sociological science

Supplement to:

Serrano-Serrat, Josep. 2026. “Quantities of Interest for Interactions and the Pitfalls of Assuming Linear Treatment Effects” Sociological Science 13: 945-970.

Online Supplementary Material

Quantities of Interest for Interactions and the Pitfalls of Assuming Linear Treatment Effects

Josep Serrano-Serrat

IPP-CSIC

Appendix A Non-linearities, Distributions, and Bias

A.1 CAMEs as a Function of CAPEs

To see how CAME should be written as a function of CAPE, note that:

$$CAPE(d_s, x) = \mathbb{E}\left[\frac{\partial Y_i(d)}{\partial d} \mid d_i = d_s, X = x\right]$$

Then, to see the equivalence in Equation 1:

$$\tau^{CAME}(x) = \frac{1}{N_x}\sum_{i=1}^{N_x}\frac{\partial Y_i(d)}{\partial d} = \mathbb{E}\left[\frac{\partial Y_i(d)}{\partial d}|X = x\right] =$$

$$= \int_d \mathbb{E}\left[\frac{\partial Y_i(d)}{\partial d} \mid d_i = d, X = x\right] f_{D|x}(d)\, dd = \int_d CAPE(d, x) \cdot f_{D|x}(d)\, dd$$

A.2 Quantities of Interest

To see the conditions under which Equation 1 may be different from Equation 2, define $\beta_j = \int_d CAPE(d,x) f_{D|X} \mathrm{d}d$ and $\tilde{\beta}_j = \int_d CAPE(d,x) f(d) \mathrm{d}d$. β_j is the CAME, and $\tilde{\beta}_j$ is the weighted-CAME: the AME of D on Y in the $j-th$ tercile of X assuming that the distribution of D in this level of the moderator is the same as the overall distribution of D. Then, $\Delta CAME$ and $AIPE$ can be written as

$$\Delta CAME(d, j', j) = \beta_{j'} - \beta_j$$

$$AIPE(d, x', j) = \tilde{\beta}_{j'} - \tilde{\beta}_j$$

Thus, generally, Equations 2 and 1 are equal when, for all j, $\beta_j = \tilde{\beta}_j$. I have argued that this happens with linear effects (i.e.,$CAPE(d,x) = ME_{Y,D}(j)$), and/or when the distribution of D in tercile j, is the same as the overall distribution.

- **Distribution:** To see this, note that, when for all j $f_{D|X}(d|X = j) = f(d)$, the result follows directly.
- **Non-Linearities.** When the relationship between D and Y is linear for all levels of the moderator, it follows that $CAPE(d,x) = ME_{Y,D}(j)$. Being this the case it can be shown

that $\beta_j = \tilde{\beta}_j = ME_{Y,D}(j)$. To see this,

$$\tilde{\beta}_j = \int_d CAPE(d,x) f(d) \mathrm{d}d = \int_d ME_{Y,D}(j) f(d) \mathrm{d}d = ME_{Y,D}(j) \int_d f(d) \mathrm{d}d = ME_{Y,D}(j)$$

$$\beta_j = \int_d CAPE(d,x) f_{D|X}(d|X=j) = ME_{Y,D}(j) \int_d f_{D|X}(d|X=j) \mathrm{d}d = ME_{Y,D}(j)$$

There are other cases where they can be the same that depend specifically on the functional form of the relationship of interest and the distribution of the treatment variable at different levels of the distribution. However, there is no theoretical prior about why it should be the case that both equations are the same when these two conditions do not hold.

