Estimation of Causal Moderated Treatment Effects Under the Potential Outcomes Framework: An Application of the Propensity Score

Authors and Affiliations:

Nianbo Dong
University of North Carolina at Chapel Hill
116 Peabody Hall, CB 3500
Chapel Hill, NC 27599
Phone: (919)843-9553
dong.nianbo@gmail.com

Ben Kelcey
University of Cincinnati
3311B RECCENTER
Cincinnati, Ohio 45221
Tel: 513-556-3608
ben.kelcey@gmail.com

Jessaca Spybrook
Western Michigan University
3571 Sangren Hall
Kalamazoo, Michigan 49008
Phone: (269) 387-3889
jessaca.spybrook@wmich.edu

Estimation of Causal Moderated Treatment Effects Under the Potential Outcomes Framework: An Application of the Propensity Score

Purpose

The purpose of this study is to provide a framework for the causal interpretation of moderator analyses based on the potential outcomes framework (Newman, 1923; Rubin, 1974), and develop estimation approaches under different assumptions and types of moderator effects.

Background

Moderation analyses examine whether the effects of an intervention differ by the moderator subgroups. Conventional moderator analyses implicitly provide “a correlational analysis” (Baron & Kenny, 1986, p.1174) because most experimental designs randomly assign the treatment but not moderator values. As a result, conventional moderation analyses typically produce descriptive reports of effect variation that plausibly conflate true effect differences among moderator subgroups with effects produced by other pretreatment variables. For example, consider an experiment designed to assess if the effects of a program differ by the students’ free or reduced lunch eligibility (FRL; focal moderator). Similar to the main effect, the treatment effect can be estimated without bias for each pretest subgroup. However, the observed difference between the subgroup treatment effects (i.e., moderation effects) does not necessarily have a causal interpretation because even in a randomized design this difference may simply represent an artifact of an unobserved pretreatment variable that is an antecedent of this subgroup difference. Applied to our example, one potential pretreatment variable that might confound the difference between the eligible and ineligible for FRL subgroups is special education status. That is, because the FRL eligibility (focal moderator) is not randomly assigned, it is possible that the subgroup of students eligible for FRL includes more students with special education needs. More specifically, because special education status precedes FRL eligibility in a given year, it may causally responsible for the observed FRL subgroup differences. For this reason, a host of literature has recently taken up causal inference for moderated treatment effects (e.g., Dong, 2012, 2015; Dong & Kelcey, 2019; Bansak, 2018). However, a general framework and corresponding statistical procedures to guide causal moderation analyses still lack.

Theoretical Framework and Methods

Conventional Moderation Under the Potential Outcomes Framework

Using the potential outcomes framework(Newman, 1923; Rubin, 1974), Hong (2015) defined the moderated treatment effect for an individual or contextual characteristic as the average treatment effect difference between two moderator subgroups. Because Hong (2015) only considers the potential outcomes with treatment but not the moderator, this definition is still under the conventional moderation analysis framework, and it is not necessarily causal in nature. Although the average treatment effect estimates for the subgroups are unbiased, the treatment effect differences between two subgroups could be due to variables other than the hypothesized moderator variable as we discussed above (Dong, 2015).

$$\text{Conventional Moderation Effect} = E[Y(1) – Y(0)|R = 1] - E[Y(1) – Y(0)|R = 0], \qquad (1)$$

where Y indicates the outcome and R indicates the moderator subgroup (Hong, 2015, p. 134).

Causal Moderation Under the Potential Outcomes Framework

The potential outcomes and paths are presented at Figure 1. Similarly as causal inference for the main effect analysis which distinguishes the average treatment effect (ATE) and the average treatment effect on the treated (ATT) (Imai, King, & Stuart, 2008; Imbens, 2004; Kurth et al., 2006; McCaffrey, Ridgeway, & Morral, 2004), there are two types of estimands for the causal moderated treatment effects. Depending on the sample of interest to make inference, there is the average moderated treatment effect (AMTE) for all the sample, and the average moderated treatment effect on certain subgroups (AMTS). The average moderated treatment effect (AMTE) for all sample can be defined as:

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)], (2)

where Y(Z, R) represents the potential outcome for treatment condition Z and moderator group R (Dong, 2012, 2015; Dong & Kelcey, 2019). This AMTE definition is same as the one for the factorial design where two concurrent treatments exist by Hong (2015), and same as the average marginal interaction effect (AMIE) by Egami and Imai (2018) and the average treatment moderation effect (ATME) by Bansak (2018).

