s=1$) is $w_{AMTE}(\boldsymbol{X}|S=1) = \frac{n_1}{N}\left(\frac{1}{\widehat{\pi_i(1)}}\right)$, where $\widehat{\pi_i(1)}$ is the estimated generalized propensity score for students being in subgroup 1. This weight can ensure that the feature distribution of the students who did not go to preschool and did not speak English at home resembles the feature distribution of the entire sample (Expression 15).

When there are only two groups ($s=1$ or 2), this stabilized IPTW-AMTE weighting is the same as the stabilized IPTW-ATE weighting for the binary treatment variable (Robins et al., 2000).

(3b) For AMTS, the direct estimate (e.g., using a regression model) is impossible because there are four potential AMTS of interest, each AMTS of interest needs four equivalent treatment-by-moderator subsamples, and it is impossible to simultaneously have four equivalent treatment-by-moderator subsamples for each of four potential AMTS of interest using the original sample without any adjustment (except with random assignment of both treatment and moderator variables where AMTS = AMTE). However, we can use the odds ratio of the generalized propensity scores as the weight, $w_i(s) = \frac{\widehat{\pi_i(s_0)}}{\widehat{\pi_i(s)}}$, to estimate AMTS. The denominator of this expression (odds ratio) is the propensity score of being in the actual subgroup (s) and the numerator is the propensity score of being in the targeted inference subgroup (s_0). For example, if the targeted sample of interest for inference is the treated moderator comparison group ($Z=1$ and $R=1$, i.e., $S=4$), the weight, $w_i(s) = \frac{\widehat{\pi_i(4)}}{\widehat{\pi_i(s)}}$, where $s=1$, 2, 3, or 4.

Similar to stabilized IPTW-AMTE weighting, alternatively we can get the stabilized IPTW-AMTS weighting. For the AMTS estimation, we are interested in a targeted sample ($S=s_0$; e.g., the students who went to preschool and did not speak English at home, i.e., $s_0 = 3$) for making inference to the population that it represents. Hence, we need to make the feature distribution of the sample in the other three groups ($s=1$ for the students who did not go to preschool and did not speak English at home, 2 for the students who did not go to preschool and spoke English at home, and 4 for the students who went to preschool and spoke English at home) resemble the feature distribution of the targeted inference sample. That is, we want to find the weights $w_{AMTS}(\boldsymbol{X}|S=s)$ such that

$$f(\boldsymbol{X}S=s_0) = w_{AMTS}(\boldsymbol{X}|S=s)f(\boldsymbol{X}|S=s). \tag{18}$$

Rearranging and applying Bayes Theorem we find

$$\begin{aligned} w_{AMTS}(\boldsymbol{X}|S=s) &= \frac{f(\boldsymbol{X}|S=s_0)}{f(\boldsymbol{X}|S=s)} \\ &= \frac{f(\boldsymbol{X})f(S=s_0|\boldsymbol{X})/f(S=s_0)}{f(\boldsymbol{X})f(S=s|\boldsymbol{X})/f(S=s)} \\ &= \frac{f(S=s)}{f(S=s_0)}\left(\frac{f(S=s_0|\boldsymbol{X})}{f(S=s|\boldsymbol{X})}\right) \\ &= \frac{n_s}{n_{s_0}}\left(\frac{f(S=s_0|\boldsymbol{X})}{f(S=s|\boldsymbol{X})}\right), \end{aligned} \tag{19}$$

where $\frac{f(S=s)}{f(S=s_0)} = \frac{n_s}{n_{s_0}}$ is the ratio of the sample size for Group s to the sample size for the targeted inference group. Note that $\frac{f(S=s_0|\boldsymbol{X})}{f(S=s|\boldsymbol{X})}$ is the odds ratio of the generalized propensity scores in the targeted inference group ($S = s_0$) to the actual/observed Group s. Hence, we can use the weight below to estimate the AMTS:

$$w_{AMTS}(\boldsymbol{X}|S=s) = \frac{n_s}{n_{s_0}}\left(\frac{\widehat{\pi_i(s_0)}}{\widehat{\pi_i(s)}}\right). \quad (20)$$

For the individuals in the inference group ($s = s_0$), $w_{AMTS}(\boldsymbol{X}|S=s_0) = 1$.

We then check the overlap of the generalized propensity scores and covariate balance based on the weights that we derived (e.g., Austin, 2008; Ridgeway et al., 2021; Rosenbaum, 2002). The means of covariates for four treatment-by-moderator subgroups (S) are estimated using the AMTE and AMTS weights, and without weights. The maximum standardized mean differences (MSMD) among four subgroups were calculated for the AMTE:

$$MSMD_{AMTE} = [Max\left(\overline{X}|S=1, \overline{X}|S=2, \overline{X}|S=3, \overline{X}|S=4\right) - Min\left(\overline{X}|S=1, \overline{X}|S=2, \overline{X}|S=3, \overline{X}|S=4\right)]/SD_x. \quad (21)$$

Similarly, the MSMD between the targeted inference subgroup and the other three subgroups were calculated for the AMTS:

$$MSMD_{AMTS} = Max\left(|(\overline{X}|S=s_1 - \overline{X}|S=s_0)|, |(\overline{X}|S=s_2 - \overline{X}|S=s_0)|, |(X|S=s_4 - \overline{X}|S=s_0)|\right)/SD_x, \quad (22)$$

where $\overline{X}|S$ is the sample mean of covariate X for subgroup S, SD_x is the pooled standard deviation among four subgroups for the unweighted sample, $|.|$ is the operation for absolute values, s_0 is the targeted inference subgroup ($s_0 = 3$ in this example), and s_1, s_2, and s_4 are the other subgroups. The MSMD with and without weights for the entire sample and targeted inference subsamples will be compared and plotted in figures.

Finally, we can estimate the AMTE and AMTS based on respective weights while controlling for covariates in the statistical models to further reduce selection bias and improve precision (refereed as "for double robustness", e.g., Austin, 2017; Kang & Schafer, 2007; Tsiatis & Davidian, 2007). We can also estimate the $AMTS_{(Z=1)}$, $AMTS_{(Z=0)}$, $AMTS_{(R=1)}$, and $AMTS_{(R=0)}$ based on Expressions in Table 2.

In addition to weighting, we can use propensity score matching (e.g., greedy matching, optimal matching) to estimate AMTS. First, we can estimate the generalized propensity score of being in the targeted inference subgroup, s_0. Then we match the sample from the other subgroups with Subgroup s_0 based on the generalized propensity score of being in Subgroup s_0. After balance checking we can estimate AMTS using the combined sample. The limitation of this matching approach is that we may not have well matched units as finding well matched units is more likely when the number of comparison units is much larger than the targeted sample. Thus the propensity score matching approach may only work well for the targeted inference subgroup with the smallest sample size among all four subgroups.

Illustration: The differential (moderated) effect of preschool

Data

The data were from the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), a nationally representative longitudinal study of children (U.S. Department of Education & National Center for Education Statistics, 2009). A total of 22,666 children attending kindergarten during the 1998–99 school year were sampled. The academic achievement measures on math and reading were administered in the fall of Kindergarten through the spring of Grade 8, and additional extensive data regarding child and family characteristics was collected at kindergarten entry.

Following Magnuson et al. (2007) we defined the treatment and comparison conditions using the parental response to the fall kindergarten survey question "primary type non-parental care at prekindergarten" (variable P1PRIMPK) (U.S. Department of Education & National Center for Education Statistics, 2009). The analytic sample includes two groups of interest: center-based preschool treatment (N = 7,367) and parental care comparison (N = 3,150). We coded the treatment variable, Preschool = 1 for children in center-based preschool, and Preschool = 0 for children in parental care.

The outcome variable is the Item Response Theory (IRT) scale score of children's math achievement in the fall of Kindergarten. The outcome measure has high reliability, with a Cronbach's alpha coefficient of 0.88 (Tourangeau et al., 2009). We standardized the outcome to a z-score to facilitate interpretation. The moderator variable is English speaking status at home: Speaking English at home (EnglishHome = 1, N = 9,239) and not speaking English at home (EnglishHome = 0, N = 1,278). For the covariates to estimate the generalized propensity scores, we

Table 3. Descriptive statistics of the sample and covariate balance among four treatment-by-moderator subgroups.