Exploring Sign of Difference between ΔCAME and AIPE

This section studies, under some conditions, when $\Delta CAME$ is expected to be higher/lower than AIPE. In a first step, I rewrite Equations 1 and 2,

$$\Delta CAME(d, j', j) = \beta_{j'} - \beta_j$$

$$AIPE(d, x', j) = \tilde{\beta}_{j'} - \tilde{\beta}_j$$

being $\beta_j = \int_d CAPE(d,x) f_{D|X} \mathrm{d}d$ and $\tilde{\beta}_j = \int_d CAPE(d,x) f(d) \mathrm{d}d$. β_j is the CAME when $X = j$, and $\tilde{\beta}_j$ is the weighted-CAME: the AME of D on Y conditional on $X = j$ assuming that the distribution of D in this level of the moderator is the same as the overall distribution of D. Since the CAME and the wieghted-CAME are defined as the multiplication between treatment effects and the distribution of D, in what follows I explore the role of each of these dimensions.

To keep the discussion simple, along this section, I will consider cases where there is first order stochastic dominance of the distribution of D at different values of X. That is, the distribution of D at $X = 0$ stochastically dominates the distribution of D at $X = 1$: $F_0(d) \geq F(d) \geq F_1(d)$ $\forall d \in R$, with F_x being the cumulative distribution function of D when $X = x$, and $F(d)$ is the overall cumulative distribution of D. $F_0(d) \geq F(d)$ implies that there are lower values of D when $X = 0$ than in the overall sample.

Turning to marginal effects, I will explore whether the function, or an interval of it, is (strictly) concave or convex. For instance, having concave (convex) treatment effects implies that the effects are smaller (larger) the higher the value of D.

In a final step, I analyze the relationship between β_j and $\tilde{\beta}_j$, this will allow us to understand the differences between the $\Delta CAME$ and the AIPE.[21] As a continuation of Table 1, I consider a two-by-two table depending on the type of distribution (based on first order stochastic dominance), and the functional form of the treatment effects (whether they are concave or convex). Note that these simplifications are useful because for any continuous and differentiable treatment effects function and distribution of D, all intervals can be classified in one of the four groups. The results are summarized in Table 3.

Table 3: Relationship CAME and Weighted CAME

		Distribution of D at different X	
		$F_j(d) \geq F_{j'}(d)$	$F_j(d) \leq F_{j'}(d)$
Treatment Effects	Concave	$\beta_j < \tilde{\beta}_j$	$\beta_j > \tilde{\beta}_j$
	Convex	$\beta_j > \tilde{\beta}_j$	$\beta_j < \tilde{\beta}_j$

The logic of this table is as follows. If treatment effects are concave (convex) on an interval, then its slope is monotonically decreasing (increasing) on that interval. Then, for example, if $F_j(d) \geq F(d)$ – which indicates that there are relatively more low values of of D when $X = j$ than for the overall distribution – it follows that the CAME will be lower than the weighted-CAME.

A.3 Empirical Estimands and Identification Assumptions

As it is well-known, the empirical estimand of the D-CAME is the following:

$$\theta^{\Delta CAME} = \frac{\partial \mathbb{E}[Y|D, X = 1]}{\partial d} - \frac{\partial \mathbb{E}[Y|D, X = 0]}{\partial d} \tag{9}$$

Then the empirical estimands of IPE and AIPE are, respectively:

$$\theta^{IPE}(d_i, 1, 0) = \frac{\partial \mathbb{E}[Y(d_i)|X = 1]}{\partial d} - \frac{\partial \mathbb{E}[Y(d_i)|X = 0]}{\partial d} \tag{10}$$

$$\theta^{AIPE}(1, 0) = \frac{1}{n} \sum_{i=1}^{n} \theta^{IPE}(d_i, 1, 0) \tag{11}$$

with n being the number of observations in the sample. The empirical estimand for the IPE tells us that we aim to capture the between-group differences in the slopes of the (expected)

[21] Concretely, $\Delta CAME - AIPE = (\beta_1 - \tilde{\beta}_1) + (\tilde{\beta}_0 - \beta_0)$.

outcome, at a given value of the treatment variable. The crucial element here is that we cannot observe the counterfactual outcome in case of experiencing an increase in the treatment dose. Instead, we compare different individuals that differ in terms of the intensity of the treatment (that is, we aim to estimate $\frac{\partial \mathbb{E}[Y(d_i)]}{\partial d}$ rather than $\frac{\partial Y_i(d)}{\partial d}$).