The average moderated treatment effect on certain subgroups (AMTS) can be defined as:

AMTS = E[Y(1,1) – Y(0,1)| Z = z, R = r] - E[Y(1,0) – Y(0,0)| Z = z, R = r], (3)

where z = 0 or 1 and r = 0 or 1. It defines the average moderated treatment effect on a subgroup (Z = z and R = r). For example, the subgroups in AMTS may be the treated moderator group 2 (Z = 1 and R = 1).

The assumptions for causal moderation include: the stable unit treatment value assumption, ignobility of treatment and moderator, independence of treatment and moderator, and treatment-by-moderator common support.

Estimation

Under the potential outcomes framework, we can use propensity score analysis to estimate AMTE and AMTS. The procedure follows. (1) We can 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. (2) We can use multinomial logistic regression to estimate the propensity scores for individual i of being in certain group given covariates (X): $\pi_i = pr(S_i = s|X_i)$, where s = 1, 2, 3, or 4. (3) Then we can use the inverse probability of treatment weighting (IPTW) to estimate the average moderated treatment effect (AMTE). The weights are $w_i = \frac{1}{\widehat{\pi_i}}$, where $\widehat{\pi_i}$ is the estimated propensity score of being in the actual subgroup.

To estimate the average moderated treatment effect on certain subgroups (AMTS), we can use the odds ratio of propensity score as weight. The denominator of the odds ratio is the propensity score of being in the actual subgroup and the nominator is the propensity score of being in the interested subgroup. For example, if the sample of interest is treated moderator group 2 (Z = 1 and R = 1, i.e., S =4), the weight, $w_i = \frac{\widehat{\pi_i}|\ S_i=4}{\widehat{\pi_i}|\ S_i=s}$ if $S_i = s$, where s = 1, 2, 3, or 4.

In addition, we can use propensity score matching (e.g., greedy matching, optimal matching) to estimate AMTS. First, we can estimate the propensity score of being in the interested subgroup, s. Then we match sample from the other subgroups with the interested subgroup based on the propensity score of being in the interested subgroup. After balance checking we can estimate AMTS using the combined sample.

Application

We applied the methods discussed above to estimate AMTE and AMTS using the data from an evaluation of the Incredible Years teacher classroom management program (Reinke, Herman, & Dong, 2018). The research questions is whether the effects of the intervention differ by the pretest (low vs. high) of the social competence measures. The descriptive statistics of covariates are presented at Table 1. The covariates are much more balanced after weighted by AMTE and AMTS (Figure 2). The conventional moderation analysis produced significant moderation effect size difference (*d*= 0.163, *p* = 0.0247) on social competence, however, the AMTE estimate based on the propensity score weighting produced insignificant moderation effect size difference (*d*= 0.132, *p* = 0.0543) (Table 2). In addition, the AMTS estimates varied by the sample of interest.

Conclusion and Significance

This paper presents the potential outcome framework and propensity score methods to estimate two types of causal moderated treatment effects: the average moderated treatment effect (AMTE) for all the sample, and the average moderated treatment effect on certain subgroups (AMTS). The results would help researchers to distinguish sample of interest and identify true moderators that differentiate treatment effects.

References:

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

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

Dong, N. (2012, March). *Causal moderation analysis using propensity score methods*. Paper presented at the Spring 2012 Conference of the Society for Research on Educational Effectiveness (SREE), Washington, DC. (ERIC #: ED530452).

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. doi: 10.1177/1098214014553261.

Dong, N. & Kelcey, B. (2019). A review of *Causality in a Social World: Moderation, Mediation, and Spill-over*. *Journal of Educational and Behavioral Statistics*. Paper accepted for publication.

Egami, N. & Imai, K. (2018) Causal Interaction in Factorial Experiments: Application to Conjoint Analysis, *Journal of the American Statistical Association*, Advance online publication. DOI: 10.1080/01621459.2018.1476246

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

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings among experimentalists and observationalists in causal inference. *J. Roy. Statist. Soc. Ser. A, 171,* 481–502.

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

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,* 262–270.