Treatment-by-Moderator (S)	1		2		3		4		
Preschool (Z)	0		0		1		1		
Speaking English at home (R)	0		1		0		1		
Variable	Mean	SD	Mean	SD	Mean	SD	Mean	SD	MSMD
Binary									
Black	0.01	0.11	0.12	0.33	0.02	0.15	0.12	0.33	0.36
Hispanic	0.82	0.38	0.13	0.33	0.60	0.49	0.08	0.27	2.38
Rural	0.03	0.18	0.26	0.44	0.03	0.17	0.16	0.37	0.63
One parent with siblings	0.13	0.33	0.13	0.34	0.09	0.29	0.11	0.31	0.13
Biological mother	0.98	0.15	0.95	0.22	0.97	0.17	0.95	0.22	0.13
Continuous									
Weight (pounds)	46.92	9.71	45.69	8.62	47.17	9.55	46.25	8.14	0.18
Age (month)	64.55	4.61	65.62	4.36	64.72	4.05	65.79	4.23	0.29
Family income ($ thousand)	27.36	33.49	43.77	41.02	50.33	54.56	68.41	64.68	0.71
Parent highest education	3.17	2.04	4.37	1.78	4.89	2.38	5.45	1.82	1.23
SES	−0.63	0.68	−0.14	0.73	0.05	0.92	0.34	0.74	1.31
Sample size	654		2,496		624		6,743		

Note: Treatment-by-Moderator (S) corresponds to the four combinations of Preschool (Z) and Speaking English at home (R). The maximum standardized mean difference (MSMD) on covariate X among four subgroups is $\left[Max\left(\bar{X}|\boldsymbol{S}=1,\bar{\boldsymbol{X}}|S=2,\bar{X}|\boldsymbol{S}=3,\bar{\boldsymbol{X}}|S=4\right)-Min\left(\bar{X}|\boldsymbol{S}=1,\bar{\boldsymbol{X}}|S=2,\bar{X}|\boldsymbol{S}=3,\bar{\boldsymbol{X}}|S=4\right)\right]/SD_x$, where $\bar{X}|S$ is the mean for subgroup S, SD_x is the pooled standard deviation among four subgroups.

considered the covariate list that Magnuson et al. (2007) used, and we chose the covariates that were correlated with the outcome, the treatment status, and the moderator (Steiner et al., 2010). These covariates included race, weight, age at the kindergarten entry, parents' educational level, income, composite SES measure, household structure (numbers of parents and siblings), and locality (rural or urban). Table 3 presents the descriptive statistics of the covariates by the treatment-by-moderator groups.

We conducted the initial covariate balance checking before the moderation analysis. Only three out of ten covariates were balanced, that is, the maximum standardized mean difference for three covariates among four treatment-by-moderator groups was smaller than 0.25 (Table 3). Multiple covariates demonstrated extremely large imbalances across the four treatment-by- moderator subgroups, e.g., Hispanic, parent highest education, and SES yielded standardized mean differences of 2.38, 1.23, and 1.31, respectively. Such covariate imbalance across treatment by moderator subgroups suggests a violation of assumption 2 such that the treatment by moderator interaction is not independent of these covariates.

Procedures for estimating AMTE and AMTS

The procedure unfolds as follows (The annotated SAS code and dataset are in the supplemental material package).

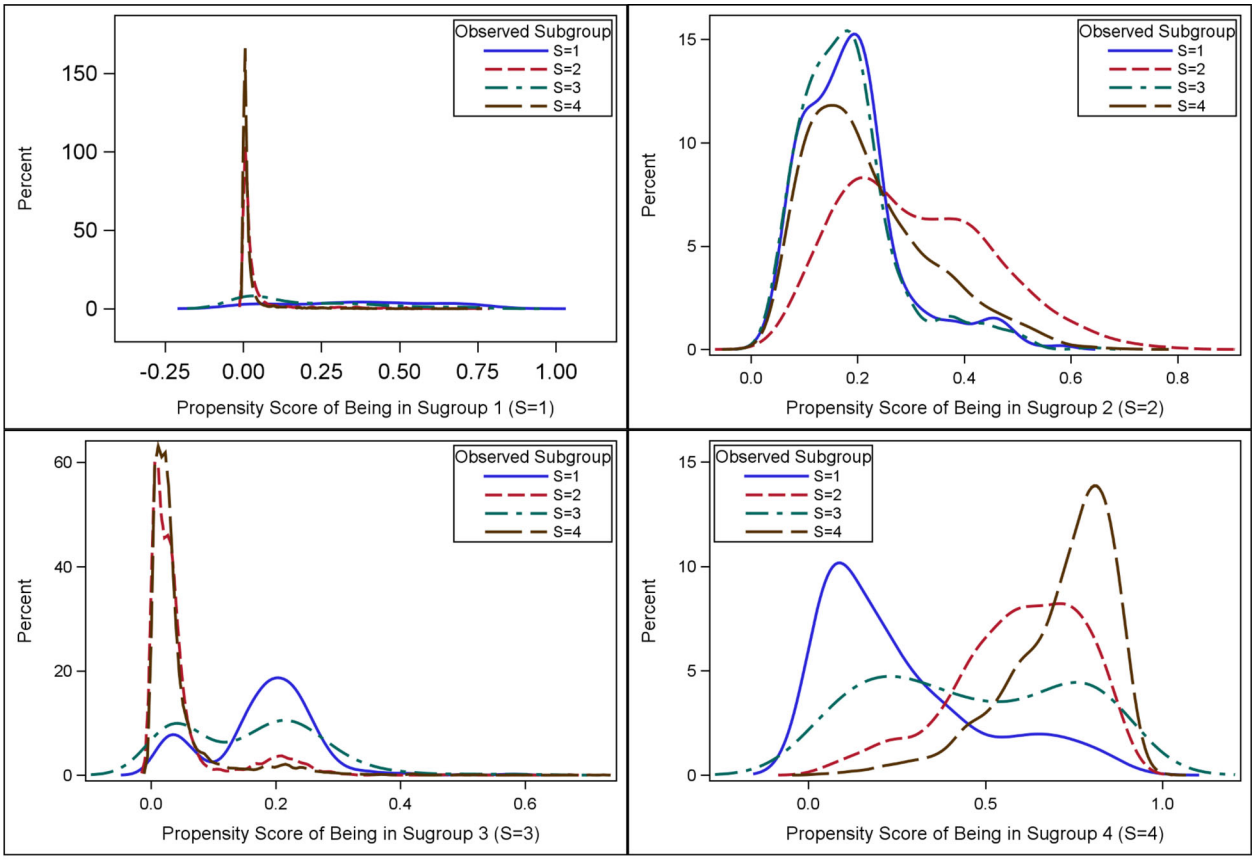
(1) We first created a new variable (S) indicating four treatment-by-moderator subgroups ($S=1$ if $Preschool=0$ and $EnglishHome=0$, $S=2$ if $Preschool=0$ and $EnglishHome=1$, $S=3$ if $Preschool=1$ and $EnglishHome=0$, and $S=4$ if $Preschool=1$ and $EnglishHome=1$).

(2) We estimated a multinomial logistic regression model to predict the generalized propensity scores for individual i of being in certain subgroup: $\pi_i(s)=Pr(S_i=s|\boldsymbol{X}_i)$, where $s=1$, 2, 3, or 4. We used an iterative process to estimate the generalizing propensity scores by assessing covariate balance and revising the model to include polynomials and interactions to explore nonlinear functional forms for achieving the best covariate balance. The covariates ($\boldsymbol{X}$) in the final model included those listed in Table 3, the interaction term of one parent with siblings and parent highest education, and several high order terms (quadratic and cubic terms of family income and SES). We checked the overlap of generalized propensity scores among subgroups. Figure 2 presents the kernel density of the generalized propensity scores among the four subgroups. There is some overlap on the generalized propensity scores among the four subgroups, but the distribution is not the same.