The identification assumptions described in the text are described as follows:

Assumption 1 (Unconfoundedness): Conditional on the moderator X and control variables Z, treatment assignment is independent of potential outcomes:

$$E[Y_i(d)|D_i = d, X_i = x, Z_i] = E[Y_i(d)|X_i = x, Z_i] \tag{12}$$

This assumption states that after conditioning on X and Z, individuals experiencing different treatment intensities are otherwise comparable. In causal inference settings, this is the standard "selection on observables" assumption.

Assumption 2 (Positivity): For all treatment values d in the support of D and both values of the moderator $x \in \{0, 1\}$, there exists a positive probability of observing that treatment level: $P(D = d|X = x, Z) > 0$ for all relevant combinations.

This assumption implies that we require common support of D at different values of the moderator X. This assumption would be violated, for instance, if the range of D is different at different values of X. Being this the case, estimates of the IPE, and therefore of the AIPE, would rely on extrapolation, misleading the results (Hainmueller et al., 2019; King and Zeng, 2006). Being this the case, researchers should trim the data, making clear that the target population has changed (King and Zeng, 2006; Liu et al., 2025).

Assumption 3 (Smoothness): the potential outcome $Y_i(d)$ is continuously differentiable in d for all values of d.

The plausibility of these assumptions, specially the unconfoundedness assumption, depends on the specific application and must be justified theoretically.

Appendix B Data Generating Process

This section described the DGPs that appear along the text.

Example Introduction

The example in the introduction is a draw from a DGP with $N = 2000$ observations with a continuous treatment variable D and a binary moderator X.

Treatment variable. The treatment variable is drawn from a uniform distribution:

$$D_i^{(0)} \sim \text{Uniform}(0, 3) \tag{13}$$

Moderator. The binary moderator X is generated as a function of D plus random noise:

$$X_i = 1\{D_i + \varepsilon_i > 1.5\}, \quad \varepsilon_i \sim \mathcal{N}(0, 4) \tag{14}$$

This specification induces correlation between D and X, mimicking realistic settings where treatment intensity and moderator status are related. This is crucial for the part of the argument about D-CAME being different than AIPE.

Outcome. The outcome is generated with a cubic treatment effect and additive noise:

$$Y_i = -3D_i + 2D_i^2 + u_i, \quad u_i \sim \mathcal{N} \sim \mathcal{N}(0, 36) \tag{15}$$

Crucially, the true DGP features positive D-CAME, and zero AIPE.

Simulation Exercise

The example uses a DGP with $N = 2000$ observations designed to illustrate settings where both D-CAME and AIPE are identical and non-zero, but linear models fail to recover either estimand.

Moderator. The binary moderator X is generated independently:

$$X_i \sim \text{Bernoulli}(0.5) \tag{16}$$

Treatment variable. The treatment variable follows a log-normal distribution, independent of X:

$$D_i \sim \text{LogNormal}(0, 1) \tag{17}$$

To avoid extreme observations (King and Zeng, 2006; Liu et al., 2025), D is trimmed at its 95th percentile.

Outcome. The outcome is generated with group-specific threshold effects. Let $q_{0.75}^{(1)}$ denote the 75th percentile of D among observations with $X = 1$. Then:

$$Y_i = \begin{cases} c + u_i & \text{if } X_i = 0 \\ c - D_i \cdot \mathbb{1}\{D_i \leq q_{0.75}^{(1)}\} + D_i \cdot \mathbb{1}\{D_i > q_{0.75}^{(1)}\} + u_i & \text{if } X_i = 1 \end{cases} \tag{18}$$

where c is a constant and $u_i \sim \mathcal{N}(0, \sigma^2)$. This specification implies that for $X = 0$, the outcome is independent of treatment, while for $X = 1$, the marginal effect is -1 for the bottom 75% of the treatment distribution and $+1$ for the top 25%.