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,* 403–425.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, section 9. (translated in 1990). *Statistical Science, 5*, 465-480.

Reinke, W. M., Herman, K. C., & Dong, N. (2018). The Incredible Years Teacher Classroom Management program: Outcomes from a group randomized trial. *Prevention Science, 19*(8), 1043–1054. doi: 10.1007/s11121-018-0932-3

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

Figure 1: Potential Outcomes and Paths

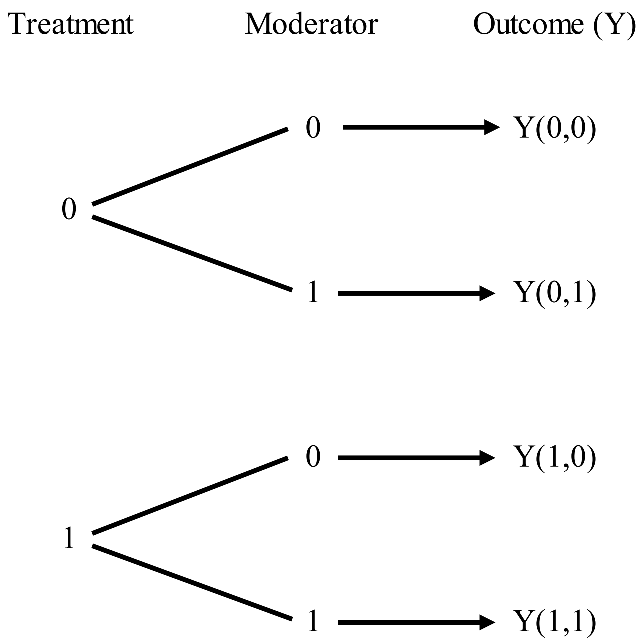

Figure 2: Covariate Balance Checking before and after Propensity Score Weighting

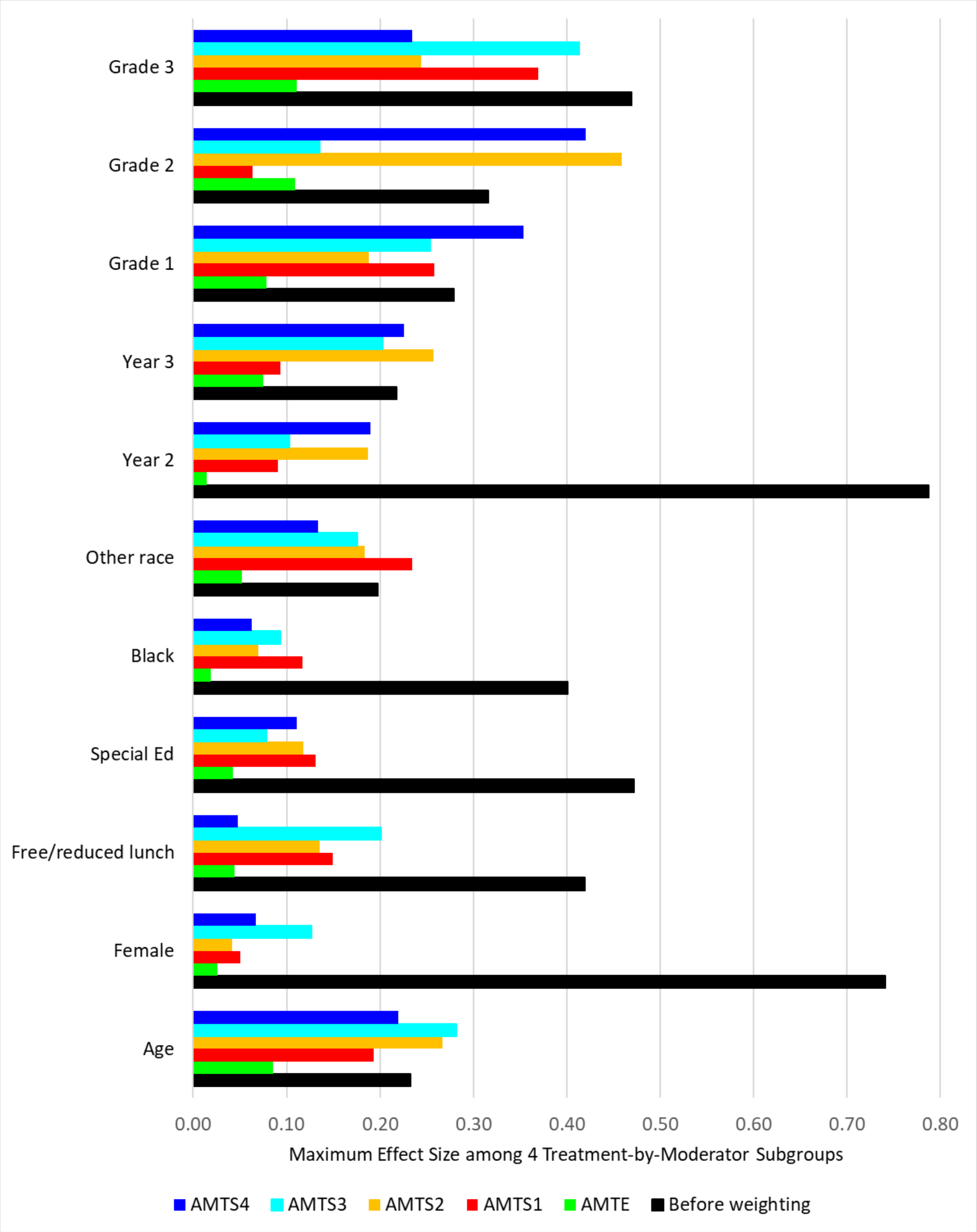

Table 1: Descriptive Statistics of Covariates among 4 Treatment-by-Moderator Subgroups

Treatment-by-Moderator (S)	1		2		3		4		Maximum Effect Size among 4 Groups
Treatment (Z)	0		0		1		1		
Low Pretest (R)	0		1		0		1		
Variable	Mean	SD	Mean	SD	Mean	SD	Mean	SD	
Age	7.06	1.09	6.99	1.10	7.25	1.03	7.21	1.34	0.23
Female	0.66	0.48	0.35	0.48	0.62	0.49	0.30	0.46	0.74
Free/reduced lunch	0.49	0.50	0.69	0.46	0.51	0.50	0.69	0.47	0.42
Special Ed	0.06	0.24	0.17	0.38	0.03	0.17	0.15	0.36	0.47
Black	0.67	0.47	0.80	0.40	0.71	0.45	0.84	0.37	0.40
Other race	0.02	0.14	0.05	0.21	0.02	0.14	0.02	0.12	0.20
Year 2	0.15	0.35	0.43	0.50	0.10	0.30	0.39	0.49	0.79
Year 3	0.43	0.50	0.33	0.47	0.39	0.49	0.41	0.49	0.22
