sociological science

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

Josep Serrano-Serrat

Spanish National Research Council (CSIC)

Abstract: Although quantitative social sciences often rely on estimating models in which treatment effects vary across groups, researchers rarely specify which causal quantity they aim to estimate or justify their empirical modeling choices. This article makes two contributions. First, it clarifies the distinct quantities of interest when studying interactions: comparisons at different treatment intensities (the difference in conditional average marginal effects) and comparisons at similar intensities (what I term the average interactive partial effect). When treatment effects are nonlinear and treatment distributions differ across groups, these quantities diverge. Second, the article assesses estimation strategies to estimate these quantities. It demonstrates that linear interaction models produce biased estimates of either quantity when treatment effects are nonlinear and explores two alternatives that explicitly accommodate such nonlinearities. This article is accompanied by an R package that implements these approaches. Through simulations, stylized examples, and an empirical application, the article shows that explicitly defining the quantity of interest and selecting appropriate models is essential for valid interaction analysis.

Keywords: interactions; nonlinear treatment; theoretical estimand

Reproducibility Package: All code necessary to replicate this study is available in an OSF repository at https://osf.io/8tx6g/

Many relationships studied in the social sciences are conditional in nature. For example, unemployed individuals generally face a higher risk of suicide than those who are employed, especially when local unemployment rates are low (Lee and Pescosolido, 2024); outdoor recreation areas increase children's well-being, particularly for individuals with low socioeconomic status (SES) (Rubio-Cabañez, 2025); and firms with more Black employees tend to have lower managerial quality, especially in socially conservative regions (Zhang, 2023). However, although researchers routinely test whether treatment effects vary across groups, they rarely specify which causal quantity they aim to estimate.[1]

Consider the *simulated* example shown in Figure 1, which illustrates the relationship between health problems and extreme temperatures (e.g., operationalized as the share of days with extreme temperatures in the past three months), distinguishing between individuals with low and high SES.[2] The histograms at the bottom of the graph display the density of extreme temperatures for different groups. For many researchers, this relationship entails no "interactive effects," as the effect of extreme temperatures on health does not appear to differ between groups when comparing at similar values of treatment intensity (Ganzach, 1997; Simonsohn, 2024). Other accounts suggest otherwise and claim that there are negative "interactive effects"—that is, the effects are smaller for high SES individuals than for low

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

Received: March 11, 2026

Accepted: June 23, 2026

Published: August 6, 2026

Editor(s): Arnout van de Rijt, Kristian B. Karlson

DOI: 10.15195/v13.a36

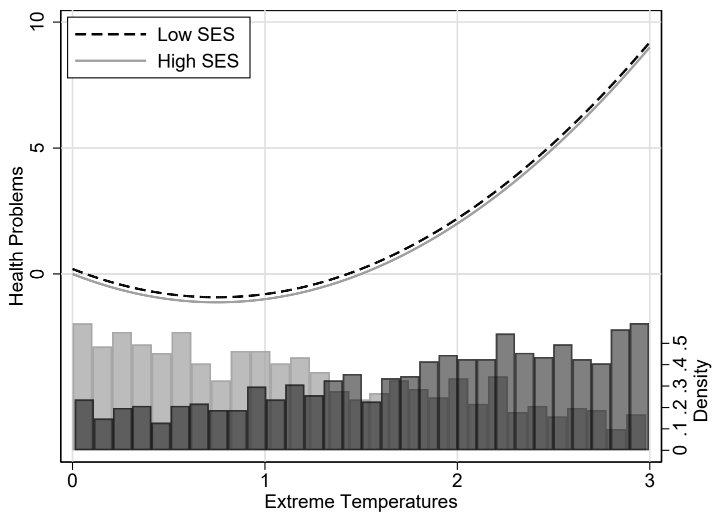


Figure 1: Simulated data: effect of extreme temperatures on health problems by socioeconomic status (SES). *Note:* Lines represent the relationship between health problems and extreme temperatures according to the data generating process (DGP) described in online supplement B. The bottom part of the figure displays the histogram of extreme temperatures for a realization of the DGP, in black for low SES and in gray for high SES.

SES individuals. They argue that this is the case because low-SES individuals face greater baseline exposure to extreme temperatures, which themselves carry larger health consequences (Hainmueller, Liu, et al., 2025; Liu et al., 2025).

Adding to this debate, the first aim of the article is to discuss different quantities of interest when considering interactions and how they relate to each other. In this context, it is argued that researchers may be interested in two types of questions (i.e., quantities of interest or theoretical estimands). For the example at hand, the first type of interactive question one may ask is: *For the distribution of extreme temperature exposure as it exists, does a change in extreme temperatures affect low- and high-SES individuals differently?* This involves comparing changes in extreme temperatures between groups at different levels of exposure. This is the relevant quantity of interest if one is concerned about how a generalized increase in temperatures would affect different socioeconomic groups, given that they are already exposed to a certain amount of extreme temperatures. In this case, I will show that the theoretical estimand is the difference in conditional average marginal effects (difference in CAMEs [D-CAME]). This compares how a unit increase in the treatment would, on average, affect different groups, given that they are already exposed to a specific level of treatment intensity (Liu et al., 2025; Long and Mustillo, 2021; Zhirnov et al., 2023).

The second type of question is: *For similar exposure levels, does a change in extreme temperatures affect low- and high-SES individuals differently?* This involves comparing between-group differences in the effects of extreme temperature on health for individuals at similar extreme temperature levels. Often, when exploring mechanisms, this is the quantity of interest. For example, if a researcher claims that high-SES individuals are more able to shelter from extreme temperatures than low-SES individuals, then he/she should compare individuals at similar levels of extreme temperatures. This second type of question has been studied less extensively. The theoretical estimand is the average interactive partial effect (AIPE), which measures whether, on average, treatment effects differ between groups when making comparisons at similar levels of the treatment. Discussing these different theoretical estimands, *Estimands when Testing Interactions: Definitions and Comparison* section shows that when the treatment and the moderator are not independent and treatment effects are nonlinear, D-CAME is different from AIPE.

The second aim of the article is to provide a better grounding for estimating different quantities of interest. When investigating how the effect of one variable changes across levels of another, researchers often rely on linear interaction models (Brambor et al., 2006; Hargens, 2009). The over-reliance on linear treatment effects comes from the combination of two beliefs: (1) interactive linear models allows researchers to estimate D-CAME; and (2) that this is the quantity of interest when testing interactions. Against this view, this article shows that models imposing linear treatment effects may not recover D-CAME (Yitzhaki, 1996), and that even when they can, in many cases they cannot estimate the AIPE, which may be the quantity of interest. By means of examples, simulation evidence, and an empirical illustration, the article explores how problematic may be assuming linear treatment effects.

To overcome problems associated with assuming linear treatment effects, the article explores how to estimate and display results of interactions when there are nonlinear treatment effects (Long and Mustillo, 2021; Mize, 2019).[3] This is done sequentially in two steps. The first step explains how to estimate and display the quantities of interest in the case of knowing the functional form of the relationship of interest. The second one explores different approaches to estimate the nonlinear relationship of interest from data. Although much of the literature tends to add quadratic or cubic terms, the article proposes two approaches, that are, arguably, more flexible: (1) to use a basis-expansion approach that selects the functional form from a large set of polynomial and spline functions; and (2) to use generalized additive models (GAMs) (Lundberg et al., 2021; Simonsohn, 2024; Wood, 2017). To this end, the article provides an `R` package that performs the proposed approach following the "best practices" for displaying nonlinear treatment effects as proposed by Mize (2019). This package allows researchers to estimate different quantities of interest, and to compare them with the linear model.

The discussion about different theoretical estimands when testing interactions reflects recent methodological debates across social science disciplines regarding what we aim to estimate when exploring interactions. Psychology researchers argue that interactive effects should compare treatment effects at similar treatment intensities (Ganzach, 1997; Simonsohn, 2024), criticizing political science approaches that assume linear treatment effects. Political scientists respond that Simonsohn (2024) does not properly define an estimand, so one cannot assess whether the results are biased. Indeed, they claim that the relevant estimand is the D-CAME, which can be estimated assuming linear treatment effects (Hainmueller, Liu, et al., 2025; Liu et al., 2025). This article defines the AIPE and analytically compares it with D-CAME, showing that both are meaningful but distinct quantities of interest. Moreover, it discusses the problem of assuming linear treatment effects when estimating either estimand.

The article also contributes to much of the existing literature which has focused on related questions in the context of limited dependent variables, typically using nonlinear models such as logit or probit (Long and Mustillo, 2021; Mize, 2019). As previous authors have noted, making group comparisons in such models introduce several challenges, including difficulties in interpreting coefficients, visualizing nonlinear treatment effects, and accounting for how all control variables interact between them (Ai and Norton, 2003; Long and Mustillo, 2021; Mize, 2019; Rainey, 2016). Because this article aims to focus exclusively on the role played by nonlinear treatment effects, it does so in a linear model framework. However, because nonlinear models inherently impose nonlinear treatment effects, the main insights of this article about defining a theoretical and empirical estimand for interactions can be also applied to those models.

