How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice

Jens Hainmueller Jonathan Mummolo Yiqing Xu*

April 20, 2018

(*Political Analysis*, forthcoming)

Abstract

Multiplicative interaction models are widely used in social science to examine whether the relationship between an outcome and an independent variable changes with a moderating variable. Current empirical practice tends to overlook two important problems. First, these models assume a linear interaction effect that changes at a constant rate with the moderator. Second, estimates of the conditional effects of the independent variable can be misleading if there is a lack of common support of the moderator. Replicating 46 interaction effects from 22 recent publications in five top political science journals, we find that these core assumptions often fail in practice, suggesting that a large portion of findings across all political science subfields based on interaction models are modeling artifacts or are at best highly model dependent. We propose a checklist of simple diagnostics to assess the validity of these assumptions and offer flexible estimation strategies that allow for nonlinear interaction effects and safeguard against excessive extrapolation. These statistical routines are available in both `R` and `STATA`.

*Jens Hainmueller is a Professor of Political Science, Stanford University, jhain@stanford.edu. Jonathan Mummolo is an Assistant Professor of Politics and Public Affairs, Princeton University, jmummolo@princeton.edu. Yiqing Xu is an Assistant Professor of Political Science, University of California, San Diego, yiqingxu@ucsd.edu. We thank Licheng Liu for excellent research assistance. We thank David Broockman, Daniel Carpenter, James Fowler, Justin Grimmer, Seth Hill, Macartan Humphreys, Kosuke Imai, Gabe Lenz, Neil Malhotra, John Marshall, Marc Ratkovic, Jas Sekhon, Sean Westwood and participants at the PolMeth, APSA and MPSA annual meetings and at methods workshops at Massachusetts Institute of Technology, Harvard University, Princeton University, Columbia University and University of California, San Diego for helpful feedback. We also thank the authors of the studies we replicate for generously sharing code and data. Replication data available in Hainmueller, Mummolo and Xu (2018).

1 Introduction

The linear regression model with multiplicative interaction terms of the form

$$Y = \mu + \alpha D + \eta X + \beta(D \cdot X) + \epsilon$$

is a workhorse model in the social sciences for examining whether the relationship between an outcome Y and a key independent variable D varies with levels of a moderator X, which is often meant to capture differences in context. For example, we might expect that the effect of D on Y grows with higher levels of X. Such conditional hypotheses are ubiquitous in the social sciences and linear regression models with multiplicative interaction terms are the most widely used framework for testing them in applied work.[1]

A large body of literature advises scholars how to test such conditional hypotheses using multiplicative interaction models. For example, Brambor, Clark and Golder (2006) provide a simple checklist of dos and don'ts.[2] They recommend that scholars should (1) include in the model all constitutive terms (D and X) alongside the interaction term ($D \cdot X$), (2) not interpret the coefficients on the constitutive terms (α and η) as unconditional marginal effects, and (3) compute substantively meaningful marginal effects and confidence intervals, ideally with a plot that shows how the conditional marginal effect of D on Y changes across levels of the moderator X.

The recommendations given in Brambor, Clark and Golder (2006) have been highly cited and are nowadays often considered the best practice in political science.[3] As our

[1] There obviously exist many sophisticated estimation approaches that are more flexible such as Generalized Additive Models (Hastie and Tibshirani 1986; Wood 2003), Neural Networks (Beck, King and Zeng 2000), or Kernel Regularized Least Squares (Hainmueller and Hazlett 2013). We do not intend to critique these approaches. Our perspective for this study is that many applied scholars prefer to remain in their familiar regression framework to test conditional hypotheses and our proposals are geared towards this audience. Also, since our replications are based on articles published in the top political science journals our conclusions about the state of empirical practice apply to political science, although similar problems might be present in other disciplines.

[2] Other advice includes Friedrich 1982; Aiken, West and Reno 1991; Jaccard and Turrisi 2003; Braumoeller 2004; Kam and Franzese Jr. 2007; Berry, Golder and Milton 2012.