Grade 1	0.32	0.47	0.29	0.45	0.34	0.48	0.22	0.41	0.28
Grade 2	0.23	0.42	0.31	0.46	0.27	0.44	0.17	0.38	0.32
Grade 3	0.17	0.38	0.11	0.32	0.23	0.42	0.30	0.46	0.47
Sample size	198		195		206		197		

Note: This study sample includes 796 students in 104 classrooms in 9 schools.

Table 2: Summary of Effect Size Differences using Different Methods

Analysis	Effect Size Difference	*p*-value	Lower Bound of 95% CI	Upper Bound of 95% CI
Conventional	0.163	0.0247	0.021	0.306
AMTE	0.132	0.0543	-0.002	0.266
AMTS1	0.120	0.1172	-0.030	0.271
AMTS2	0.172	0.0550	-0.004	0.348
AMTS3	0.080	0.2489	-0.056	0.217
AMTS4	0.177	0.0038	0.057	0.297

Note: Conventional refers to the 3-level HLM moderation analysis (students nested within classrooms, and classrooms nested within schools) without propensity score weighting; AMTE refers to the 3-lefel HLM moderation analysis weighted by the AMTE weights; AMTS1 refers to the 3-lefel HLM moderation analysis weighted by the AMTS weights, where is the sample of interest S = 1; AMTS2 refers to the 3-lefel HLM moderation analysis weighted by the AMTS weights, where is the sample of interest S = 2; AMTS3 refers to the 3-lefel HLM moderation analysis weighted by the AMTS weights, where is the sample of interest S = 3; AMTS4 refers to the 3-lefel HLM moderation analysis weighted by the AMTS weights, where is the sample of interest S = 4.