(3) We calculated various weights based on the generalized propensity scores:

The IPTW-AMTE weight, $w_i(s)=\frac{1}{\widehat{\pi_i(s)}}$, and the stabilized IPTW-AMTE weight, $w_{AMTE}(\boldsymbol{X}|S=s)=\frac{n_s}{N}\left(\frac{1}{\widehat{\pi_i(s)}}\right)$, where $\widehat{\pi_i(s)}$ is the estimated generalized propensity score of being in the actual subgroup, s.

The odds ratio of the generalized propensity scores serves as weight for AMTS, $w_i(s)=\frac{\widehat{\pi_i(s_0)}}{\widehat{\pi_i(s)}}$, and the stabilized AMTS weight, $w_{AMTS}(\boldsymbol{X}|S=s)=\frac{n_s}{n_{s_0}}\left(\frac{\widehat{\pi_i(s_0)}}{\widehat{\pi_i(s)}}\right)$,



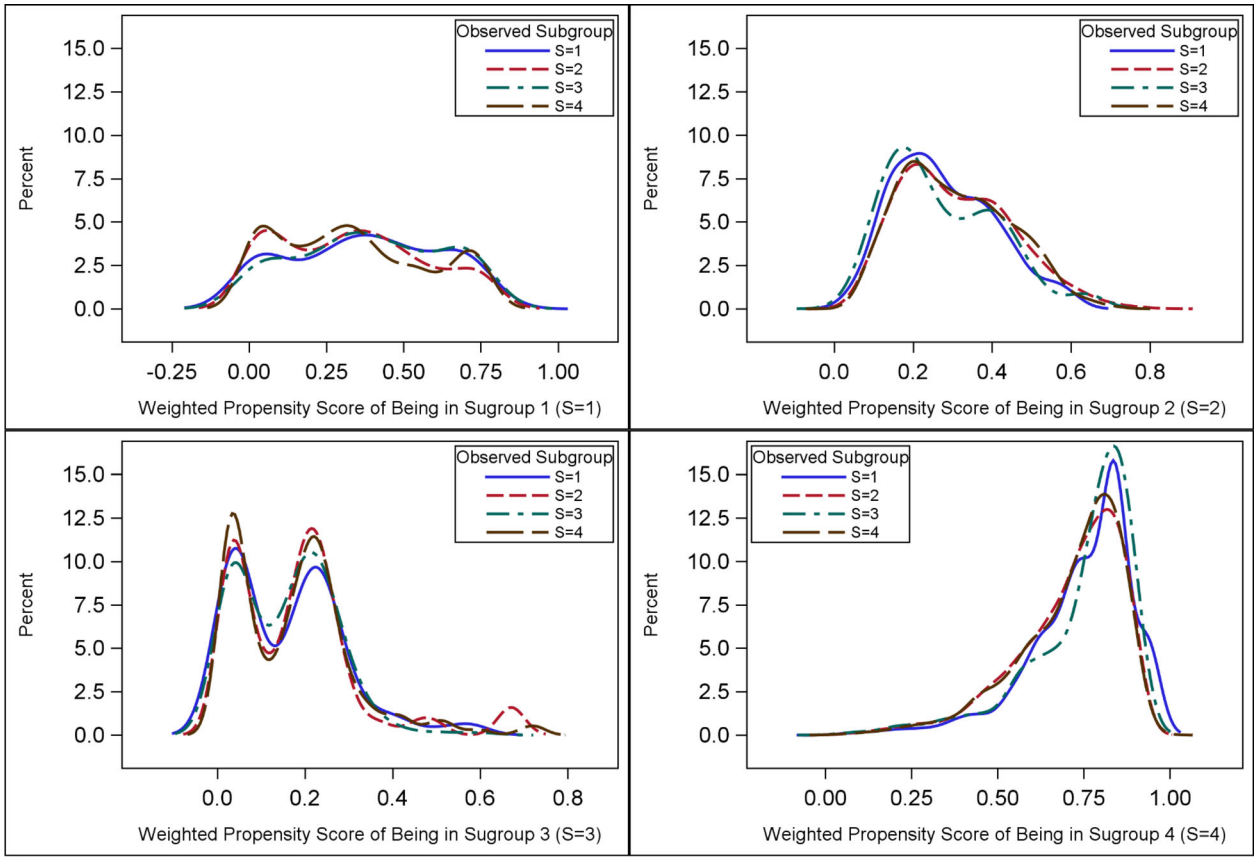
Note: S = 1 if Preschool = 0 and Speaking English at home = 0, S = 2 if Preschool = 0 and Speaking English at home = 1, S = 3 if Preschool = 1 and Speaking English at home = 0, and S = 4 if Preschool = 1 and Speaking English at home = 1.

Figure 2. Kernel density of the generalized propensity scores among four treatment-by-moderator subgroups before weighting.

where $S = s_0$ indicates the targeted inference subgroup. For example, if the actual/observed subgroup of an individual is $S = 2$, then her stabilized AMTS weights are: $w_{AMTS}(\boldsymbol{X}|S=1) = \frac{n_2}{n_1}\left(\frac{\widehat{\pi_i(1)}}{\widehat{\pi_i(2)}}\right)$, for resembling targeted inference Subgroup 1; $w_{AMTS}(\boldsymbol{X}|S=2) = \frac{n_2}{n_2}\left(\frac{\widehat{\pi_i(2)}}{\widehat{\pi_i(2)}}\right) = 1$, for being in targeted inference Subgroup 2; $w_{AMTS}(\boldsymbol{X}|S=3) = \frac{n_2}{n_3}\left(\frac{\widehat{\pi_i(3)}}{\widehat{\pi_i(2)}}\right)$, for resembling targeted inference Subgroup 3; $w_{AMTS}(\boldsymbol{X}|S=4) = \frac{n_2}{n_4}\left(\frac{\widehat{\pi_i(4)}}{\widehat{\pi_i(2)}}\right)$, for resembling targeted inference Subgroup 4.

(4) We assessed the overlap of the generalized propensity scores and covariate balance. The kernel density of the generalized propensity scores among the four subgroups after weighting by AMTS in Figure 3 indicates much better overlap than without weighting (Figure 2). The maximum standardized mean differences (MSMD) with and without weights for the entire sample and targeted inference subsamples were plotted in Figures 4 and 5, respectively. All the covariates were much more balanced when weighted by the AMTE and AMTS weights (dots) than without weighting (circles). For instance, the MSMD for all covariates were below 0.25 for AMTS (S = 1, 2, and 3), only one covariate was above 0.25 for AMTE (0.31 for one parent with siblings), and two covariates were above 0.25 for AMTS (S = 4) (0.31 for Black and 0.41 for one parent with sibling).

(5) We estimated the AMTE and AMTS for the four treatment-by-moderator subgroups using the general linear model including the respective weights and controlling for covariates for double robustness. The statistical model is below:



Note: S = 1 if Preschool = 0 and Speaking English at home = 0, S = 2 if Preschool = 0 and Speaking English at home = 1, S = 3 if Preschool = 1 and Speaking English at home = 0, and S = 4 if Preschool = 1 and Speaking English at home = 1.

Figure 3. Kernel density of the generalized propensity scores among four treatment-by-moderator subgroups after weighting.