[3] As of February 2018, Brambor, Clark and Golder (2006) has been cited over 4,200 times according

survey of five top political science journals from 2006-2015 suggests, most articles with interaction terms now follow these guidelines and routinely report interaction effects with the marginal effects plots recommended in Brambor, Clark and Golder (2006). In addition, scholars today rarely leave out constitutive terms or misinterpret the coefficients on the constitutive terms as unconditional marginal effects. Clearly, empirical practice improved with the publication of Brambor, Clark and Golder (2006) and related advice.

Despite these advances, we contend that the current best practice guidelines for using multiplicative interaction models fail to address key issues, especially in the common scenario where at least one of the interacted variables is continuous. In particular, we emphasize two important problems that are currently often overlooked and not detected by scholars using the existing guidelines.

First, while multiplicative interaction models allow the effect of the key independent variable D to vary across levels of the moderator X, they maintain the important assumption that the interaction effect is linear and follows the functional form given by $\frac{\partial Y}{\partial D} = \alpha + \beta X$. This linear interaction effect (LIE) assumption states that the effect of D on Y can only linearly change with X at a constant rate given by β. In other words, the LIE assumption implies that the heterogeneity in effects is such that as X increases by one unit, the effect of D on Y changes by β and this change in the effect is constant across the whole range of X. Perhaps not surprisingly, this LIE assumption often fails in empirical settings because many interaction effects are not linear and some may not even be monotonic. In fact, replicating nearly 50 interaction effects that appeared in 22 articles published in the top five political science journals between 2006 and 2015, we find that the effect of D on Y changes linearly in only about 48% of cases. In roughly 70% of cases, we cannot even reject the null that the effect of the key independent variable of interest is equal at typical low and typical

to Google Scholar, which makes it one of the most cited political science articles in recent decades.

high levels of the moderator once we relax the LIE assumption that underlies the claim of an interaction effect in the original studies. This suggests that a large share of published work across all empirical political science subfields using multiplicative interaction models draws erroneous conclusions that rest on a modeling artifact that goes undetected even when applying the current best practice guidelines. It is worth noting that researchers can use a regression model as a linear approximation for the unknown true model. However, the linear marginal effects plots in the studies that we review, as well as the accompanying discussions therein, show that many authors take the LIE assumption quite literally and treat the linear interaction model as the true model. That is, both in text and in their marginal effects plots, researchers move beyond on-average conclusions and instead claim to have estimated the marginal effect of the treatment at specific values of the moderator, results which rely heavily on the linear functional form being correct.[4]

Second, another important problem that is often overlooked is the issue of lack of common support. Scholars using multiplicative interaction models routinely report the effect of D on Y across a wide range of X values by plugging the X values into the conditional marginal effects formula $\frac{\partial Y}{\partial D} = \alpha + \beta X$. However, often little attention is paid as to whether there is sufficient common support in the data when computing the conditional marginal effects. Ideally, to compute the marginal effect of D at a given value of the moderator, X_o, there needs to be (1) a sufficient number of observations whose X values are close to X_o and (2) variation in the treatment, D, at X_o. If either of these two conditions fails, the conditional marginal effects estimates are based on

[4] For example, in Bodea and Hicks (2015*b*), the authors wrote: "At low levels of POLITY, the marginal effect of CBI is positive but statistically insignificant. Similarly, the marginal effect of CBI is negative and significant only when the FREEDOM HOUSE score is greater than about 5" (p. 49). Similarly, in Clark and Leiter (2014), the authors write that when the moderator, "party dispersion," is set to one standard deviation above its mean, "a change from the minimum value of valence to the maximum value. . . " corresponds to "a 10-point increase in predicted vote share—more than double that of predicted change in vote share for the mean value of party dispersion, and a sufficient change in vote share to move a party from government to opposition." (p. 186). We thank the Editor and anonymous reviewers for highlighting this point.