Derivation of D-CAME. Since D is independent of X, the D-CAME equals the AIPE. The difference in D-CAME across groups is:

$$\tau^{D-CAME} = \tau^{AIPE} = \mathbb{E}\left[\frac{\partial Y}{\partial D} \,\middle|\, X = 1\right] - \mathbb{E}\left[\frac{\partial Y}{\partial D} \,\middle|\, X = 0\right] = \mathbb{E}\left[\frac{\partial Y}{\partial D} \,\middle|\, X = 1\right] - 0 = \tag{19}$$

$$= \Pr(D \leq q_{0.75}^{(1)} \mid X = 1) \cdot (-1) + \Pr(D > q_{0.75}^{(1)} \mid X = 1) \cdot (+1) =$$

$$= 0.75 \cdot (-1) + 0.25 \cdot (+1) = -0.75 + 0.25 = -0.5$$

Example in Section 4

The example uses a DGP with $N = 3000$ observations with a continuous treatment variable D and a binary moderator X.

Moderator. The binary moderator X is generated independently of treatment:

$$X_i \sim \text{Bernoulli}(0.5) \tag{20}$$

Treatment variable. The treatment variable is generated with different distributions across moderator groups. For the $X = 0$ group:

$$D_i^{(0)} \sim \text{Uniform}(0, 1) \text{ if } X_i = 0 \tag{21}$$

For the $X = 1$ group, we first draw from a Weibull distribution and then apply transformations:

$$\tilde{D}_i \sim \text{Weibull}(k = 1.6, \lambda = 1.4) \text{ if } X_i = 1 \tag{22}$$

with shape parameter $k = 1.6$ and scale parameter $\lambda = 1.4$. Values exceeding 3 are redrawn from Uniform(0,1). We then rescale within each group to $[0, 1]$.

Outcome. The outcome is generated with group-specific functional forms:

$$Y_i = \begin{cases} 2 + 4D_i - 4D_i^2 + u_i & \text{if } X_i = 0 \\ 2 + 4D_i - 4D_i^2 - (1 - D_i)^3 + 0.1(1 - D_i) + u_i + v_i & \text{if } X_i = 1 \end{cases} \tag{23}$$

where $u_i \sim \mathcal{N}(0, 0.25)$ and $v_i \sim \mathcal{N}(0, 0.25)$ are independent normal errors.

Appendix C Estimation

First, researchers should display *predicted values* of Y as a function of D for different groups. Concretely, at the mean level of the covariates ($\overline{Z}$), the predicted values for given values of D and X take the following form:

$$\hat{Y}(d, x) = \hat{\mu} + \hat{\alpha}x + \hat{g}_0(d) + \hat{g}_1(d) + \hat{\delta}'\overline{Z} \tag{24}$$

With binary X, displaying $\hat{Y}(d, 1)$ and $\hat{Y}(d, 0)$ for different values of d allows researchers to see more clearly the relationship between the outcome variable and the treatment, which is crucial. As stated by Long and Mustillo (2021), we get intuition on the marginal effect of D on Y and also about the differences between distinct groups. This is our estimated $E[Y|D, X]$. However, getting a sense of the significance of the interactive effects is hard, if not impossible, because we are interested in whether the slopes of the relationship between D and Y are statistically different at different levels of the moderator (Clark and Golder, 2023).

D-CAME. To estimate CAMEs for different groups, one needs to take the average of the slope of the relationship between Y and D, for a given group $X = x$. From the predicted value in Equation 24:

$$\hat{\theta}_x^{CAME} = \frac{1}{n_x}\sum_{i=1}^{n_x}\frac{\partial\hat{Y}(d_i,x)}{\partial d} = \frac{1}{n_x}\sum_{i=1}^{n_x}\hat{g}'_x(d_i) \tag{25}$$

being $g'_x(.)$ the partial derivative of $g_x(.)$ with respect to its argument, for the group $X = x$. Then, the estimated D-CAME is the difference of the estimated CAMEs of both groups, that is, $\hat{\theta}^{D-CAME} = \hat{\theta}_1^{CAME} - \hat{\theta}_0^{CAME}$.