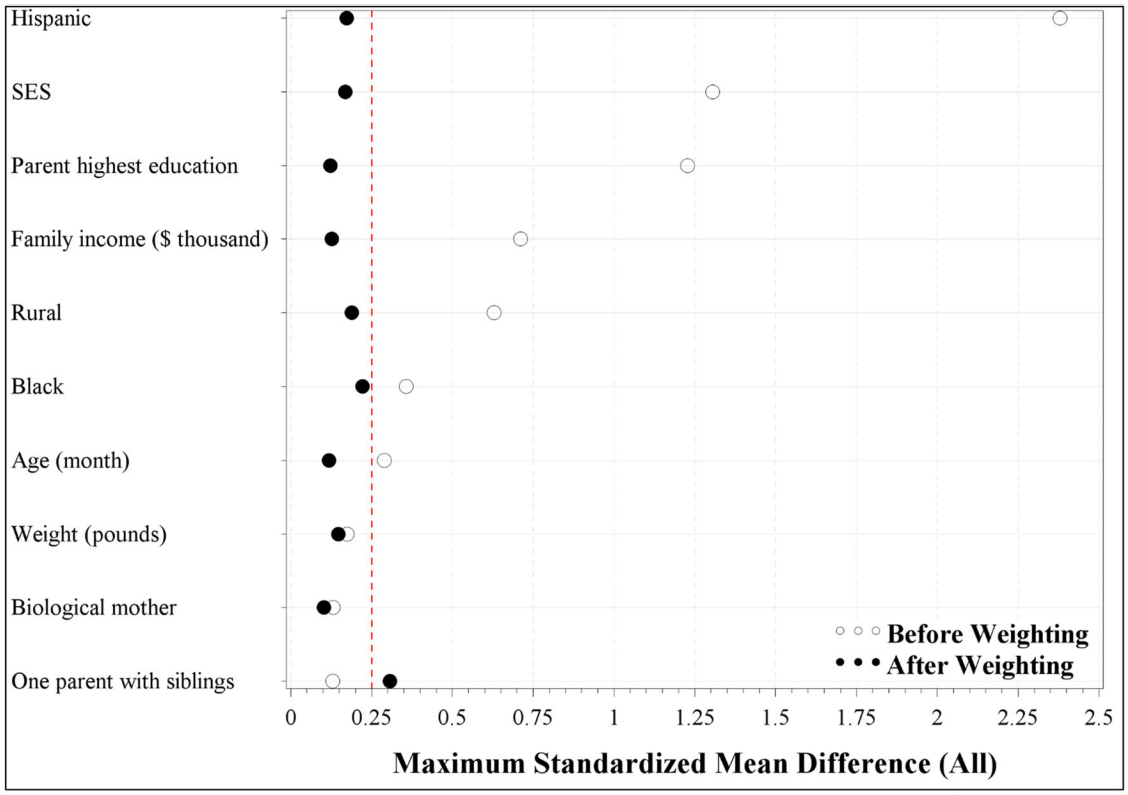
$$Y_i = \beta_0 + \beta_1(Preschool)_i + \beta_2(EnglishHome)_i + \beta_3(Preschool)_i*(EnglishHome)_i + \sum_{q=4}^{Q} \beta_q X_{qi} + e_i, e_i \sim N(0, \sigma^2). \quad (23)$$

Y_i represents the z-score of the math achievement for student i in the fall of Kindergarten. $(Preschool)_i$ represents the student's preschool experience ($Preschool = 1$ for being in the preschool, and 0 in parental care). $(EnglishHome)_i$ represents the student's English speaking status at home ($EnglishHome = 1$ for speaking English at home, and $EnglishHome = 0$ for not speaking English at home). X_{qi} represents the other covariates listed in Table 1, which include Black, Hispanic, rural, one parent with siblings, biological mother, weight (pounds), age (month), family income ($thousand), parent highest education, and SES. The covariates are included for further reduction of bias (double robustness) and improved precision. The parameter, β_1, is the average effect of preschool on the math achievement in the fall of kindergarten for the students who did not speak English at home. The parameter, β_3, is the moderator (additional) effects of preschool on the math achievement for the students who spoke English at home compared with the students who did not speak English at home. The average effect of preschool for the student who spoke English at home can be calculated using $(\beta_1 + \beta_3)$. Because the outcome measure was a z-score, the parameters, β_1 and β_3 are the standardized regression coefficients and indicate the effect sizes in the unit of a standard deviation. We estimated the $AMTS_{(Z=1)}$, $AMTS_{(Z=0)}$, $AMTS_{(R=1)}$, and $AMTS_{(R=0)}$ based on Expressions in Table 2.

For comparison purposes, in addition to the weighted analysis with controlling for covariates for double robustness (AMTE), we conducted the conventional moderation analysis without weighting without controlling for covariates (conventional w/o covariates), with controlling for covariates (conventional), and the weighted analysis of the entire sample without controlling for covariates (AMTE w/o covariates).

Results

The detailed results of the analyses (conventional w/o covariates, conventional, AMTE w/o covariates,

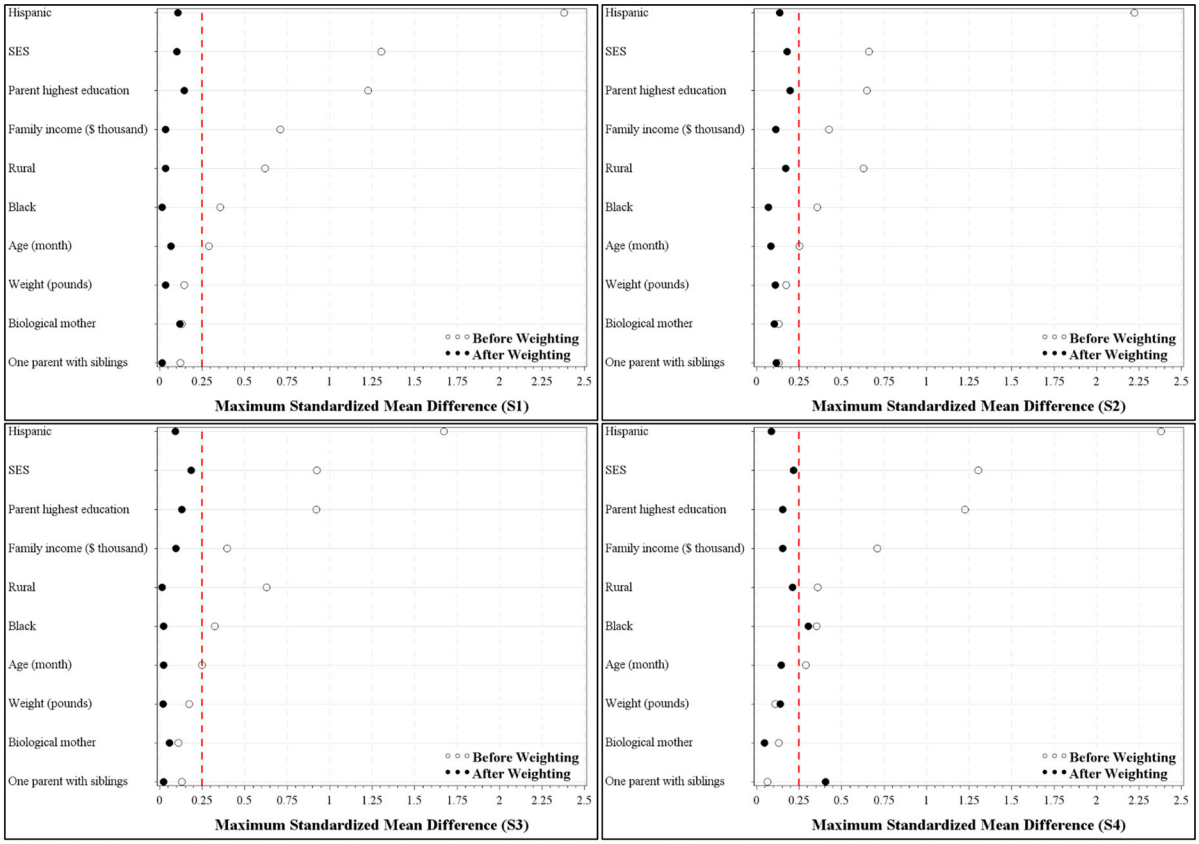


Note: Variables are sorted from highest to lowest maximum standardized mean difference (MSMD) prior to weighting. Dots to the right of the vertical red dashed line indicates variables with MSMD > 0.25 imbalance among four subgroups.

Figure 4. Covariate balance checking before and after propensity score weighting for all sample.

AMTE, AMTS) are presented in Table 4. The bolding represent the parameters of interest. Both β_1 and β_3 are the standardized regression coefficients indicating the effect sizes in the unit of a standard deviation as in Expression 23. β_1 indicates the average effect of preschool on the math achievement for students who did not speak English at home; β_3, indicates the moderator (additional) effects of preschool on the math achievement for the students who spoke English at home compared with the students who did not speak English at home. The average effect of preschool for the student who spoke English at home can be calculated using $(\beta_1 + \beta_3)$. Figure 6 presents the moderator effect sizes and 95% confidence intervals from the different analyses. Figure 7 presents the effect sizes and 95% confidence intervals of the effects of preschool by moderator subgroups (speaking and not speaking English at home).