Estimands When Testing Interactions: Definitions and Comparison

Quantities of interest with interactions

This section aims to discuss which are the estimands that are of theoretical interest for applied researchers in the context of studying effect heterogeneity (Keele and Stevenson, 2021; Lundberg et al., 2021). Even though the theoretical estimand is the "object of inquiry" in quantitative articles (Lundberg et al., 2021), it has not received the proper attention by applied researchers using interactions.

Theoretical estimands are made up of two components: the unit-specific quantity and the target population. The former element is the specific quantity/effect for each observation. Being $Y_i(d)$ the potential outcome of individual i for a given value of the treatment d,[4] the unit-specific quantity when studying causal effects is the change in the potential outcome for a marginal change in the treatment, that is, $\frac{\partial Y_i(d)}{\partial d}$ (Hirano and Imbens, 2004). If the quantity of interest is the average marginal effect (AME), then the unit-specific quantity is averaged over all the population of interest (target population). Conditional AME, on the other hand, has the same unit-specific quantity but averages over a specific group. Specifically, being N_x the number of observations belonging to group $X = x$ in the population:

$$\underbrace{\tau_x^{CAME}}_{\substack{\text{Theoretical}\\ \text{estimand}}} = \underbrace{\frac{1}{N_x}\sum_{i=1}^{N_x}}_{\text{target pop.}} \underbrace{\frac{\partial Y_i(d)}{\partial d}}_{\substack{\text{unit-spec.}\\ \text{quantity}}}$$

Then, in the literature, heterogeneous effects are usually defined as the difference in CAMEs, setting the target population to different values of the moderator. However, although usually employed in the literature, this may not be the quantity of interest for researchers. In particular, researchers should acknowledge that D-CAME confounds differences in treatment effects at different values of the moderator, and differences in treatment intensity. In Figure 1, this implies comparing the marginal effects of extreme temperatures for low-SES individuals, who were already highly exposed to extreme temperatures, to high-SES individuals who were less exposed to extreme temperatures.

To examine this point more analytically, it is useful to express the D-CAME as a function of conditional average partial effects (CAPEs), which represent the average marginal effect of D on Y at a given value of D and X (i.e., $CAPE_x(d)$). Or what is the same, the slope of the relationship of interests for a specific group, at a given value of D. For the example about extreme temperatures, $CAPE_{low}(2)$ is the average marginal effect of increasing extreme temperatures for individuals who have already experienced two days of extreme temperatures and are of low SES. Then, having defined CAPE, it can be shown that (see online supplement A.1):

$$\tau^{D-CAME} = \tau_1^{CAME} - \tau_0^{CAME} =$$

$$= \int_d \underbrace{\mathrm{CAPE}_1(d)}_{\substack{\text{effect of } D \text{ on } Y \\ \text{for group 1 at } D=d \\ \text{(i.e., the slope)}}} \overbrace{f_{D|1}(d)}^{\substack{\text{“share” in group 1} \\ \text{with } D=d}} \, \mathrm{d}d - \int_d CAPE_0(d) f_{D|0}(d) \, \mathrm{d}d, \tag{1}$$

where $f_{D|x}$ is the conditional density function of D for group $X = x$. This is just an application of the law of iterated expectations, and states that CAME can be defined as the average of the effect of D on Y for given values of D, weighted by their relative frequency in the group. In the extreme temperature example, the CAME is the average marginal effect of extreme temperatures at each level of exposure, weighted by the share of cases at each intensity of exposure. The main takeaway from Equation (1) is that D-CAME can be understood as the combination of the difference in marginal effects and the difference in the intensity of the treatment.

Diversely, researchers may want to do comparisons of the treatment effects at similar levels of the treatment. In this instance, the unit-specific component should be the difference in CAPE at different levels of the moderator. This is what I call interactive partial effects (IPEs), which is the difference in the treatment effects between groups, *at the same values of* D (i.e, $IPE(d)$). In the extreme temperature example, $IPE(2)$ is the difference in average marginal effects of extreme temperatures between low- and high-SES individuals for individuals who are already exposed to two days of extreme temperatures. In Figure 1, IPEs are zero for all values of D.

By averaging IPEs over the population of interest, one gets the AIPE as follows:

$$\tau^{AIPE} = \int_d \Big(\underbrace{CAPE_1(d) - CAPE_0(d)}_{\substack{IPE(d) \equiv \text{between-group differences} \\ \text{in the effect of } D \text{ on } Y \text{ at } D=d \\ \text{(i.e., differences in the slopes)}}} \Big) \overbrace{f(d)}^{\substack{\text{total} \\ \text{“share”} \\ \text{with } D=d}} \, \mathrm{d}d \tag{2}$$

with $f(d)$ being the density function of D. Intuitively, Equation (2) represents the average between-group differences in the slopes of the relationship between Y and D. That is, it answers questions such as: Does increasing extreme temperatures affect more low-SES than high-SES individuals when both are exposed to similar amounts of extreme temperatures?

At this point it should be noted that theoretical estimands cannot be directly estimated because they involve unobserved quantities, such as counterfactual outcomes for marginal changes in the treatment variable. Rather than observing the counterfactual value of the outcome variable for different doses of the treatment, $Y_i(d)$, we can only observe $Y_i(d_i)$. Online supplement A.3 discusses the identification assumptions to make the connection between these two quantities, and describes the associated empirical estimands—the target of empirical analysis—accordingly (Lundberg et al., 2021).

When D-CAME = AIPE?

To explore when the two theoretical estimands may be equal, it should be noted that the relationship between D-CAME and AIPE can be made explicit through a simple decomposition. Starting from Equation (1), each group-specific density can be written as the sum of the marginal density and a group-specific deviation, that is, $f_{D|x}(d) = f(d) + \left[f_{D|x}(d) - f(d)\right]$. Substituting into the D-CAME expression and rearranging yields

$$\tau^{D\text{-}CAME} = \underbrace{\int \Big[\mathrm{CAPE}_1(d) - \mathrm{CAPE}_0(d)\Big] f(d)\,\mathrm{d}d}_{\tau^{AIPE}} + \underbrace{\Delta}_{\substack{\text{distributional}\\ \text{discrepancy}}} \tag{3}$$

where

$$\Delta = \int \mathrm{CAPE}_1(d)\left[f_{D|1}(d) - f(d)\right]\mathrm{d}d - \int \mathrm{CAPE}_0(d)\left[f_{D|0}(d) - f(d)\right]\mathrm{d}d. \tag{4}$$

Equation (3) shows that the two estimands coincide if and only if $\Delta = 0$. The residual term Δ captures the *distributional differences in treatment intensity across groups*. Generally, Δ will be equal to zero whenever one of the following two conditions are met:[5]

1. **The treatment intensity between groups does not differ (i.e., $f_{D|1}(d) = f_{D|0}(d) = f(d)$).** When the two groups share the same treatment-intensity distribution, $f_{D|x}(d) - f(d) = 0$, making both elements in Equation (4) equal to zero. Intuitively, when the distribution of D is independent of X, confounding between the difference in marginal effects and the difference in the treatment distribution is no longer a concern. Moreover, in this case, X is not a confounder of the relationship between Y and D.

2. **The treatment effects for both groups are linear (i.e., $\mathrm{CAPE}_x(d) = \beta_x$).** In this scenario, each term in Equation (4) becomes $\beta_x \int \left[f_{D|x}(d) - f(d)\right]\,\mathrm{d}d$, which is equal to zero because $f_{D|x}$ and f are density functions that integrate to one. Intuitively, because the difference between D-CAME and AIPE is that the former does not compare the marginal effect of D at similar values of D, this distinction is not relevant when treatment effects are linear, because they are the same for all values of D.

Table 1: Relationship between the D-CAME and AIPE.

		D and X	
		Independent	Nonindependent
Treatment Effects	Linear	$\Delta CAME = AIPE$	$\Delta CAME = AIPE$
	Nonlinear	$\Delta CAME = AIPE$	$\mathbf{\Delta CAME \neq AIPE}$

Table 1 summarizes the discussion, indicating that both estimands will be different when there are nonlinear treatment effects and the distribution between D is different for different groups.

The Perils of Assuming Linear Effects

So far, most of the discussion has been about theoretical estimands. Indeed, there has not yet been a direct connection to any empirical model or specification. This section explores the risks of using a model that assumes linear treatment effects to estimate the quantities of interest defined in the previous section (in particular the sample analogs of these quantities; see online supplement A.3).

Set Up Model with Linear Treatment Effects

Being i the unit of analysis (e.g., individuals or countries), the standard model for testing interactions takes the following form (e.g., Brambor et al., 2006):

$$Y_i = \gamma + \alpha D_i + \eta X_i + \beta^L (D_i \times X_i) + \delta' \mathbf{Z}_i + e_i, \tag{5}$$

where $\mathbf{Z}$ is a set of controls and e is the error term. In this setting, D is continuous while X is binary (0 or 1), indicating whether the observation belongs to group 0 or to group 1. γ, α, η, β^L, and δ is a vector of unknown parameters, estimated by Ordinary Least Squares (OLS).

In this setting, scholars consider α and β^L their parameters of interest. They would take the partial derivative of Equation (5) with respect to D and interpret $\alpha + \beta^L x$ as the effects of D for different groups. Then β^L is interpreted as the parameter capturing the interactive effect, which is the difference in the effects between group 1 ($\alpha + \beta^L$) and group 0 (α).