extrapolation or interpolation of the functional form to an area where there is no or only sparse data and therefore the effect estimates are fragile and model dependent (King and Zeng 2006). In our replications we find that this type of extrapolation is very common in empirical practice. Typically articles report conditional marginal effect estimates for the entire range of the moderator which often includes large intervals where there are no or very few observations. Similarly, some articles report conditional marginal effects estimates for values of the moderator where there is no variation in the key independent variable of interest. Overall, our replications suggest that scholars are not sufficiently aware of the lack of common support problem and draw conclusions based on highly model dependent estimates. And according to our replications, these problems are common to all empirical subfields in political science.

Our goal is not to point fingers. Indeed, in the vast majority of studies we replicate below researchers were employing the accepted best practices at the time of publication. Our goal is to improve empirical practice. To this end we develop a set of simple diagnostic tests that help researchers to detect these currently overlooked and important problems. In addition, we offer simple semi-parametric modeling strategies that allow researchers to remain in their familiar regression framework and estimate conditional marginal effects while relaxing the LIE assumption and avoiding model dependency that stems from excessive extrapolation.

Our diagnostics and estimation strategies are easy to implement using standard software packages. We propose a revised checklist that augments the existing guidelines for best practice. We also make available the code and data that implements our methods and replicates our figures in `R` and `STATA`.[5]

While the focus of our study is on interaction models, we emphasize that the issues of model misspecification and lack of common support are not unique to these models

[5] You can install `R` package `interflex` from CRAN and `STATA` package `interflex` from SSC. For more information, see `http://yiqingxu.org/software.html#interflex`.

and also apply to regression models without interaction terms. However, we find that these issues may more often go overlooked in interaction models because marginal effects estimates involve three key variables—the treatment, moderator and response—requiring different diagnostic approaches to assess both functional form and common support.

In fact, as we show below, the LIE assumption implies that the conditional effect of D is the difference between two linear functions in X and therefore the assumption is unlikely to hold unless both of these functions are indeed linear. Similarly, there is often insufficient common support in X across different values of D if the distribution of D and/or X is highly skewed, or if one of the variables does not vary in some regions of the joint support of D and X.

The rest of the article proceeds as follows. In the next section we discuss the problems with the multiplicative interaction model. In the third section we introduce our diagnostic tools and estimation strategies. In the fourth section we apply them to the replication data. The last section provides our revised guidelines for best practice and concludes.

2 Multiplicative Interaction Models

Consider the classical linear multiplicative interaction model that is often assumed in empirical work and is given by the following regression equation:

$$Y = \mu + \eta X + \alpha D + \beta (D \cdot X) + Z\gamma + \epsilon. \tag{1}$$

In this model Y is the outcome variable, D is the key independent variable of interest or "treatment", X is the moderator—a variable that affects the direction and/or strength of the treatment effect,[6] $D \cdot X$ is the interaction term between D and X, Z is a vector

[6] A moderator is different from a mediator, which is a variable that accounts for at least part of the treatment effect (see, for example, Imai, Keele and Yamamoto (2010)).

of control variables, and μ and ϵ represent the constant and error terms, respectively.[7]

We focus on the case where the treatment variable D is either binary or continuous and the moderator X is continuous. When D and X are both binary or discrete with few unique values one should employ a fully saturated model that dummies out the treatment and the moderator and includes all interaction terms to obtain the treatment effect at each level of X. Moreover, in the following discussion we focus on the interaction effect components of the model (D, X, and $D \cdot X$). When covariates Z are included in the model, we maintain the typical assumption that the model is correctly specified with respect to these covariates.[8]