The findings are summarized below. First, regarding the average treatment effects on the subgroups (speaking English at home or not), there were statistically significantly positive effects for preschool compared with parental care in all analyses (Figure 7). This suggests that preschool is more effective than parental care in improving students' math achievement regardless of home language background. Specifically, the analysis without controlling for covariates tend to produce larger effect sizes, and this is more obvious for the conventional moderation analysis. The AMTE analysis with double robustness produced slightly smaller estimates than the AMTE analysis without controlling for covariates. In addition, the AMTE analysis with double robustness produced larger but non-significantly different estimates of the effects of preschool on the students not speaking English at home ($d = 0.24$, $p < 0.001$) than the conventional moderation analysis with controlling for covariates ($d = 0.18$, $p < 0.001$), and there was no difference between these two analyses on the students speaking English at home. In addition, the AMTS analysis produced similar estimates of the effect of preschool on the subgroups as the AMTE except for the targeted subgroup $S = 1$ (the students who received parental care and did not speak English at home). For the students in this subgroup, the effect size of preschool was 0.05 ($p = 0.002$) if they had attended preschool and spoke English at home, and the effect size of preschool was 0.15 ($p < 0.001$) if they had attended preschool but did not speak English at home. Both effect sizes for this subgroup were smaller than the other subgroups and the entire sample.



Note: Variables are sorted from highest to lowest maximum standardized mean difference (MSMD) prior to weighting. Dots to the right of the vertical red dashed line indicates variables with MSMD > 0.25 imbalance among four subgroups. S = 1 if Preschool = 0 and Speaking English at home = 0, S = 2 if Preschool = 0 and Speaking English at home = 1, S = 3 if Preschool = 1 and Speaking English at home = 0, and S = 4 if Preschool = 1 and Speaking English at home = 1.

Figure 5. Covariate balance checking before and after propensity score weighting for targeted subsample.

Second, the moderator effect in the analysis of the entire sample was non-significant for both the conventional moderation analysis with controlling for covariates ($d = 0.01$, $p = 0.881$) and the AMTE estimate ($d = -0.06$, $p = 0.274$), although the students who did not speak English at home ($d = 0.24$, $p < 0.001$) benefited more from preschool than their peers who spoke English at home ($d = 0.18$, $p < 0.001$) in the AMTE estimate. In addition, for the analysis of targeted subgroups, none of AMTS (S = 2, 3, and 4) estimates produced a significant moderation effect size difference; however, the AMTS for S = 1 (the students who received parental care and did not speak English at home) is statistically significant ($d = -0.10$, $p = 0.008$). This suggests that it helped to improve the students' math achievement more if they went to preschool but did not speak English at home than spoke English at home.

Finally, the AMTE based on the weighted average of AMTS (Expression 14) is −0.06, which is the same as the direct estimate. The $AMTS_{(Z=1)}$, $AMTS_{(Z=0)}$, $AMTS_{(R=1)}$, and $AMTS_{(R=0)}$ estimates based on Expressions in Table 2 are −0.06, −0.05, −0.05, and −0.07, respectively, indicating very little difference on the moderator effect estimates among the targeted subgroups solely based on preschool or English speaking status.

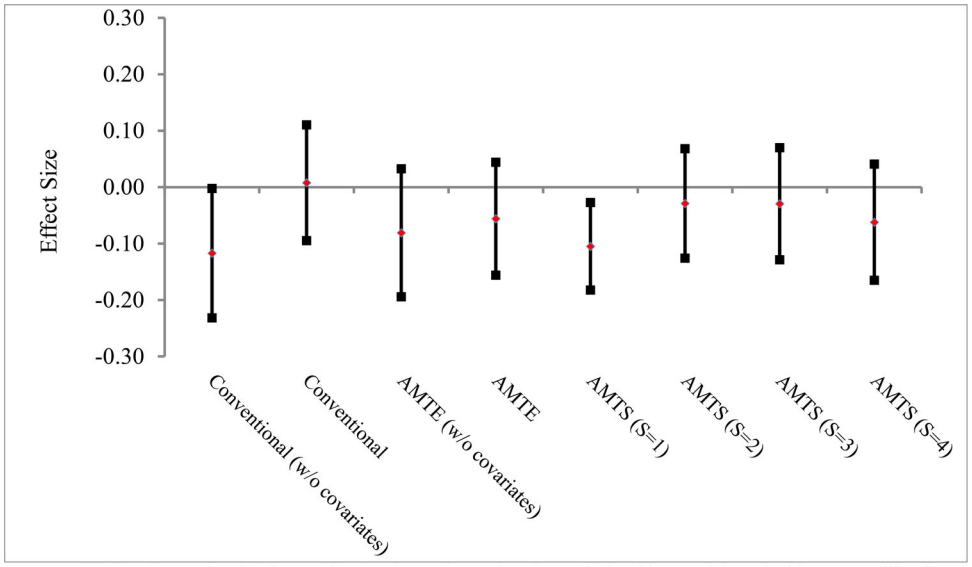
Discussion and conclusion

In this study, we proposed an extended causal moderation analysis framework based on potential outcomes. We defined and proposed two types of estimands (AMTE and AMTS) for making inferences to different populations of interest to estimate the moderator effects and main treatment effects. These estimands provide more options to study policy relevant subgroups, e.g., the children who did not speak English at home with parental care. Furthermore, we used the (stabilized) IPTW-AMTE weight to estimate the AMTE and the (stabilized) AMTS weight (odds ratio of generalized propensity scores rescaled by sample sizes) to estimate the AMTS. We derived these

Table 4. Results of conventional moderation analysis, AMTE, and AMTS.

	Conventional (w/o covariates)			Conventional			AMTE (w/o covariates)			AMTE			AMTS (S = 1)			AMTS (S = 2)			AMTS (S = 3)			AMTS (S = 4)		
Variable	*b*	*SE*	*p*	*b*	*SE*	*p*	*b*	*SE*	*p*	*b*	*SE*	*p*	*b*	*SE*	*p*	*b*	*SE*	*p*	*b*	*SE*	*p*	*b*	*SE*	*p*
Intercept	−0.718	0.038	<0.001	−4.003	0.147	<0.001	−0.091	0.035	0.009	−3.884	0.149	<0.001	−3.359	0.126	<0.001	−3.735	0.138	<0.001	−3.371	0.150	<0.001	−4.001	0.155	<0.001
Black	NA			−0.263	0.029	<0.001	NA			−0.220	0.029	<0.001	−0.233	0.059	<0.001	−0.215	0.026	<0.001	−0.225	0.056	<0.001	−0.218	0.028	<0.001
Hispanic	NA			−0.261	0.028	<0.001	NA			−0.275	0.025	<0.001	−0.294	0.018	<0.001	−0.226	0.025	<0.001	−0.339	0.018	<0.001	−0.250	0.033	<0.001
Rural	NA			−0.121	0.023	<0.001	NA			−0.147	0.023	<0.001	−0.113	0.039	0.003	−0.152	0.019	<0.001	−0.108	0.050	0.032	−0.140	0.025	<0.001
One parent with siblings	NA			−0.115	0.028	<0.001	NA			−0.121	0.028	<0.001	−0.059	0.020	0.004	−0.107	0.025	<0.001	−0.102	0.030	0.001	−0.141	0.029	<0.001
Biological mother	NA			0.166	0.039	<0.001	NA			0.160	0.040	<0.001	0.073	0.043	0.090	0.125	0.036	<0.001	0.145	0.050	0.004	0.180	0.040	<0.001
Weight (pounds)	NA			0.004	0.001	<0.001	NA			0.004	0.001	<0.001	0.004	0.001	<0.001	0.003	0.001	0.002	0.002	0.001	0.006	0.004	0.001	<0.001
Age (month)	NA			0.047	0.002	<0.001	NA			0.046	0.002	<0.001	0.034	0.002	<0.001	0.043	0.002	<0.001	0.038	0.002	<0.001	0.049	0.002	<0.001
Family income ($ thousand)	NA			0.001	<0.001	<0.001	NA			0.001	<0.001	<0.001	0.002	<0.001	<0.001	0.001	<0.001	<0.001	0.001	<0.001	<0.001	0.001	<0.001	<0.001
Parent highest education	NA			0.079	0.009	<0.001	NA			0.092	0.009	<0.001	0.126	0.006	<0.001	0.107	0.007	<0.001	0.090	0.008	<0.001	0.079	0.010	<0.001
SES	NA			0.216	0.023	<0.001	NA			0.192	0.023	<0.001	0.058	0.017	0.001	0.145	0.019	<0.001	0.122	0.021	<0.001	0.236	0.025	<0.001
Preschool (β_1)	**0.551**	**0.054**	**<0.001**	**0.179**	**0.049**	**<0.001**	**0.269**	**0.053**	**<0.001**	**0.241**	**0.047**	**<0.001**	**0.154**	**0.036**	**<0.001**	**0.200**	**0.046**	**<0.001**	**0.221**	**0.047**	**<0.001**	**0.263**	**0.048**	**<0.001**
Speaking English at home (β_2)	0.463	0.042	<0.001	0.084	0.042	0.047	−0.027	0.040	0.501	−0.024	0.036	0.496	0.158	0.029	<0.001	0.013	0.034	0.692	0.109	0.037	0.003	−0.066	0.037	0.072
Preschool * Speaking English at home (β_3)	**-0.117**	**0.059**	**0.046**	**0.008**	**0.052**	**0.881**	**-0.081**	**0.058**	**0.162**	**-0.056**	**0.051**	**0.274**	**-0.105**	**0.040**	**0.008**	**-0.029**	**0.049**	**0.559**	**-0.029**	**0.051**	**0.562**	**-0.062**	**0.053**	**0.237**