Regarding controls, in the types of models considered—both linear and nonlinear—they enter additively, while still allowing interaction with the moderator (Blackwell and Olson, 2021). Specifically, these specifications assume that covariates Z are additively separable from D, which may be a restrictive assumption. When D and Z interact in the true DGP, the additive model is misspecified, and the OLS estimate of the D slope is a variance-weighted average of covariate-specific slopes, with weights proportional to the variance of D at different values of Z.[6] Consequently, misspecification may be particularly consequential when the variance of D varies substantially across covariates. The *Discussion and Conclusion* section elaborates on this.

Assuming Linear Treatment Effects: What We Estimate and What We Do Not

Although it is common in the literature to interpret β^L as the "interactive effect," there is usually no proper assessment of the substantive interpretation of the coefficient.

When the treatment effects are linear for both groups, β^L is an unbiased estimate of the D-CAME and, therefore, following the discussion in the *Estimands when Testing Interactions: Definitions and Comparison* section, of AIPE. However, this is rather a strong empirical assumption. In fact, in the more general case when the treatment effects are not linear at least for one of the groups, it is not entirely clear from a substantive point of view what is the interpretation of β^L (Yitzhaki, 1996). As shown by Yitzhaki (1996), OLS coefficients are a weighted sum of the slopes of the relationship between D and Y (i.e., CAPEs), weighting each observation i by the distance of D_i to the median value of D and the variation at its neighborhood values. Certainly, this does not speak to any of the theoretical estimands described in *Quantities of interest with interactions* section: while it has the same unit-specific quantity, by introducing these theoretically ungrounded weights, it is not clear what is the population of interest.

To illustrate the implication of this point, Figure 2 shows the distribution of coefficients from 1,000 Monte Carlo simulations generated by a Data Generating Process (DGP) with different slopes for different groups. Specifically, for group 1, the relationship between D and Y is a piecewise linear function, with a slope of -1 up to the 75th percentile of D, and a slope of 1 for the remaining 25th percentile. For group 0, Y is independent of D. D follows a log-normal distribution, is trimmed to the 95th percentile to avoid having extreme observations (King and Zeng, 2006; Liu et al., 2025), and is independent of the moderator, so D-CAME is equal to AIPE. In this DGP, both estimands are equal -0.5.[7]

The figure shows that the interaction coefficients from the linear model are not centered around -0.5. In fact, the average of the coefficients estimated is of the *opposite* sign (dashed vertical blue line). Therefore, researchers should be aware that assuming linear treatment effects may be problematic, as it does not allow them to estimate either the D-CAME or the AIPE.

Estimation of Interactions with Nonlinear Treatment Effects

Having stated that models assuming linear treatment effects may lead to biased estimates of D-CAME and AIPE, the next logical step is to explore nonlinearities. This section explores this in three parts. The first one sets up the type of model considered in this article. The second one discusses how to estimate the quantities of interest and display the results in case of knowing the functional form of the relationship of interest (Mize, 2019; Long and Mustillo, 2021). Finally, the last section discusses how to approximate the nonlinear relationship between D and Y, proposing two different approaches.

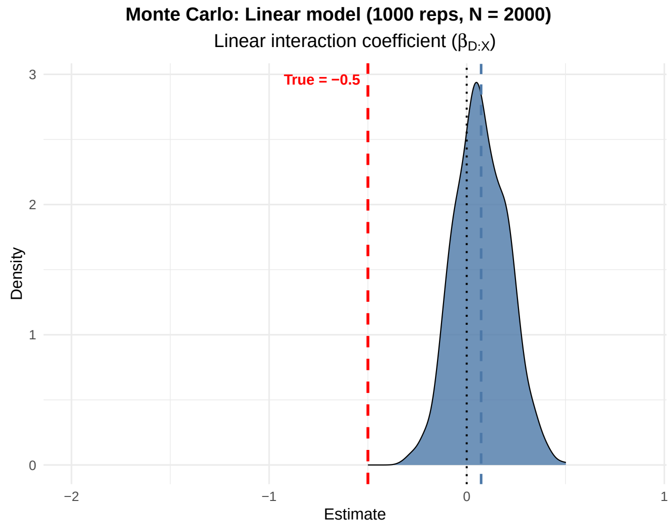


Figure 2: Simulation evidence assuming linear treatment effects.
Note: Coefficients' kernel density estimate of 1,000 Monte Carlo simulations assuming linear treatment effects. The red dashed line shows the true D-CAME and AIPE ($= -0.5$). The dotted black vertical line states for 0.

Set Up Model with NonLinear Treatment Effects

Rather than assuming a linear relationship between Y and D, the type of model that will be considered in this article takes the following form:

$$Y_i = \mu + \alpha X_i + g_0(D_i) + g_1(D_i) + \delta' \mathbf{Z}_i + e_i \tag{6}$$

With $g_x(.)$ being a potentially nonlinear function linking D and Y when $X = x$. Again, the controls enter additively into the model. The main empirical problem in Equation (6) is to determine the functional form of these functions. First, I consider the case in which the functional form is known, discussing how to estimate the quantities of interest with nonlinearities and explaining how to present the results. The next section explores different approaches to estimate g_x. It is assumed that g_x is continuously differentiable, which means that the case of continuous, but not discrete, variables is considered. The *Discussion and Conclusion* section elaborates on the similarities and differences with the case with discrete treatment variables.

Estimating Quantities of Interest Knowing g_x

Before discussing how to estimate g_x, this section examines how to estimate the quantities of interest when g_x is known. Although this is generally unfeasible, it provides a better understanding of how to estimate the quantities of interest when dealing with nonlinear treatment effects.

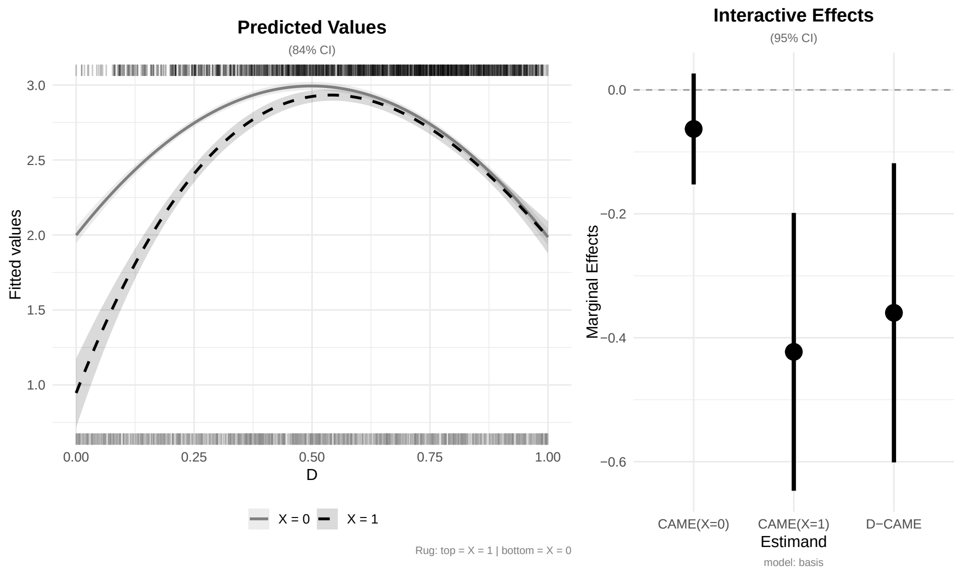


Figure 3: Proposed approach for showing results of estimated D-CAME.
Note: The left panel shows predicted values and 84 percent confidence intervals from robust standard errors. The right panel shows the estimate CAME for each group and their difference with 95 percent confidence intervals.

D-CAME

When regressing a specification with nonlinear treatment effects, researchers should present the information in the most intuitive way for readers. Indeed, in these instances, displaying the estimated parameters is not informative (Clark and Golder, 2023; Mize, 2019; Simonsohn, 2024), as there is no single parameter that summarizes the estimated quantities of interest (either D-CAME or AIPE).

To better understand the relationship between Y and D for different groups, researchers should present the predicted values of the dependent variable as a nonlinear function of D for each group. This should be complemented with a "rugged" histogram of D at the top of the panel for $X = 1$ and at the bottom for $X = 0$.[8] For confidence intervals, I recommend displaying them at the 84 percent level, as this generally allows for better between-group comparisons than intervals at the 95 percent level (Armstrong II and Poirier, 2025; Browne, 1979).[9]

The left panel in Figure 3 shows the estimated predicted values, after estimating a model knowing the functional form, with the DGP described in online supplement B. It can be seen that for both groups there is an inverse u-shaped relationship, and that, although predicted values of Y are lower for group 1 than for group 0, the differences monotonically decrease for higher values of D and eventually fade away.

Second, in another graph, researchers should report the estimated CAME for each group and their difference. This provides information about which specific groups have significant CAME and whether the difference between the groups is significant. From the estimated model capturing nonlinear treatment effects, researchers should obtain the estimated average slope of the relationship of interest for each group. In particular, the estimated CAME for a given group is the average of the estimated slopes for each group, weighting them by the share of observations at each value of D in that group. Accordingly, estimated D-CAME is the difference in estimated CAMEs (see online supplement C for a more analytical discussion).