AIPE. When aiming the estimate the AIPE, it has been argued that besides predicted values, researchers should also show the estimated difference of slopes between groups as a function of D,, that is, the *interactive partial effects,* $IPE(d_i, 1, 0)$.

$$\hat{\theta}_i^{IPE}(d_i,1,0) = \frac{\partial\hat{Y}(d_i,1)}{\partial d} - \frac{\partial\hat{Y}(d_i,0)}{\partial d} = \hat{g}'_1(d_i) - \hat{g}'_0(d_i) \tag{26}$$

This provides researchers with two crucial insights: i) which values of D drive the interactive effects; ii) whether the interactive effects go in the expected direction for all values of D. Consider the following example. We want to analyze how the age of an individual (D) is related to their political ideology (Y) depending on whether the individual is member of a trade union ($X = 1$) or not ($X = 0$), for individuals being 25 years old ($d = 25$). In this instance, $\hat{g}'_1(25) - \hat{g}'_0(25)$ indicates how becoming older is associated with ideology depending on trade union membership, when individuals are 25 years old. That is we are comparing individuals with the same age. It should be noted that in case of assuming linear treatment effects (i.e., $g(d) = d$), then in expectation $\hat{\theta}^{IPE} = \hat{\theta}^{D-CAME}$, for all values of D.

The final step is to estimate the *average interactive partial effects*, $AIPE(x, x')$. This should be done taking the sample average of the estimated IPE:

$$\hat{\theta}^{AIPE}(x,x') = \frac{1}{n}\sum_{i}^{n}\hat{\theta}_i^{IPE}(d_i,x,x') \tag{27}$$

Following the example above about age and ideology, the estimation of $\hat{\theta}^{IPE}$ is repeated for all observations in the sample, and then taken the mean sample value. For instance, this allows us to get whether, on average, there are differences in the relationship between age and ideology depending on trade union membership.

Appendix D Cross-Validation for Basis-Expansion

D.1 Why conventional CV may be problematic for estimating AIPE?

Model selection in the proposed approach requires choosing the functional forms $g_0(\cdot)$ and $g_1(\cdot)$ that best approximate the conditional expectation functions $\mathbb{E}[Y \mid D = d, X = x]$ for each group $x \in \{0, 1\}$. Standard 10-fold cross-validation minimizes the average squared prediction error within each group,

$$\mathrm{CV}(g_x) \;=\; \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_{x,k}}\sum_{i\in\mathcal{S}_{x,k}}\left(Y_i - \hat{g}_x^{(-k)}(D_i)\right)^2 \tag{28}$$

where $K = 10$ denotes the number of folds, $\mathcal{S}_{x,k}$ is the set of observations in group x assigned to fold k, $n_{x,k}$ is the number of such observations, and $\hat{g}_x^{(-k)}$ denotes the fit trained on all folds except fold k.

While this is appropriate for estimating D-CAME, this may not be the case for AIPE. Recall that the AIPE integrates the interactive partial effect (IPE) against the marginal density of D in the full sample. When these group-specific densities differ from the pooled density—for instance, when group 1 has many more observations in one region of the dose support than group 0—the standard CV loss assigns importance to regions in proportion to each group's sample size there, rather than in proportion to the pooled sample.

To see why this matters, suppose that $f_{D|0}$ is concentrated on $[0, 1]$ while $f_{D|1}$ spreads more evenly over $[0, 3]$ (even the range is the same). Standard CV for g_0 will invest nearly all of its error budget fitting $[0, 1]$ well, since that is where most of the $X = 0$ data lie. But the AIPE averages the derivative difference over the *pooled* density, thus, placing a non-trivial importance on $[1, 3]$. A specification that fits g_0 nicely on $[0, 1]$ but extrapolates poorly beyond that range may nonetheless win the standard CV race, even though its contribution to the AIPE at higher dose values is badly estimated.