Note: $N = 10{,}517$. Conventional (w/o covariates) refers to the moderation analysis without weight and without controlling for covariates for the entire sample. AMTE (w/o covariates) refers to the analysis weighted by the stabilized AMTE weights but without controlling for covariates. All the other analysis controlled for covariates. AMTS (S = s) refers to the analysis weighted by the stabilized AMTS weights, where s is the targeted inference sample of interest. S = 1 if Preschool = 0 and Speaking English at home = 0, S = 2 if Preschool = 0 and Speaking English at home = 1, S = 3 if Preschool = 1 and Speaking English at home = 0, and S = 4 if Preschool = 1 and Speaking English at home = 1. The bolding represent the parameters of interest. Both β_1 and β_3 are the standardized regression coefficients as in Expression 21. β_1 indicates the average effect of preschool on the math achievement for students who did not speak English at home; β_3, indicates the moderator (additional) effects of preschool on the math achievement for the students who spoke English at home compared with the students who did not speak English at home.



Note: $N = 10{,}517$. Conventional (w/o covariates) refers to the moderation analysis without weight and without controlling for covariates for the entire sample. AMTE (w/o covariates) refers to the analysis weighted by the stabilized AMTE weights but without controlling for covariates. All the other analysis controlled for covariates. AMTE refers to the analysis weighted by the stabilized AMTE weights; AMTS (S=s) refers to the analysis weighted by the stabilized AMTS weights, where is the targeted inference sample of interest S = 1, 2, 3, and 4, respectively. Unweighted (S=s) refers to the analysis without weights, where all the other subgroups are compared with the targeted inference sample of interest S. S = 1 if Preschool = 0 and Speaking English at home = 0, S = 2 if Preschool = 0 and Speaking English at home = 1, S = 3 if Preschool = 1 and Speaking English at home = 0, and S = 4 if Preschool = 1 and Speaking English at home = 1.

Figure 6. Moderator effect sizes (preschool effect size differences between students speaking and not speaking English at home) and 95% confidence intervals by different analysis.

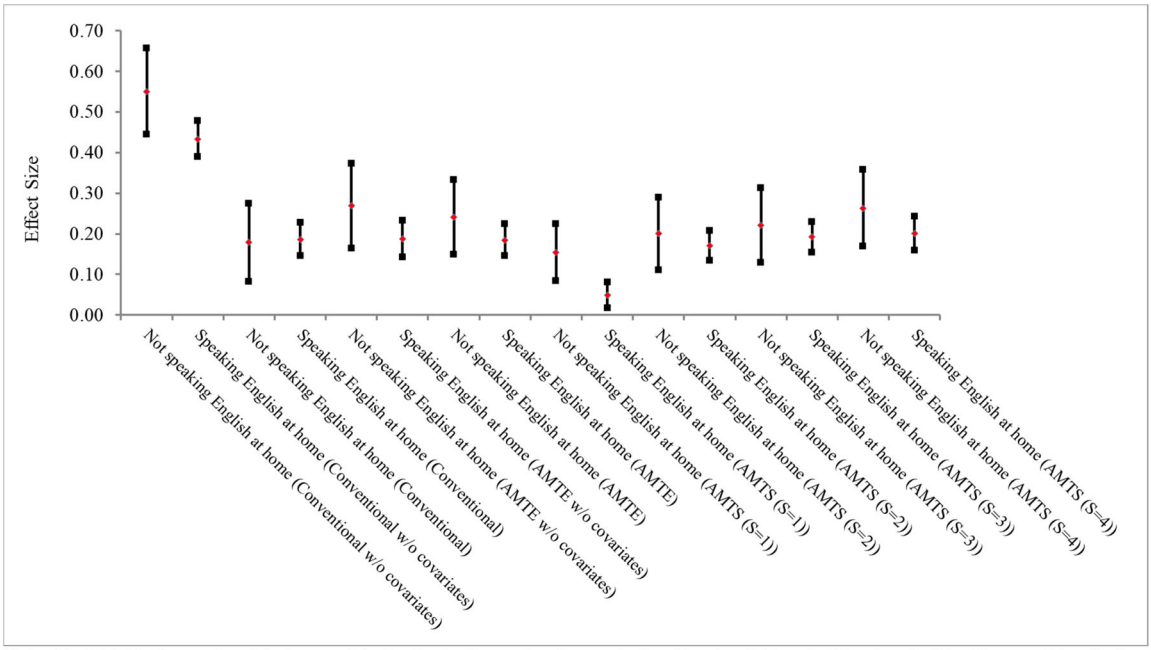
weights aiming to make the feature distribution of the sample in other subgroups resemble the feature distribution of the inference sample of interest. This weighting approach makes it feasible to make causal inferences for moderator effects to targeted populations.

We demonstrated the application of the new causal moderation analysis framework through the preschool example. Several key findings emerged. First, the covariates are much more balanced using the AMTE and AMTS weights than without weights. This suggests our proposed method can reduce selection bias due to nonrandom assignment of the treatment (preschool) and the moderator (home language status). In addition, the weighting approach can balance all the other potential moderators that are included in the propensity score model, hence, the AMTE and AMETS estimates of the moderator of interest are still valid even if there are other moderators.

Second, the non-significant AMTE estimate on the entire sample suggests that the home language status is not a moderator, which is consistent with the findings from the conventional moderation analysis. In our example, although the AMTE estimation does not draw a different conclusion from the conventional moderation analysis, it provides evidence of good covariate balance. It suggests that the conventional moderation analysis based on the regression model that controls for covariates may sometimes work well to reduce selection bias (e.g., in this case), but the AMTE estimation can reduce selection bias through balancing covariates in general.

Third, the AMTS estimates demonstrate some variation among four targeted subgroups. The AMTS estimates for $S = 2$, 3, and 4 were non-significant; however, the AMTS estimate for $S = 1$ (not speaking English at home with parental care) was statistically significant ($d = -0.10$, $p = 0.008$). This finding suggests that for the students who had a similar background to this subgroup ($S = 1$), i.e., most were Hispanic students (82%) with lower family income, lower parent education, and lower SES (Table 3), the preschool was more effective if their status was not speaking English at home ($d = 0.15$, $p < 0.001$) than if it were speaking English at home ($d = 0.05$, $p = 0.002$). This finding is consistent with Lipsey et al. (2013). This provides additional evidence of the effects of preschool in improving the math achievement for students with low SES and not speaking English at home (e.g., Hispanic). It implies that preschool may be a policy tool to improve the academic achievement for economically disadvantaged immigrant children.