The right panel of Figure 1 shows the estimated CAME for each group and the estimated D-CAME. For group 1, the estimated CAME assigns greater importance to high values of D compared to group 0, and at high values of D, the relationship between D and Y is more negative. The estimated CAME for group 0 is close to zero and not statistically significant, whereas the estimate for group 1 is negative and highly significant, making the difference between CAMEs also negative and significant.

AIPE

In case of being interested in the AIPE, the logic for estimating and presenting the results is similar to that of the D-CAME. The empirical estimand is again defined as the sample analog of the theoretical estimand. To estimate the AIPE, one must first estimate the *difference* in slopes between groups for each value of D, which are the estimated IPEs. By taking the weighted average of these differences, one then estimates the AIPE (see online supplement C for a more detailed analytical discussion).

Regarding how to present the results, the logic is also similar to the case for D-CAME. In this case, researchers should show three pieces of information (Mize, 2019). First, as before, they should show the estimated predicted values. The second relevant piece of information is the estimated IPEs, which show how different the treatment effects are between groups for given values of the treatment variable. Finally, the estimated AIPE, which is the weighted average of IPEs. By providing these three pieces of information, researchers can understand the key features of the interactive relationship between the treatment and the dependent variable.

Figure 4 shows the results of this exercise using the same example as in the previous section. The left panel displays the estimates of the predicted values, the middle panel shows the IPE, and the right panel compares the estimates of the interactive parameter of the linear model, as well as the estimated D-CAME and AIPE from the nonlinear specification.

As before, the left panel intuitively shows the relationship between Y and D. Additionally, the histogram indicates that the distribution of D differs markedly between groups, suggesting that D-CAME may differ from AIPE. The result displayed in the middle panel shows that the IPEs are positive at low values of D but approach zero and are not statistically significant at high values. The right panel shows that, although the linear model estimates a nonsignificant effect, the estimated D-CAME is negative (as shown previously). In this case, the coefficient of the linear model is lower than the estimated D-CAME because, for group 1, the linear model assigns more weight to low values of D, where variability is higher. Because the relationship is more positive at low values of D, the coefficient estimated from the linear model is also more positive. In contrast, the estimated AIPE is positive and significant because the differences between group 1 and group 0

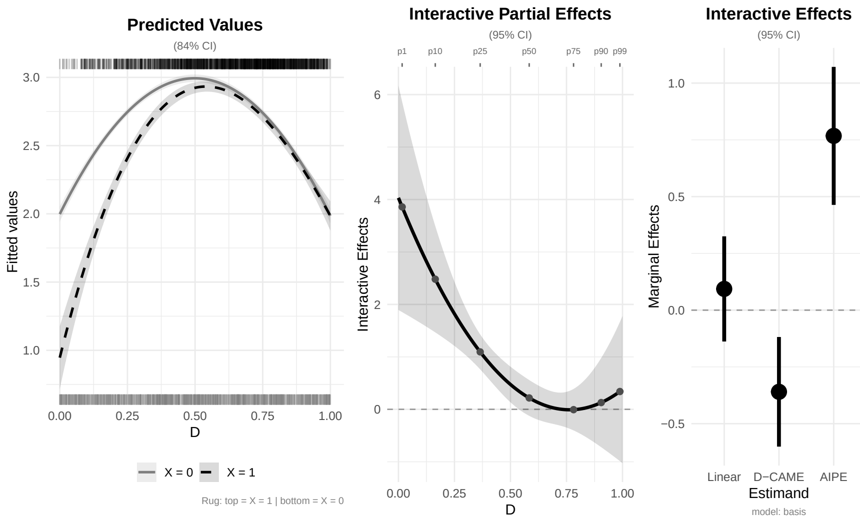


Figure 4: Proposed approach for presenting results of estimated AIPE.
Note: The left panel displays predicted values and the associated 84 percent confidence intervals. The middle panel shows estimated IPEs with associated 95 percent confidence intervals and the right panel presents the coefficient estimate from a linear model, as well as the estimated D-CAME and AIPE using the proposed approach.

Table 2: Pros and Cons of Different Empirical Estimands.

Empirical Estimands	Pros	Cons
Predicted values	Useful for visualizing the relationship between treatment and the dependent variable at different moderator levels.	Does not provide insight into whether slope differences are statistically significant.
Interactive partial effects (IPEs)	Identifies the specific treatment values driving the interactive effects and verifies whether they are present across all treatment levels.	Does not provide information on whether there are interactive effects on average, which depend on the distribution of the treatment variable.
Average interactive partial effects (AIPEs)	Provides an overall summary of whether there are interactive effects on average.	Similar AIPE values may arise from different underlying IPE patterns.

decrease monotonically with D. This example highlights the importance of deciding both which model to estimate and which is the quantity of interest.

Table 2 displays the pros and cons of showing the estimated values of different empirical estimands, making clear the usefulness of combining all of them.

Empirical Approaches to Estimate g_x: Basis Regression and GAM

This section elaborates on how to estimate g_x in models such as the one in Equation (6). Usually the literature employs quadratic or cubic effects of D at different values of X (Mize, 2019; Long and Mustillo, 2021). However, imposing a given polynomial relationship between D and Y may be too restrictive. Diversely, this section considers two distinct approaches that allow for flexible, ex-ante unknown functional form of the relationship of interest. The first one is a parametric approach that selects over a set of polynomial and splines functions. The second uses GAMs (Hastie and Tibshirani, 1986; Keele, 2008; Simonsohn, 2023). As has been stated, all these models assume that g_x is continuously differentiable. The *Discussion and Conclusion* section provides a brief discussion for the case with discrete multivalued treatment variables. These approaches are available as an R package (`dcameaipe`).[10]

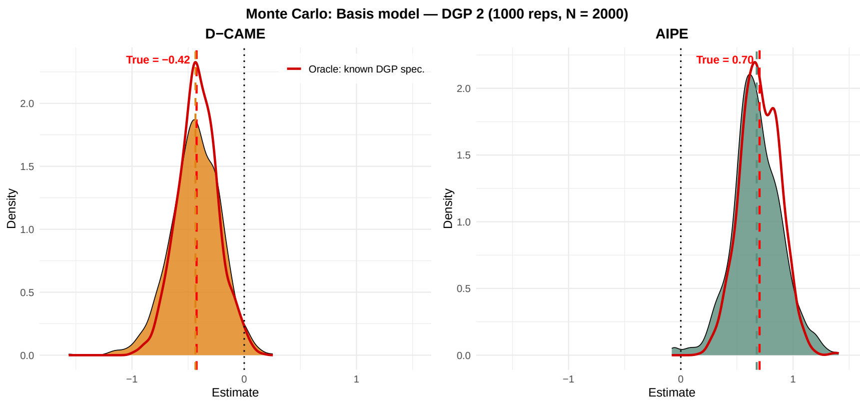
Basis regression with cross-validation

Estimation. This approach approximates the unknown relationship between Y and D selecting over a set of basis functions, specifically, polynomials and splines of varying complexity. The default model considers from simple polynomial functions (up to degree 5) to relatively complex spline functions (cubic splines with up to 6 knots, with knots position in empirical quintiles). In case that the authors aim to explore more flexible models, the `R` package provided allows researchers to decide among a wider set of functions. The final specification is selected by minimizing the cross-validated mean square error (CV-MSE) using 10-fold cross-validation (Lundberg et al., 2021).[11,12]

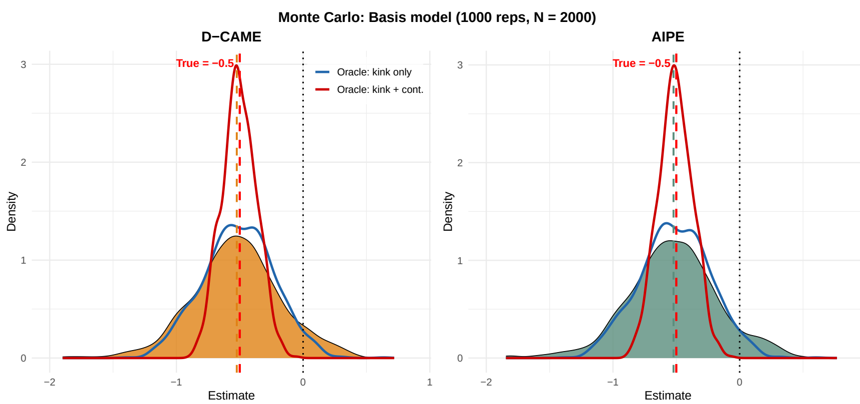
Although it selects the model from a prespecified discrete set of parametric functions, this approach considers a broad range of candidates and can approximate functional forms not directly included among the basis functions. Figure 5 shows the results of a Monte Carlo simulation analogous to those in the *Assuming Linear Treatment Effects: What We Estimate and What We Do Not* section. Panel (A) shows results for the DGP discussed in the *Estimating Quantities of Interest Knowing g_x* section, and Panel (B) shows results for the DGP in *Assuming Linear Treatment Effects: What We Estimate and What We Do Not* section. The red densities correspond to an oracle estimator that knows the true functional form, whereas the blue density in Panel (B) corresponds to a partially informed oracle that knows the relationship is linear on each side of the kink and knows the kink location, but does not know whether the function is continuous at the kink.