Diversely, for the AIPE, we instead want this cost to be a function of the *total* number of observations—pooled across groups—at each dose level, because that is the distribution over which the estimand integrates.

D.2 Targeted CV weights for AIPE

To align the CV criterion with the AIPE estimand, the `R` function re-weights the squared residuals so that each group's loss function reflects the pooled density rather than its own. Specifically, we assign each observation i in group x the importance weight

$$w_i \;=\; \frac{\hat{f}_D(D_i)}{\hat{f}_{D|X=X_i}(D_i)},$$

where $\hat{f}_D$ is a kernel density estimate of the pooled dose distribution and $\hat{f}_{D|X=x}$ is the corresponding estimate within group x. Then, the targeted CV criterion then becomes

$$\mathrm{CV}_{\mathrm{targeted}}(g_x) \;=\; \frac{1}{\sum_{i:\,X_i=x} w_i} \sum_{i:\,X_i=x} w_i \left(Y_i - \hat{g}_x^{(-i)}(D_i)\right)^2.$$

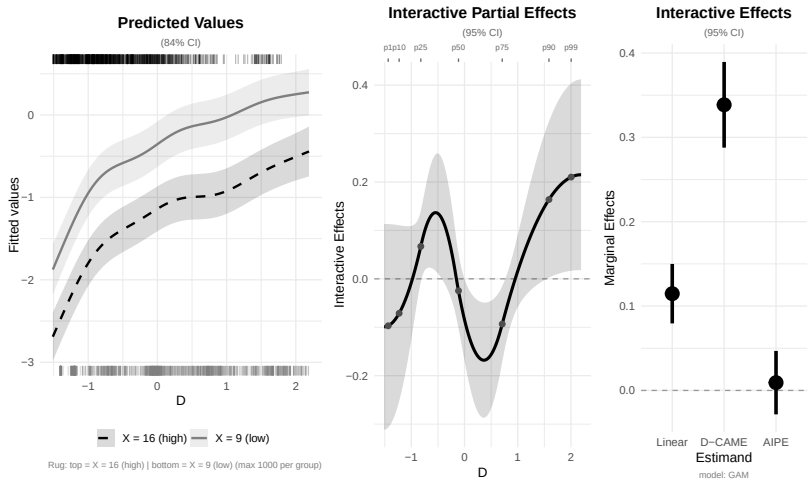
Which is the same as Equation 28 but assigning weights. Observations in group x that lie in dose regions where the pooled density exceeds the group-specific density receive higher weight and those in regions where the group is over-represented relative to the pool receive lower weight. The aim is to penalize misspecification in proportion to its contribution to the AIPE, rather than in proportion to within-group sample size.

Appendix E Application: REML rather than GCV

When discussing GAMs it has been stated that it is usually recommended to use REML for selecting the smoothing penalization parameter rather than GCV. This was the case because, when the number of observations is small, GCV may under-smooth the data. This was not the case in the application, as the number of observations was large, more than 150,000. Moreover, REML tends to be much more computationally demanding, making it almost prohibitive for 500 bootstraps.

Figure 8 shows the results using REML (with analytic standard errors, not cluster-bootstrapped standard errors) rather than GCV. In that sense, what matters is the comparison of point estimates which are essentially equivalent to the case using GCV.

Figure 8: Application using REML rather than GCV



Appendix F Discussion: Discrete Multi-valued Treatment

Rather than considering the case where D is continuous, this section expands the discussion to the case where D is measured on a ratio scale, meaning that differences between its values are substantively meaningful and comparable.

F.1 Unit-specific quantities: From Derivatives to First Differences

As has been stated in the main text, we considered two unit specific quantities: the slope of the relationship between D and Y (the CAPE), and the difference in the slopes between groups (the IPE).