Fourth, the AMTS estimates can help with the investigation of treatment effect heterogeneity. For instance, the largest effect ($d = 0.26$, $p < 0.001$, Table 4) of preschool compared to parental care was for the students



Note: $N = 10{,}517$. Conventional (w/o covariates) refers to the moderation analysis without weight and without controlling for covariates for the entire sample. AMTE (w/o covariates) refers to the analysis weighted by the stabilized AMTE weights but without controlling for covariates. All the other analysis controlled for covariates. AMTE refers to the analysis weighted by the stabilized AMTE weights; AMTS (S=s) refers to the analysis weighted by the stabilized AMTS weights, where is the targeted inference sample of interest S = 1, 2, 3, and 4, respectively. Unweighted (S=s) refers to the analysis without weights, where all the other subgroups are compared with the targeted inference sample of interest S. S = 1 if Preschool = 0 and Speaking English at home = 0, S = 2 if Preschool = 0 and Speaking English at home = 1, S = 3 if Preschool = 1 and Speaking English at home = 0, and S = 4 if Preschool = 1 and Speaking English at home = 1.

Figure 7. Preschool effect sizes and 95% confidence intervals by moderator subgroups.

who attended preschool and spoke English at home (S = 4) should they not speak English at home, i.e., among four subgroups the preschool had the largest effect for the students with the same characteristics as S = 4 if they did not speak English at home. In contrast, the smallest effect of preschool was for the students with the same characteristics as S = 1 (did not speak English at home with parental care): the effect was $d = 0.15$ ($p < 0.001$, Table 4) should they attend preschool and not speak English at home, and the effect was $d = 0.05$ ($p = 0.002$, Table 4) should they attend preschool and speak English at home. The treatment effect heterogeneity may be due to the sample difference, e.g., proportion of Hispanic, family income, parent education, and SES (Table 3), and suggest existence of the other moderators. Applying the similar analysis to other potential moderators may help identify the source of treatment effect heterogeneity.

Finally, the AMTS estimate has an advantage in that it reduces assumptions compared to the AMTE estimate. For example, if our targeted subgroup is S = 1 (Z = 0, R = 0), the ignorability of the treatment and moderator assumption only requires that the potential outcome is independent of the treatment and moderator variables given observed covariates for all other three subgroups (S = 2, 3, and 4) whereas the AMTE requires for all four subgroups. Hence, it is more likely to produce unbiased AMTS estimates than the AMTE estimate.

Limitations

As in all propensity score analyses, the veracity of causal inferences are potentially susceptible to hidden bias due to unmeasured variables. When either the treatment or the hypothesized moderator variable is not randomly assigned, the interaction term of the treatment and the hypothesized moderator variables cannot be assured to be independent of other measured and unmeasured variables. Like all other applications, our proposed approach is limited in that it only balances the measured variables among the treatment-by-moderator subgroups through weighting to make the interaction term independent of the measured variables. It is very important to for researchers to plan and use as many variables as possible that are associated with the outcome, treatment, and hypothesized moderator variables to reduce hidden bias due to omitted variables (e.g., Steiner et al., 2010).

Another limitation of this study that is common to other propensity score methods is that the results may be subject to bias from the propensity score model misspecification. In the demonstration, we used

multinomial logistic regression model and used the iterative process to revise our model by including the interaction terms and higher order terms of covariates to reach best covariate balance. Some other methods for estimating propensity scores, e.g., random forests or boosted regression (Cham & West, 2016; McCaffrey et al., 2004), might produce better covariate balance.

In addition, although the double robustness adjustment using propensity score weighting while controlling for covariates in the outcome model generally reduces selection bias if either the propensity score model, the outcome model, or both are correctly specified, it may fail to reduce bias if both models are mis-specified. Other adjustment approaches should be considered (see Kang & Schafer, 2007, and Tsiatis & Davidian, 2007, for deeper discussion).

Future work

One important direction for future work on this track is to explore sensitivity analysis methods to assess the robustness of inferences when there is an unmeasured moderator variable. Researchers may extend Rosenbaum's (2002) Gamma parameter based on Wilcox rank statistics, or other statistics based on regression (Frank, 2000; Frank et al., 2013; Hong & Raudenbush, 2006; Lin et al., 1998; Pan & Frank, 2003) to the causal moderation analysis framework.

The second direction for future work is to extend the study to non-binary moderators. The current framework can be easily extended for a multi-valued treatment variable ($z > 2$) and a multi-valued categorical moderator ($r > 2$) by converting the two dimensions (treatment by moderator, $k = z \times r$) of design to one dimension of design with k categories and using the procedure discussed in this article. Future work includes developing a causal framework and approaches to conducting a moderated treatment effect analysis for targeted subgroups defined by continuous moderators. For instance, the generalized propensity score method that was developed for analyzing continuous treatment variables (Hirano & Imbens, 2004; Imai & van Dyk, 2004) can be extended for the analysis of moderated treatment effect with continuous moderators using stratification.

In addition, all estimands in our framework are expressed on an additive scale by looking at mean differences. Another direction for future work is to consider risk ratios or odds ratios for binary outcomes.

In summary, we provide a causal moderation analysis framework and estimation approach for eliminating the influence of the other measured covariates/moderators on the estimate of AMTE. In addition, the AMTS estimates provide an approach for identifying the targeted policy-relevant subgroup for effective intervention.

Article information

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was supported by Grant [1913563, 1552535, 1760884] from the National Science Foundation.

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

References

Abadie, A., & Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, *76*(6), 1537–1557. https://doi.org/10.3982/ECTA6474

Ahern, J. (2018). Start with the "C-word," follow the roadmap for causal inference. *American Journal of Public Health*, *108*(5), 621–621. https://doi.org/10.2105/AJPH.2018.304358

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Sage.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444–472. https://doi.org/10.2307/2291629

Austin, P. C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, *27*(12), 2037–2049. https://doi.org/10.1002/sim.3150

Austin, P. C. (2017). Double propensity-score adjustment: A solution to design bias or bias due to incomplete matching. *Statistical Methods in Medical Research*, *26*(1), 201–222. https://doi.org/10.1177/0962280214543508

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, *34*(28), 3661–3679. https://doi.org/10.1002/sim.6607

Bansak, K. (2018). *A generalized framework for the estimation of causal moderation effects with randomized treatments and non-randomized moderators*. Working paper. https://arxiv.org/abs/1710.02954

Barnett, W. S. (2011). Effectiveness of early educational intervention. *Science (New York, N.Y.)*, *333*(6045), 975–978. https://doi.org/10.1126/science.1204534

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*(6), 1173–1182. https://doi.org/10.1037//0022-3514.51.6.1173

Bun, M. J. G., & Harrison, T. D. (2019). OLS and IV estimation of regression models including endogenous interaction terms. *Econometric Reviews*, *38*(7), 814–827. https://doi.org/10.1080/07474938.2018.1427486

Cham, H., & West, S. G. (2016). Propensity score analysis with missing data. *Psychological Methods*, *21*(3), 427–445. https://doi.org/10.1037/met0000076

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. https://doi.org/10.1177/1098214014553261

Dong, N., & Kelcey, B. (2020). A review of *Causality in a social world: Moderation, mediation, and spill-over*. *Journal of Educational and Behavioral Statistics*, *45*(3), 374–378. https://doi.org/10.3102/1076998619881791x10060

Dong, J., Zhang, J. L., Zeng, S., & Li, F. (2020). Subgroup balancing propensity score. *Statistical Methods in Medical Research*, *29*(3), 659–676. https://doi.org/10.1177/0962280219870836

Dong, N. (2012, March). *Causal moderation analysis using propensity score methods* [Paper presentation]. Paper Presented at the Spring 2012 Conference of the Society for Research on Educational Effectiveness (SREE), Washington, DC.