Panel (A) considers a DGP whose functional form is included in the basis expansion. The basis-regression approach performs almost as well as the oracle. This is evident because the two densities are nearly identical, indicating negligible bias and minimal efficiency loss. Panel (B) uses a DGP where the true relationship is piecewise linear with a kink at the 75th percentile of D. In this case, the true functional form is not among the basis functions. The estimator remains approximately unbiased for both the D-CAME and the AIPE, but the sampling distributions are wider than in Panel (A). The basis-expansion estimates are comparable in spread to the partially informed oracle (blue).[13]

Brief note on inference. A problem arises when deciding the model's functional form using the same data that will be used to test the specification, a problem known as pretesting. The issue is that this approach does not account for the uncertainty



(A) Functional form included in the basis functions.



(B) Functional form not included in the basis functions.

Figure 5: Simulation evidence basis regression approach.
Note: Results of 1,000 Monte Carlo simulations. (A) The results for the DGP discussed in the *Estimating Quantities of Interest Knowing* g_x section, and (B) the results for the DGP in the *Assuming Linear Treatment Effects: What We Estimate and What We Do Not* section.

introduced in this initial step. This would not be an issue if researchers had decided in advance which functional form to use (e.g., quadratic or cubic effects).

For the case at hand, I propose two approaches researchers can use to recover standard errors that account for the uncertainty from the first step. The first option is to use cross-fitting (Chernozhukov et al., 2017). Although this approach was primarily designed to address overfitting bias, it also serves to obtain correct standard errors (Ratkovic and Tingley, 2023). This approach involves splitting the sample into two random parts: one part is used to determine the functional form through cross-validation, and the other part is used to estimate the model. Because the data used for estimation are independent of the data used to select the functional form, the estimation step can be treated as if the functional form were fixed *a priori*, and standard inference on the estimated coefficients is valid without correction for model selection uncertainty. Naively using this approach would result in a significant loss of statistical power, as it uses only half of the sample. To regain almost full efficiency, cross-fitting repeats the process with the roles reversed and aggregates the results.

The second alternative is to use bootstrapping, using each bootstrap sample both to select the functional form and to estimate the model. By obtaining estimates from different bootstrap samples, this method accounts for the uncertainty generated in the first step. When observations are clustered, for instance, by country or school, researchers should compute block cluster bootstrap standard errors (Cameron et al., 2008). Although both approaches generally work well, bootstrapping tends to be more computationally demanding as the number of observations increases. Conversely, when the number of observations is low, cross-fitting may be problematic because splitting the sample can penalize more complex functional forms.

Continuous moderators. The model proposed can be extended to the case of having a continuous moderator. In this case, however, it may not be enough to simply add the (linear) interaction between the moderator and the function linking D to Y (Hainmueller, Mummolo, et al., 2019). Differently, I follow Hainmueller, Mummolo, et al. (2019) and propose to estimate a binning specification where treatment effects are computed at different terciles of the moderator. Considering nonlinearities in the treatment effects, the proposed specification takes the following form,

$$Y_i = \sum_{j=1}^{3} \left\{\mu_j + g_j(D_i) + \alpha_j X_i\right\} G_j + \delta' \mathbf{Z}_i + u. \tag{7}$$

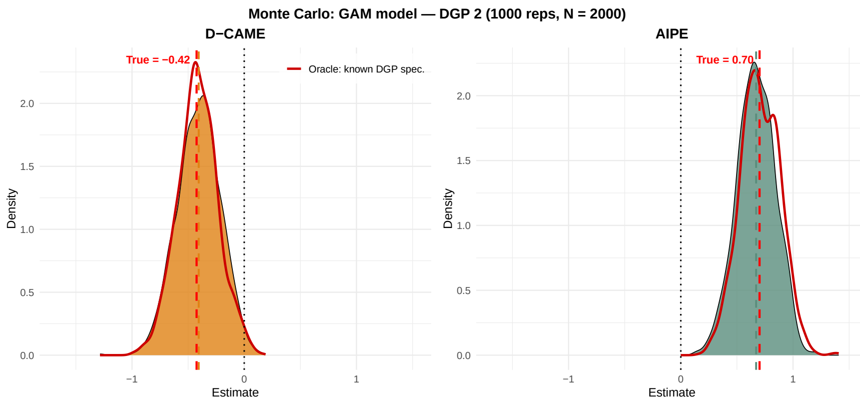
Here, G_j is a dummy variable that equals one if X falls within its jth tercile, and zero otherwise. With this specification, I recommend presenting the same key results, but evaluated at low and high values of the moderator (i.e., the lowest and highest terciles). Although terciles is usually enough, researchers can split the moderator in other ways.

GAM

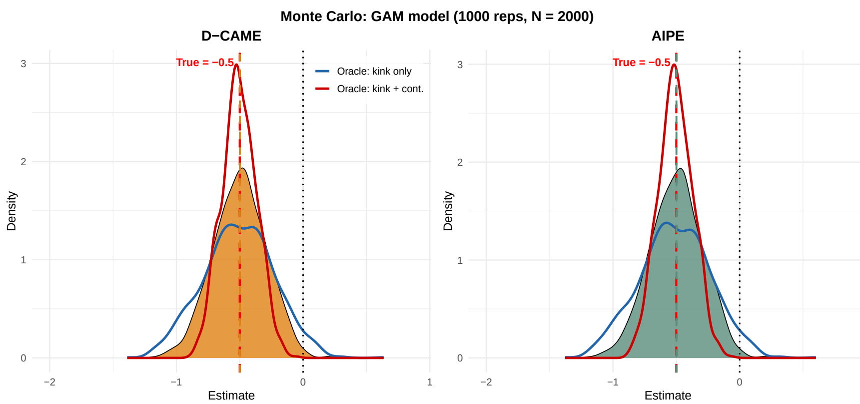
Estimation. Rather than selecting from a set of prespecified parametric functions, the second alternative estimates g_x using GAMs (Keele, 2008; Simonsohn, 2024).[14] GAMs impose no global parametric form on the relationship between D and Y, instead representing g_x as a smooth function learned directly from the data. In practice, g_x is modeled as a penalized regression spline, with a penalty parameter controlling the trade-off between fit and smoothness, which is selected by maximizing restricted maximum likelihood (REML) criterion (Wood, 2017). In this approach, the estimated function adapts to the data, becoming smoother when the signal is weak and more flexible where the relationship is clearly nonlinear.[15]

Figure 6 shows simulation evidence analogous to Figure 5, now estimating the D-CAME and the AIPE using GAMs.[16] In the first instance, where the DGP follows a polynomial function, GAMs perform well: the estimates are approximately unbiased, and the sampling distribution closely matches that of the infeasible oracle that knows the functional form in advance.

The second instance is more demanding. Because GAMs impose smoothness—assuming the relationship is continuous and differentiable—they are poorly suited



(A) Continuous and differentiable DGP.



(B) Continuous but not differentiable DGP.

Figure 6: Simulation evidence basis expansion approach.
Note: See Note in Figure 5. The DGPs considered in each panel are the same as in Figure 5; the change in name reflects a change in the logic of each approach.

to a DGP with a kink in the relationship of interest. Even so, the estimates remain approximately unbiased for both estimands. The sampling distributions are wider than in the fully informed oracle case, as expected, but narrower than those of the partially informed oracle that knows the relationship is piecewise linear and knows the kink location, yet does not know whether the function is continuous there.

Brief note on inference. A well-known limitation of GAMs is that the confidence intervals may be too narrow (Marra and Wood, 2012; Simonsohn, 2024; Wood, 2017). This occurs because these intervals are computed treating the selected smoothing parameter as known. Thus, it fails to account for the uncertainty in the smoothing selection step itself.

To address this, researchers should estimate bootstrapped confidence intervals for the parameters of interest. Specifically, each bootstrap sample is used to refit the GAM (including the degree of smoothness) and to recompute the estimand of interest. Bootstrap estimates capture the uncertainty arising from the smoothing selection step. When the data have a clustered structure, researchers should use the block cluster bootstrap (Cameron et al., 2008).

Continuous moderators. The GAM framework offers a natural extension to the case of a continuous moderator X. Rather than estimating separate smooth functions g_x for discrete groups, one can directly model the joint interaction for all values of D_i and X_i :

$$Y_i = \mu + g(D_i, X_i) + \delta' \mathbf{Z}_i + u_i, \tag{8}$$

where $g(D, X)$ is a bivariate smooth function that captures how the effect of D on Y varies continuously with X. This specification is more flexible and tends to have more power than using terciles of the moderator (Simonsohn, 2023).[17]

Although the bivariate smooth function can be evaluated at any point of X, estimating the D-CAME requires defining the subgroups to be compared. As discussed earlier, I recommend comparing the bottom and top sample terciles of X. Specifically, I propose evaluating predicted values at two representative values of the moderator: the median of X in the bottom tercile and the median of X in the top tercile (Hainmueller, Mummolo, et al., 2019). The CAME at each moderator value is then obtained by averaging the estimated slopes over the observed distribution of D among observations in the respective tercile. The D-CAME is the difference between these two CAMEs. Thus, although the interaction between D and X is fully flexible, slope aggregation must be performed over meaningfully defined subgroups.

To estimate the AIPE, researchers can use the median values of the extreme terciles of the moderator, estimate the IPEs, and take the weighted average of the IPEs over the full sample distribution of D.