The coefficients of Model (1) are consistently estimated under the usual linear regression assumptions which imply that the functional form is correctly specified and that $\mathbb{E}[\epsilon|D, X, Z] = 0$. In the multiplicative interaction model this implies the linear interaction effects (LIE) assumption which says that the *marginal effect* of the treatment D on the outcome Y is

$$ME_D = \frac{\partial Y}{\partial D} = \alpha + \beta X, \tag{2}$$

which is a linear function of the moderator X. This LIE assumption implies that the effect of D on Y can only linearly change with X, so if X increases by one unit, the effect of D on Y changes by β and this change in the effect is constant across the whole range of X. This is a strong assumption, because we often have little theoretical or empirical reason to believe that the heterogeneity in the effect of D on Y takes such a linear form. Instead, it might well be that the effect of D on Y is non-linear or

[7] Note that the designation of one of the independent variables as the treatment and the other as the moderator is done without loss of generality. The typical approach in most empirical studies in our replication sample is to designate one variable as the treatment of interest and another variable as the moderator, for example in randomized experiments or observational studies where one variable is (quasi) randomly assigned and a pre-treatment covariate moderates the treatment effect on the outcome. In other designs, such as multi-factorial experiments, there might be two variables that can be viewed as treatments and the same diagnostics and estimation strategies that we propose here can be applied to estimate how the effect of one treatment is moderated by the other and vice versa.

[8] Note that our kernel estimator below relaxes the linearity assumption on Z.

non-monotonic. For example, the effect might be small for low values of X, large at medium values of X, and then small again for high values of X.

The LIE assumption in Equation (2) means that the relative effect of treatment $D = d_1$ versus $D = d_2$ can be expressed by the difference between two linear functions in X:

$$\begin{aligned} \mathit{Eff}(d_1, d_2) =& Y(D = d_1|X, Z) - Y(D = d_2|X, Z) \\ =& (\mu + \alpha d_1 + \eta X + \beta d_1 X) - (\mu + \alpha d_2 + \eta X + \beta d_2 X) \\ =& \alpha(d_1 - d_2) + \beta(d_1 - d_2)X. \end{aligned} \tag{3}$$

This decomposition makes clear that under the LIE assumption, the effect of D on Y is the difference between two linear functions, $\mu + \alpha d_1 + (\eta + \beta d_1)X$ and $\mu + \alpha d_2 + (\eta + \beta d_2)X$, and therefore the LIE assumption will be most likely to hold if both functions are linear for all modeled contrasts of d_1 versus d_2.[9][10]

This illustrates how attempts to estimate interaction effects with multiplicative interaction models are susceptible to misspecification bias because the LIE assumption will fail if one or both functions are misspecified due to non-linearities, non-monotonicities, a skewed distribution of X and/or D resulting in outliers or bad influence points, etc. As our empirical survey shows below, in practice this LIE assumption often fails because at least one of the two functions is not linear.[11]

The decomposition in Equation (3) also highlights the issue of common support. Since the conditional effect of D on Y is the difference between two linear functions, it

[9] The LIE assumption would also hold in the special case where both functions are non-linear but the difference between both of these functions is a linear function (e.g. they diverge at a constant rate). This is unlikely in empirical settings and never occurs in any of our replications of nearly 50 recently published interaction effects.

[10] Note that in the special case of a binary treatment variable (say, $d_1 = 1$ and $d_2 = 0$), the marginal effect of D on Y is: $ME_D = \mathit{Eff}(1, 0) = Y(D = 1|X, Z) - Y(D = 0|X, Z) = \alpha + \beta X$, which is consistent with Equation (2). The term γZ is left out given the usual assumption that the specification is correct in both equations with respect to the control variables Z.