When D is discrete—taking values in $\{d_0, d_1, \ldots, d_K\}$—derivatives, and hence slopes, are not defined. Under this assumption, the natural unit-specific quantities of interest are expressible as *first differences*. Then, the Conditional Average Partial Effect of moving from level d_k to d_{k+1} for group x is:

$$\text{CAPE}_x(d) = E[Y \mid D = d_{k+1}, X = x] - E[Y \mid D = d_k, X = x] \tag{29}$$

And the Interactive Partial Effect captures the differential impact of moving from level d_k to d_{k+1} for group 1, relative to group 0 is:

$$\text{IPE}(d,1,0) = \Big\{E[Y \mid D = d_{k+1}, X = 1] - E[Y \mid D = d_k, X = 1]\Big\} -$$

$$-\Big\{E[Y \mid D = d_{k+1}, X = 0] - E[Y \mid D = d_k, X = 0]\Big\} \tag{30}$$

F.2 The Weighting Problem

Even though the unit-specific quantities are clear and follows directly from the main text, there are more problems when deciding the population of interest, that is, in setting the weights.

The standard approach weights each transition from d_k to d_{k+1} by the probability of being at the starting level (d_k). For a one-unit *increase*:

$$\tau_x^{\text{CAME, inc.}} = \sum_{k=0}^{K-1} \text{CAPE}_x(d_k) \cdot p_x(d_k) \tag{31}$$

At the sample level, $p_x(d_k)$ is the share of observations in group $X = x$ with a value of $D = d_k$.

This answers the following type of question: "If you randomly select someone from group x and increase their D by one unit, what is the expected effect?"

Crucially, however, this makes that observations with the highest observed value is not considered to compute the weights, and accordingly, $\sum_{k=0}^{K-1} p_{D|x}(d_k) < 1$.

Now consider framing the same question as a one-unit *decrease*. The effect of moving from d_{k+1} to d_k is $-\text{CAPE}_x(d_k)$, and the natural weight is $p_{D|x}(d_{k+1})$, that is, the share of observations at the starting level of the decrease. Taking the negative to express it as the effect of an increase:

$$\tau_x^{\text{CAME, dec.}} = \sum_{k=0}^{K-1} \text{CAPE}_x(d_k) \cdot p_{D|x}(d_{k+1}) \tag{32}$$

Therefore, in general, $\tau_x^{\text{CAME, inc.}} \neq \tau_x^{\text{CAME, dec.}}$. The underlying unit-specific quantity—the effect of a transition between d_k and d_{k+1}—is the same regardless of direction. Only the weights change. This means that the CAME, and consequently the D-CAME and AIPE, depend on an arbitrary framing choice (increase vs. decrease) that has no theoretical justification.

How to address this problem is beyond the scope of the paper. Researchers could try to take the average of an increase and a decrease in d; however, they should acknowledge that even in

this case, the weights do not sum to 1.

References

Clark, William Roberts and Golder, Matt (2023). *Interaction Models: Specification and Interpretation*. Cambridge University Press.

Hainmueller, Jens, Mummolo, Jonathan, and Xu, Yiqing (2019). "How much should we trust estimates from multiplicative interaction models? Simple tools to improve empirical practice". *Political Analysis* 27.2, pp. 163–192.

King, Gary and Zeng, Langche (2006). "The dangers of extreme counterfactuals". *Political analysis* 14.2, pp. 131–159.

Liu, Jiehan, Liu, Ziyi, and Xu, Yiqing (2025). "A Practical Guide to Estimating Conditional Marginal Effects: Modern Approaches". *arXiv preprint arXiv:2504.01355*.

Long, J Scott and Mustillo, Sarah A (2021). "Using predictions and marginal effects to compare groups in regression models for binary outcomes". *Sociological Methods & Research* 50.3, pp. 1284–1320.