A comparison of the two approaches

The previous sections have described two alternatives that researchers may use to estimate the D-CAME and the AIPE. This section compares them across several dimensions. Basis regression approximates the unknown function using a prespecified class of basis functions, such as polynomials or splines. Selecting the model by cross-validation, a discrete model-selection method, helps prevent overfitting. Although this approach is similar to the continuous penalization used by GAM, which employs a continuous penalized smoothing mechanism (estimated via REML or GCV), researchers must commit to a functional family in advance. This may be problematic when the relationship of interest takes nonconventional forms.

One of the main advantages of using basis regression is that the model is *linear* in its parameters and can therefore be estimated using OLS. Therefore, this approach is familiar to most quantitative social scientists and greatly facilitates integration with standard extensions such as instrumental variables, multilevel models, multiple imputation for missing data, and high-dimensional covariate adjustment via adaptive lasso. Although some of these advances are available for GAMs, each requires a specific implementation, often preventing the use of more innovative methods developed for OLS regressions.[18]

Besides being more flexible in terms of capturing nonlinear treatment effects, one of the main advantages of GAMs is that they more naturally incorporate the estimation of interactive effects with a continuous moderator (Simonsohn, 2024). Rather than having to discretize the moderator, GAMs are able to estimate a bivariate function linking D and X to Y.

Empirical Illustration: Robotization, Education and Anti-Immigration Attitudes

This section applies the proposed approach to analyze how anti-immigration attitudes and exposure to robotization are related, depending on individuals' education levels. To do so, I use the replication data from Anelli et al., 2021. Specifically, they use data from the first seven rounds of the European Social Survey (ESS), which is pooled cross-sectional data from 2002 to 2015. Data on anti-immigration attitudes come from self-reported scores on several questions addressing economic and cultural concerns about immigration, as well as whether respondents believe immigrants make their country a better place to live. The ESS includes self-reported information on years of education, which I use as a (continuous) moderator. The treatment variable, robotization exposure, is operationalized as the weighted sum of the routinization of tasks in each occupation multiplied by the change in the number of robots in that country-year, with the weights being the individual probability of working in each occupation (for a given age, gender, education, and region in a pretreatment period).[19] Moreover, I include a small set of pretreatment sociodemographic control variables, in particular, gender and age, as well as NUTS2 region and country-year fixed-effects.[20] I use a sample of working-age population from 18 to 65 years, having around 155,000 observations. All regressions use poststratification weights provided by the ESS.

Figure 7 shows the results of estimating a GAM, calculating block cluster standard errors (500 bootstrap samples), and selecting the smoothing parameter using GCV.[21] Examining the predicted values, robotization appears to monotonically increase anti-immigration preferences, and the relationship is similar for both the low educated and the high educated. Robotization seems to increase more anti-immigration attitudes at low levels of robotization, and it flattens at high levels. When analyzing the estimated IPEs, their sign varies across cases—positive in some and negative in others—tends to be close to zero, and is not significant in any case.

When analyzing the different estimands, the approaches differ significantly. The coefficient for the linear specification is 0.11, indicating that one standard deviation increase in robotization exposure raises anti-immigration concerns by 0.11 standard deviations more for highly educated individuals than for low-educated individuals. However, when accounting for nonlinearities, the estimated D-CAME is three times as large as the linear coefficient—with a coefficient of 0.33. This suggests that assuming linear treatment effects may not be appropriate if researchers are interested in estimating D-CAME. Finally, the estimated AIPE is close to zero and not statistically significant, indicating that at similar levels of robotization exposure, low- and high-educated individuals respond similarly. The D-CAME is larger than

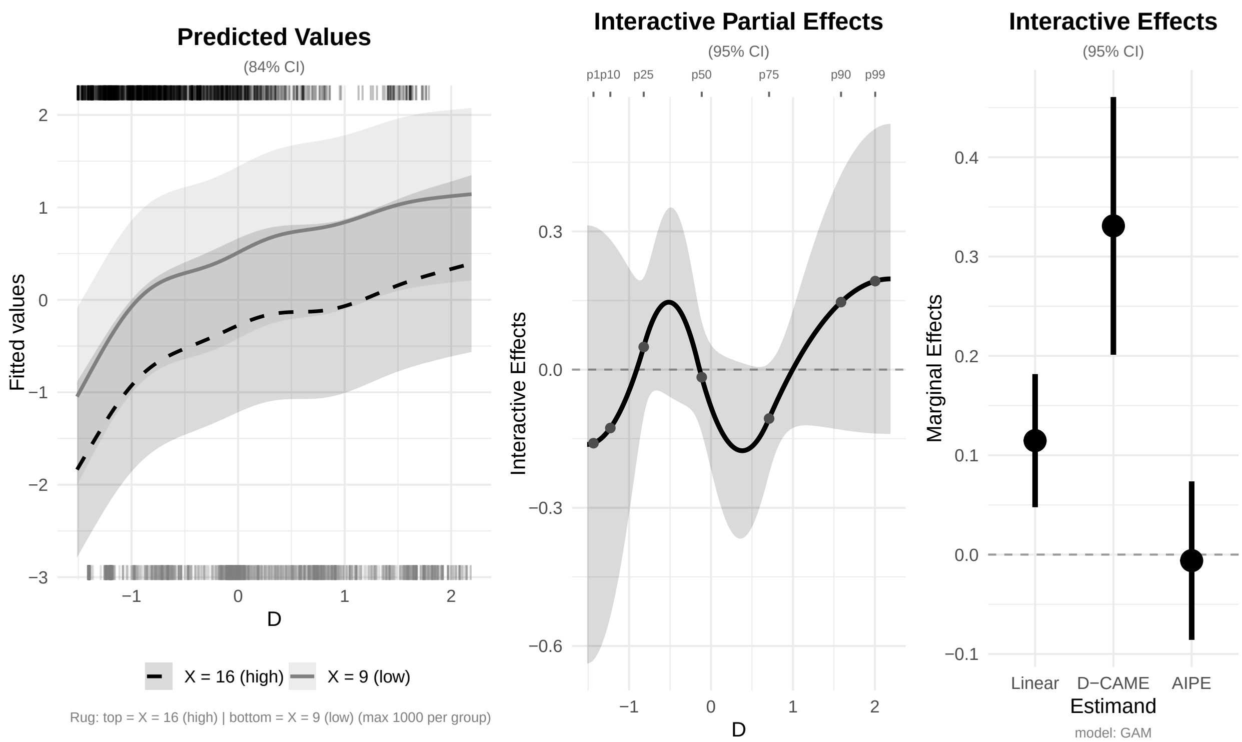


Figure 7: Application (GAM): robotization, education, and anti-immigration attitudes.
Note: GAM results using GCV to select the smoothing parameter. Data are from Anelli et al. (2021). Standard errors are computed from 500 cluster bootstrap samples. Clustering is at the region (NUTS2)-year level. $N = 154,646$.

the AIPE because there are decreasing marginal effects of robotization exposure, and highly educated individuals are less exposed to robotization.

Discussion and Conclusion

Testing how the relationship between two variables varies by different groups is common practice in sociology and other social sciences. However, it is often unclear what the quantity of interest is doing this type of analysis or how it should be estimated. Comparing *between* groups implies that these comparisons are typically descriptive, but the fact that they are descriptive does not mean they should not be linked to the theory one aims to test (De Kadt and Grzymala-Busse, 2025; Savage, 2020; Savage, 2024). In this sense, “good description” “must be grounded in social scientific theory,” which implies that researchers should be clear about what they are “attempting to describe, and the status of their questions and claims” (De Kadt and Grzymala-Busse, 2025, p. 6).

In these lines, the first contribution of the article is to discuss two types of questions that may arise when considering interactions. The first one, the D-CAME, concerns how increasing the treatment variable would affect different groups, given a specific distribution of treatment intensity. Alternatively, researchers may want to compare between-group differences in treatment effects at similar levels of treatment intensity. For this case, researchers are interested in what I have called the AIPE. It turns out that AIPE and D-CAME are different when there are nonlinear treatment effects and the treatment variable is differently distributed at different levels of the moderator.

The differences discussed in the article are not merely semantic, as failing to distinguish between the two estimands can lead to erroneous conclusions, cause researchers to dismiss valid theories, or generate not-that-puzzling puzzles. Consider the example in the introduction. One could estimate a regression and conclude that extreme temperatures have a stronger effect on low-SES individuals than on high-SES individuals. The researcher might then claim that these interactive effects occur because high-SES individuals can buffer themselves from the negative consequences of extreme temperatures. If that were the mechanism, then at similar levels of extreme temperatures, high-SES individuals would experience fewer negative effects. However, according to the DGP, this is not the case, because the effects of extreme temperatures on health are the same for both groups when comparing similar levels of treatment exposure.

The second part of the article explores how assuming linear treatment effects—which is common practice in the literature—may prevent researchers to estimate any of the two quantities of interest. As an alternative, the article proposes different approaches to deal with nonlinearities of the treatment effects, with an accompanying `R` package. By means of simulations it is shown that it outperforms the linear models, and that the same nonlinear specification may allow researchers to estimate D-CAME and AIPE. Through simulations, examples, and an empirical application, the importance of choosing the estimand of interest and specifying the empirical assumptions is demonstrated.