Egami, N., & Imai, K. (2019). Causal interaction in factorial experiments: Application to conjoint analysis. *Journal of the American Statistical Association*, *114*(526), 529–540. https://doi.org/10.1080/01621459.2018.1476246

Frank, K. A. (2000). The impact of a confounding variable on a regression coefficient. *Sociological Methods & Research*, *29*(2), 147–194. https://doi.org/10.1177/0049124100029002001

Frank, K. A., Maroulis, S. J., Duong, M. Q., & Kelcey, B. M. (2013). What would it take to change an inference? Using Rubin's causal model to interpret the robustness of causal inferences. *Educational Evaluation and Policy Analysis*, *35*(4), 437–460. https://doi.org/10.3102/0162373713493129

Frazier, P. A., Tix, A. P., & Barron, K. E. (2004). Testing moderator and mediator effects in counseling psychology research. *Journal of Counseling Psychology*, *51*(1), 115–134. https://doi.org/10.1037/0022-0167.51.1.115

Hernán, M. A. (2018). The C-word: Scientific euphemisms do not improve causal inference from observational data. *American Journal of Public Health*, *108*(5), 616–619. https://doi.org/10.2105/AJPH.2018.304337

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, Ltd. https://doi.org/10.1002/0470090456.ch7.

Hong, G. (2015). *Causality in a social world: Moderation, mediation, and spill-over*. Wiley-Blackwell.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, *101*(475), 901–910. https://doi.org/10.1198/016214506000000447

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings among experimentalists and observationalists in causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *171*(2), 481–502. https://doi.org/10.1111/j.1467-985X.2007.00527.x

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. https://doi.org/10.1198/016214504000001187

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, *87*(3), 706–710. https://doi.org/10.1093/biomet/87.3.706

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, *86*(1), 4–29. https://doi.org/10.1162/003465304323023651

Kang, J. D. Y., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*(4), 523–539. https://doi.org/10.1214/07-STS227

Kraemer, H. C., Kiernan, M., Essex, M., & Kupfer, D. J. (2008). How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychology : Official Journal of the Division of Health Psychology, American Psychological Association*, *27*(2S), S101–S108. https://doi.org/10.1037/0278-6133.27.2(Suppl.).S101

Kraemer, H. C., Stice, E., Kazdin, A., Offord, D., & Kupfer, D. (2001). How do risk factors work together? Mediators, moderators, and independent, overlapping, and proxy risk factors. *The American Journal of Psychiatry*, *158*(6), 848–856. https://doi.org/10.1176/appi.ajp.158.6.848

Kurth, T., Walker, A. M., Glynn, R. J., Chan, K. A., Gaziano, J. M., Berger, K., & Robins, J. M. (2006). Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology*, *163*(3), 262–270. https://doi.org/10.1093/aje/kwj047

Lenis, D., Nguyen, T. Q., Dong, N., & Stuart, E. A. (2019). It's all about balance: Propensity score matching in the context of complex survey data. *Biostatistics (Oxford, England)*, *20*(1), 147–163. https://doi.org/10.1093/biostatistics/kxx063

Lin, D. Y., Psaty, B. M., & Kronmal, R. A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, *54*(3), 948–963. https://doi.org/10.2307/2533848

Lipsey, M. W., Hofer, K. G., Dong, N., Farran, D. C., & Bilbrey, C. (2013). *Evaluation of the tennessee voluntary prekindergarten program: Kindergarten and first grade follow-up results from the randomized control design (Research Report)*. Vanderbilt University, Peabody Research Institute.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Sage Publications, Inc.

Magnuson, K. A., Ruhm, C., & Waldfogel, J. (2007). Does prekindergarten improve school preperation and performance? *Economics of Education Review*, *26*(1), 33–51. https://doi.org/10.1016/j.econedurev.2005.09.008

Mayer, A., Dietzfelbinger, L., Rosseel, Y., & Steyer, R. (2016). The EffectLiteR approach for analyzing average and conditional effects. *Multivariate Behavioral Research*, *51*(2–3), 374–391. https://doi.org/10.1080/00273171.2016.1151334

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. https://doi.org/10.1037/1082-989X.9.4.403

Moreno-Serra, R. (2007). *Matching estimators of average treatment effects: a review applied to the evaluation of health care programmes*. Health, Econometrics and Data Group (HEDG) Working Paper, ISSN 1751-1976. University of York. Retrieved November 17, 2020, from https://www.york.ac.uk/media/economics/documents/herc/wp/07_02.pdf

Neyman, J. (1923/1990). On the application of probability theory to agricultural experiments: Essay on principles, section 9. (translated in 1990). *Statistical Science*, *5*, 465–480. https://doi.org/10.1214/ss/1177012031

Pan, W., & Frank, K. A. (2003). A probability index of the robustness of a causal inference. *Journal of Educational and Behavioral Statistics*, *28*(4), 315–337. https://doi.org/10.3102/10769986028004315

Ridgeway, G., Kovalchik, S. A., Griffin, B. A., & Kabeto, M. U. (2015). Propensity score analysis with survey weighted data. *Journal of Causal Inference*, *3*(2), 237–249. https://doi.org/10.1515/jci-2014-0039

Ridgeway, G., McCaffey, D. F., Morral, A., Burgette, L., & Griffin, B. A. (2021). *Toolkit for Weighting and Analysis of Nonequivalent Groups: A tutorial for the twang package*. Retrieved on June 16, 2021, from http://cran.r-project.org/web/packages/twang/vignettes/twang.pdf

Robins, J. M., Hernán, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology (Cambridge, Mass.)*, *11*(5), 550–560. https://doi.org/10.1097/00001648-200009000-00011

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). 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. https://doi.org/10.1093/biomet/70.1.41

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Rubin, D. B. (1980). Bias reduction using Mahalanobis metric matching. *Ets Research Bulletin Series*, *36*(2), 293–298. https://doi.org/10.1002/j.2333-8504.1978.tb01164.x

Rubin, D. B. (1986). Comment. *Journal of the American Statistical Association*, *81*(396), 961–962. https://doi.org/10.1080/01621459.1986.10478355

Rubin, D. B. (2010). Reflections stimulated by the comments of Shadish (2010) and West and Thoemmes (2010). *Psychological Methods*, *15*(1), 38–46. https://doi.org/10.1037/a0018537

Sagarin, B. J., West, S. G., Ratnikov, A., Homan, W. K., Ritchie, T. D., & Hansen, E. J. (2014). Treatment noncompliance in randomized experiments: Statistical approaches and design issues. *Psychological Methods*, *19*(3), 317–333. https://doi.org/10.1037/met0000013

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, *15*(3), 250–267. https://doi.org/10.1037/a0018719

Tourangeau, K., Nord, C., Le, T., Sorongon, A. G., & Najarian, M. (2009). *Early childhood longitudinal study, Kindergarten class of 1998-99 (ECLS-K), combined user's manual for the ECLS-K eight-grade and K-8 full sample data files and electronic codebooks* (NCES 2009-004).

Tsiatis, A. A., & Davidian, M. (2007). Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete Data. *Statistical Science : a Review Journal of the Institute of Mathematical Statistics*, *22*(4), 569–573. https://doi.org/10.1214/07-STS227

U.S. Department of Education, National Center for Education Statistics. (2009). *Early childhood longitudinal study, kindergarten class of 1998-99 (ECLS-K) kindergarten through eighth grade full sample public-use data and documentation* (NCES 2009005).

Xu, S., Ross, C., Raebel, M. A., Shetterly, S., Blanchette, C., & Smith, D. (2010). Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. *Value in Health : The Journal of the International Society for Pharmacoeconomics and Outcomes Research*, *13*(2), 273–277. https://doi.org/10.1111/j.1524-4733.2009.00671.x

Yang, S., Lorenzi, E., Papadogeorgou, G., Wojdyla, D. M., Li, F., & Thomas, L. E. (2021). Propensity score weighting for causal subgroup analysis. *Statistics in Medicine*, *40*(19), 4294–4309. https://doi.org/10.1002/sim.9029