It should be noted that the two empirical approaches proposed in this article are only two of many ways researchers can address nonlinearities. Alternatively, they can use models such as local linear regressions or recent developments in machine learning to study heterogeneity with continuous treatment variables (Ratkovic and Tingley, 2023). Rather than providing an innovative approach to estimate nonlinear treatment effects, the key contribution of the article is to offer a tractable framework for estimating different quantities of interest when treatment effects vary across groups. A substantive limitation of the approach is that controls enter additively into the model, without including the interaction between D and Z. In this regard, the proposed approach could be improved by including $D \times Z$ in the model and averaging predictions over the empirical distribution of Z (Hanmer and Ozan Kalkan, 2013), which directly marginalizes over Z. In any case, if the interaction between D and Z is included, researchers should not keep Z fixed, as this would be more related to estimating a causal interaction—how treatment effects vary under an exogenous manipulation of Z—which is a distinct estimand from the effect heterogeneity quantities considered here (Bansak, 2021; VanderWeele, 2009).

Interestingly, the proposed discussion of nonlinear treatment effects can be extended to cases with multivalued discrete treatment variables measured on a ratio scale. However, researchers should exercise caution, particularly when determining the population of interest. In this context, rather than slopes, the unit-specific quantity is the first difference (i.e., a unit change in the treatment variable), but the logic remains similar. When defining the population of interest, additional challenges arise. For example, suppose that D has three values, $D = \{1, 2, 3\}$, which could capture the number of days in the last week with extreme temperatures. If a researcher wants to analyze how a unit increase in D affects health problems, there are only two cases to consider: an increase from 1 to 2 days and from 2 to 3 days. Because an increase from 3 to 4 days is not observed, the observations with $D_i = 3$ receive zero weight. This has two consequences: the weights no longer sum to one, and a unit increase in D generally yields different estimated D-CAME and AIPE than a unit decrease. Although a full discussion of these issues is beyond the scope of this article, researchers should be aware of them when applying the proposed approach with this type of treatment variable (see online supplement F for a more thorough discussion about this issue).

Moreover, some of the logic of this article also applies to the case of having a discrete treatment variable and a continuous moderator (Hainmueller, Mummolo, et al., 2019; Liu et al., 2025). In this case, the only quantity of interest is D-CAME, which by definition implies that comparisons are made at similar values of the moderator. Thus, nonlinearities should be taken seriously. Although the literature on nonlinear moderators shows only the estimated CAME function, this article also recommends reporting whether, on average, treatment effects are more positive or negative at higher values of the moderator.

Building on the discussion in this article, there are several guidelines that researchers analyzing interaction effects may benefit from:

1. *Clearly define the quantity of interest.* Driven by theoretical considerations, researchers should define which is the research question. In the introduction's example, researchers may be interested in how a generalized increase in extreme temperatures affect on average different groups, given that their given exposure to extreme temperatures is already different. Being this the case, researchers' quantity of interest is D-CAME. Diversely, if they want to estimate how the effect of extreme temperatures differ between groups when they are exposed at similar levels of treatment intensity, then their quantity of interest is the AIPE. These two quantities may be different when there are nonlinear treatment effects, and the distribution of the treatment differs between groups.

2. *When estimating interactions, take nonlinear treatment effects seriously.* Sometimes the literature considers that their goal is to estimate D-CAME, which can be estimated by models assuming linear treatment effects. However, this article shows even if the D-CAME is the quantity of interest, the linear model may not be able to estimate it when the treatment effects are nonlinear. Therefore, researchers should examine the role of nonlinearities. Two alternatives have been proposed, but researchers could also use recent developments in social sciences literature dealing with nonlinearities.

3. *When estimating nonlinear treatment effects, provide different relevant pieces of information.* After deciding on the theoretical estimand and the model to be estimated, researchers should present complementary information that allows understanding the functional form of the relationship of interest, as well as

the nature of interactive effects. Each of these pieces of information has its own advantages and limitations; therefore, they should be combined (e.g., see Table 2).

Declaration of Competing Interests. The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Notes

1 Recent work highlights differences between effect modification—how treatment effects differ *between* groups—and causal interactive effects—how treatment effects change when the moderator is exogenously changed (Giesselmann and Schmidt-Catran, 2022; Lundberg et al., 2021; Keele and Stevenson, 2021). Although this distinction clarifies the type of comparison made regarding the moderator, it remains largely silent about comparisons across different levels of treatment intensity, which is the aim of the article.

2 The package to replicate all figures and results of this article is available at https://osf.io/8tx6g/.

3 The article explores the case of continuous treatment variables. The case of discrete multivalued treatment variables is explored in the *Discussion and Conclusion* section.

4 This definition of potential outcomes follows from the stable unit treatment value assumption. It implies that the potential outcome for any unit does not depend on the treatment intensity of other units and that there are no different forms of each treatment intensity that lead to distinct potential outcomes. Moreover, it is assumed to be continuously differentiable (Hirano and Imbens, 2004; Liu et al., 2025).

5 See online supplement A for an alternative analytical discussion, which also explores the direction of the difference between $\Delta CAME$ and the $AIPE$.

6 It should be noted that this does not invalidate the estimands (both, D-CAME and AIPE), which are defined at the population level.

7 Online supplement B describes the DGP and derives the value of D-CAME and AIPE.

8 In cases with too many observations, I suggest two alternatives: either show the rugged histogram for a random sample of 1,000 observations from each group, or display the percentiles (p1, p10, p25, ..., p99).

9 84 percent intervals will capture between-group differences at the 0.05 significance level when estimates are normally distributed, independent, and have equal standard errors. These are quite restrictive assumptions. If directly interested in these group differences, researchers should directly test them, as 84 percent confidence intervals can sometimes be misleading (Armstrong II and Poirier, 2025). It should be noted that the purpose of reporting predicted values is not to assess whether they differ from zero, so 95 percent confidence intervals are not very informative in this context.

10 To download the package, researchers should execute the following command: `remotes::install_github(''SerranoSerrat/dcameaipe'')`.

11 Online supplement D discusses slight variations on how this is performed depending on the theoretical estimand of interest.

12 There are other ways in which researchers could select the functional form of the relationship. Mize and Kincaid (2025) run a regression with a discretized version the focal variable to get a sense of the relationship before running a model imposing a functional form. In a similar vein, researchers could use least squares binscatters, adding controls,

and selecting the optimal number of bins as in Cattaneo et al. (2024). From this, researchers can choose the preferred functional form. Note that either approach may be problematic, as both are forms of pretesting; that is, they use the same data to infer the functional form and to estimate the model.

13 Of course, this does not mean there are no specific functional forms that this approach may fail to capture. The aim of the graph is to show that including the specific functional form is not strictly necessary to remain unbiased.

14 This is an intentionally brief discussion. See Keele (2008) for an introduction to semiparametric models, including GAMs, and Wood (2017) for how GAMs are used in `R`. Simonsohn (2024) provides a useful discussion of estimating interactions with GAMs. What is new about this article is not the use of any specific method, but how these methods can be used to estimate the proposed quantities of interest.

15 The `R` package calls the *mgcv* package when estimating GAMs, using their default options (Wood, 2001). However, it allows researchers to choose the basis dimension, which controls the maximum wiggliness of the smooth, as well as the smoothness selection method. In particular, rather than using REML as the smoothness selection method, researchers can select generalized cross-validation, which provides a computationally efficient approximation to leave-one-out cross-validation (Keele, 2008). When the number of observations is not large, REML outperforms Generalized Cross-Validation (GCV), which tends overfit the data.

16 The names of the panels have been adapted. Instead of distinguishing between functional forms included in the basis expansion (which does not apply here), the distinction is whether the true functional form is differentiable at all points or not.

17 In this setting, researchers may decide between two specifications. First, the full tensor product, estimates $g(D, X)$ as a single joint surface, absorbing main effects and interaction into one term. Second, the tensor product interaction, instead decomposes the surface as $g(D, X) = g_1(D) + g_2(X) + g_{12}(D, X)$, where g_{12} captures only the pure interaction component net of the marginal smooths. In practice, for the proposed approach on how to estimate D-CAME and AIPE, the differences are small. The `R` package uses tensor product interaction as the default, but allows researchers to decide which approach to use.

18 One of this instances is discussed in online supplement D, where the cross-validation problem can be adapted to better estimate the AIPE.

19 Specifically, I use their instrument, which captures arguably exogenous exposure to robotization, and report the reduced-form results. See Anelli et al., 2021 for a more detailed description of the variables.

20 In particular, regions are defined at the second level of European Union's Nomenclature of Territorial Units for Statistics (NUTS2).

21 I use GCV instead of REML to estimate the smoothing parameters because REML is computationally intensive in large samples, especially when bootstrapping standard errors. Although GCV may undersmooth the data, this issue is mainly relevant in smaller samples. Online supplement E shows that the point estimates are similar when using REML.

References

Ai, Chunrong and Edward C. Norton. 2003. "Interaction Terms in Logit and Probit Models." *Economics letters* 80(1):123–9. https://doi.org/10.1016/S0165-1765(03)00032-6

Anelli, Massimo, Italo Colantone, and Piero Stanig. 2021. "Individual Vulnerability to Industrial Robot Adoption Increases Support for the Radical Right." *Proceedings of the National Academy of Sciences* 118(47):e2111611118. https://doi.org/10.1073/pnas.2111611118

Armstrong II, David A. and William Poirier. 2025. "Decoupling Visualization and Testing when Presenting Confidence Intervals." *Political Analysis* 33(3):252–7. https://doi.org/10.1017/pan.2024.24

Bansak, Kirk. 2021. "Estimating Causal Moderation Effects with Randomized Treatments and Non-Randomized Moderators." *Journal of the Royal Statistical Society Series A: Statistics in Society* 184(1):65–86. https://doi.org/10.1111/rssa.12614

Blackwell, Matthew and Michael P. Olson. 2021. "Reducing Model Misspecification and Bias in the Estimation of Interactions." *Political Analysis* 30(4):495–514. https://doi.org/10.1017/pan.2021.19

Brambor, Thomas, William Roberts Clark, and Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14(1):63–82. https://doi.org/10.1093/pan/mpi014

Browne, Richard H. 1979. "On Visual Assessment of the Significance of a Mean Difference." *Biometrics* 35(3):657–65. https://doi.org/10.2307/2530259

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *The Review of Economics and Statistics* 90(3):414–27. https://doi.org/10.1162/rest.90.3.414

Cattaneo, Matias D., Richard K. Crump, Max H. Farrell, and Yingjie Feng. 2024. "On Binscatter." *American Economic Review* 114(5):1488–514. https://doi.org/10.1257/aer.20221576

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. 2017. "Double/Debiased/Neyman Machine Learning of Treatment Effects." *American Economic Review* 107(5):261–5. https://doi.org/10.1257/aer.p20171038

Clark, William Roberts and Matt Golder. 2023. *Interaction Models: Specification and Interpretation*. Cambridge University Press. https://doi.org/10.1017/9781108241762

De Kadt, Daniel and Anna Grzymala-Busse. 2025. "Good Description". *SocArXiv Preprint. Preprint*. (https://osf.io/preprints/socarxiv/e74a9_v1)

Ganzach, Yoav. 1997. "Misleading Interaction and Curvilinear Terms." *Psychological Methods* 2(3):235. https://doi.org/10.1037/1082-989X.2.3.235

Giesselmann, Marco and Alexander W. Schmidt-Catran. 2022. "Interactions in Fixed Effects Regression Models." *Sociological Methods & Research* 51(3):1100–27. https://doi.org/10.1177/0049124120914934

Hainmueller, Jens, Jiehan Liu, Ziyi Liu, Jonathan Mummolo, and Yiqing Xu. 2025. "A Response to Recent Critiques of Hainmueller, Mummolo and Xu (2019) on Estimating Conditional Relationships." *arXiv preprint arXiv:2502.05717*.

Hainmueller, Jens, Jonathan Mummolo, and Yiqing Xu. 2019. "How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice." *Political Analysis* 27(2):163–92. https://doi.org/10.1017/pan.2018.46

Hanmer, Michael J. and Kerem Ozan Kalkan. 2013. "Behind the Curve: Clarifying the Best Approach to Calculating Predicted Probabilities and Marginal Effects from Limited Dependent Variable Models." *American Journal of Political Science* 57(1):263–77. https://doi.org/10.1111/j.1540-5907.2012.00602.x

Hargens, Lowell L. 2009. “Product-Variable Models of Interaction Effects and Causal Mechanisms.” *Social Science Research* 38(1):19–28. https://doi.org/10.1016/j.ssresearch.2008.05.003

Hastie, Trevor and Robert Tibshirani. 1986. “Generalized Additive Models.” *Statistical Science* 1(3):297–318. https://doi.org/10.1214/ss/1177013604

Hirano, Keisuke and Guido W. Imbens. 2004. “The Propensity Score with Continuous Treatments.” Pp. 73–84 in *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin’s Statistical Family*, edited by Andrew Gelman and Xiao-Li Meng. Chichester, UK: Wiley. https://doi.org/10.1002/0470090456.ch7

Keele, Luke and Randolph T. Stevenson. 2021. “Causal Interaction and Effect Modification: Same Model, Different Concepts.” *Political Science Research and Methods* 9(3):641–9. https://doi.org/10.1017/psrm.2020.12

Keele, Luke John. 2008. *Semiparametric Regression for the Social Sciences*. John Wiley & Sons. https://doi.org/10.1002/9780470998137

King, Gary and Langche Zeng. 2006. “The Dangers of Extreme Counterfactuals.” *Political Analysis* 14(2):131–59. https://doi.org/10.1093/pan/mpj004

Lee, Byungkyu and Bernice A. Pescosolido. 2024. “Misery Needs Company: Contextualizing the Geographic and Temporal Link between Unemployment and Suicide.” *American Sociological Review* 89(6):1104–40. https://doi.org/10.1177/00031224241288179

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

Long, J. Scott and Sarah A. Mustillo. 2021. “Using Predictions and Marginal Effects to Compare Groups in Regression Models for Binary Outcomes.” *Sociological Methods & Research* 50(3):1284–320. https://doi.org/10.2139/ssrn.5201566

Lundberg, Ian, Rebecca Johnson, and Brandon M. Stewart. 2021. “What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory.” *American Sociological Review* 86(3):532–65. https://doi.org/10.1177/0049124118799374

Marra, Giampiero and Simon N. Wood. 2012. “Coverage Properties of Confidence Intervals for Generalized Additive Model Components.” *Scandinavian Journal of Statistics* 39(1):53–74. https://doi.org/10.1177/00031224211004187

Mize, Trenton D. 2019. “Best Practices for Estimating, Interpreting, and Presenting Nonlinear Interaction Effects.” *Sociological Science* 6:81–117. https://doi.org/10.1111/j.1467-9469.2011.00760.x

Mize, Trenton D. and Reilly Kincaid. 2025. “Role-Accumulation and Mental Health across the Life Course.” *American Sociological Review* 90(2):226–56. https://doi.org/10.15195/v6.a4

Rainey, Carlisle. 2016. “Compression and Conditional Effects: A Product Term Is Essential When Using Logistic Regression to Test for Interaction.” *Political Science Research and Methods* 4(3):621–39. https://doi.org/10.1177/00031224241313394

Ratkovic, Marc and Dustin Tingley. 2023. “Estimation and Inference on Nonlinear and Heterogeneous Effects.” *The Journal of Politics* 85(2):421–35. https://doi.org/10.1017/psrm.2015.59

Rubio-Cabañez, Maria. 2025. “Stratifying Cities: the Effect of Outdoor Recreation Areas on Children’s Well-Being.” *European Sociological Review* 41(3):472–86. https://doi.org/10.1086/723811

Savage, Mike. 2020. "What Makes for a Successful Sociology? A Response to "Against a Descriptive Turn"." *The British Journal of Sociology* 71(1):19–27. https://doi.org/10.1093/esr/jcae028

Savage, Mike. 2024. "In Defence of Sociological Description: A 'World-Making' Perspective." *British Journal of Sociology* 75(3):360–65. https://doi.org/10.1111/1468-4446.12713

Simonsohn, Uri. 2023. "Don't Bin, GAM Instead: Hainmueller et al.'s Binning & Kernel Estimators Are Only Valid for Experimental Data." *Available at SSRN 4638476*.

Simonsohn, Uri. 2024. "Interacting with Curves: How to Validly Test and Probe Interactions in the Real (Nonlinear) World." *Advances in Methods and Practices in Psychological Science* 7(1):25152459231207787. https://doi.org/10.1177/25152459231207787

VanderWeele, Tyler J. 2009. "On the Distinction between Interaction and Effect Modification." *Epidemiology* 20(6):863–71. https://doi.org/10.1097/EDE.0b013e3181ba333c

Wood, Simon N. 2001. "mgcv: GAMs and generalized ridge regression for R." *R news* 1(2):20–25.

Wood, Simon N. 2017. *Generalized Additive Models: An Introduction with R*. CRC Press. https://doi.org/10.1201/9781315370279

Yitzhaki, Shlomo. 1996. "On Using Linear Regressions in Welfare Economics." *Journal of Business & Economic Statistics* 14(4):478–86. https://doi.org/10.1080/07350015.1996.10524677

Zhang, Letian. 2023. "Racial Inequality in Work Environments." *American Sociological Review* 88(2):252–83. https://doi.org/10.1177/00031224231157303

Zhirnov, Andrei, Mert Moral, and Evgeny Sedashov. 2023. "Taking Distributions Seriously: On the Interpretation of the Estimates of Interactive Nonlinear Models." *Political Analysis* 31(2):213–34. https://doi.org/10.1017/pan.2022.9

Acknowledgments: An earlier version of this article was presented at the 2024 Annual Meeting of the American Political Science Association in Philadelphia and the 2025 Annual Conference of the Swiss Political Science Association in Genève. I am grateful to Cesc Amat, Ignacio Jurado, Patrick W. Kraft, Sergi Martínez, and Marco Steenbergen for their valuable comments and suggestions, which substantially improved the article. I also thank Felix Elwert, as Editor of *Sociological Methods & Research*, Arnout van de Rijt, as Editor of *Sociological Science*, and an anonymous reviewer for their insightful feedback and constructive guidance throughout the revision process.

Josep Serrano-Serrat: Institute for Public Goods and Policies (IPP), Spanish National Research Council (CSIC). E-mail: j.serrano.serrat@csic.es.