[11] Although the linear regression framework is flexible enough to incorporate higher order terms of X and their interaction with D (see Kam and Franzese Jr. 2007; Berry, Golder and Milton 2012) this is rarely done in practice. In fact, not a single study incorporated higher order terms in our replication sample of nearly 50 recently published interaction effects (see below).

is important that the two functions share *common support* over X. In other words, at any given value of the moderator $X = x_0$, there should be (1) a sufficient number of data points in the neighborhood of $X = x_0$ and (2) those data points need to exhibit variation in the treatment, D. If, for example, in the neighborhood of $X = x_0$ all data points are treated units ($D = 1$), we have a lack of common support and, since there are no control units ($D = 0$) in the same region at all, the estimated conditional effect will be entirely driven by interpolation or extrapolation and thus be highly model dependent.[12]

Multiplicative interaction models are again especially susceptible to the lack of common support problem because if the goal is to estimate the conditional effect of D across the range of X then this requires common support across the entire joint distribution of D and X. Otherwise, estimation of the conditional marginal effect will rely on interpolation or extrapolation of at least one of the functions to an area where there is no or only very few observations. It is well known that such interpolation or extrapolation purely based on an assumed functional form results in fragile and highly model dependent estimates. Slight changes in the assumed functional form or data can lead to very different answers (King and Zeng 2006). In our empirical survey below we show that such interpolation or extrapolation is common in applied work using multiplicative interaction models.

In sum, there are two important problems with multiplicative interaction models. The LIE assumption states that the interaction effect is linear, but if this assumption fails, the conditional marginal effects estimates are inconsistent and biased. In addition, the common support condition suggests that we need sufficient data on X and D because otherwise the estimates will be highly model dependent. Both problems are currently overlooked because they are not detected by scholars following the cur-

[12] Of course, if the model happens to be correct, estimated conditional effects will still be consistent and unbiased despite the common support issue. We thank an anonymous reviewer for highlighting this point.

rent best practice guidelines. In the next section we develop simple diagnostic tools and estimation strategies that allow scholars to diagnose these problems and estimate conditional marginal effects while relaxing the LIE assumption.

3 Diagnostics

Before introducing the diagnostic tools, we present simulated data samples to highlight three scenarios: (1) linear marginal effect with a dichotomous treatment, (2) linear marginal effect with a continuous treatment, and (3) nonlinear marginal effect with a dichotomous treatment.

The data generating process (DGP) for both samples that contain a linear marginal effect is as follows:

$$Y_i = 5 - 4X_i - 9D_i + 3D_iX_i + \epsilon_i, \qquad i = 1, 2, \cdots, 200.$$

Y_i is the outcome for unit i, the moderator is $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(3,1)$, and the error term is $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,4)$. Both samples share the same sets of X_i and ϵ_i, but in the first sample, the treatment indicator is $D_i \overset{\text{i.i.d.}}{\sim} \mathit{Bernoulli}(0.5)$, while in the second one it is $D_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(3,1)$. The marginal effect of D on Y therefore is $ME_D = -9 + 3X$.

The DGP for the sample with a nonlinear marginal effect is:

$$Y_i = 2.5 - X_i^2 - 5D_i + 2D_iX_i^2 + \zeta_i, \qquad i = 1, 2, \cdots, 200.$$

Y_i is the outcome, the moderator is $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}(-3,3)$, the treatment indicator is $D_i \overset{\text{i.i.d.}}{\sim} \mathit{Bernoulli}(0.5)$, and the error term is $\zeta_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,4)$. The marginal effect of D on Y therefore is $ME_D = -5 + 2X^2$. For simplicity, we do not include any control variables. Note that all three samples have 200 observations.

We now present a simple visual diagnostic to help researchers to detect potential problems with the LIE assumption and the lack of common support. The diagnostic that we recommend is a scatterplot of raw data. This diagnostic is simple to implement and powerful in the sense that it readily reveals the main problems associated with the LIE assumption and lack of common support.

If the treatment D is binary, we recommend plotting the outcome Y against the moderator X separately for the sample of treatment group observations ($D = 1$) and the sample of control group observations ($D = 0$). In each sample we recommend superimposing a linear regression line as well as LOESS fits in each group (Cleveland and Devlin 1988).[13] The upper panel of Figure 1 presents examples of such a plot for the simulated data with the binary treatment in cases where the marginal effect is (a) linear and (b) nonlinear.

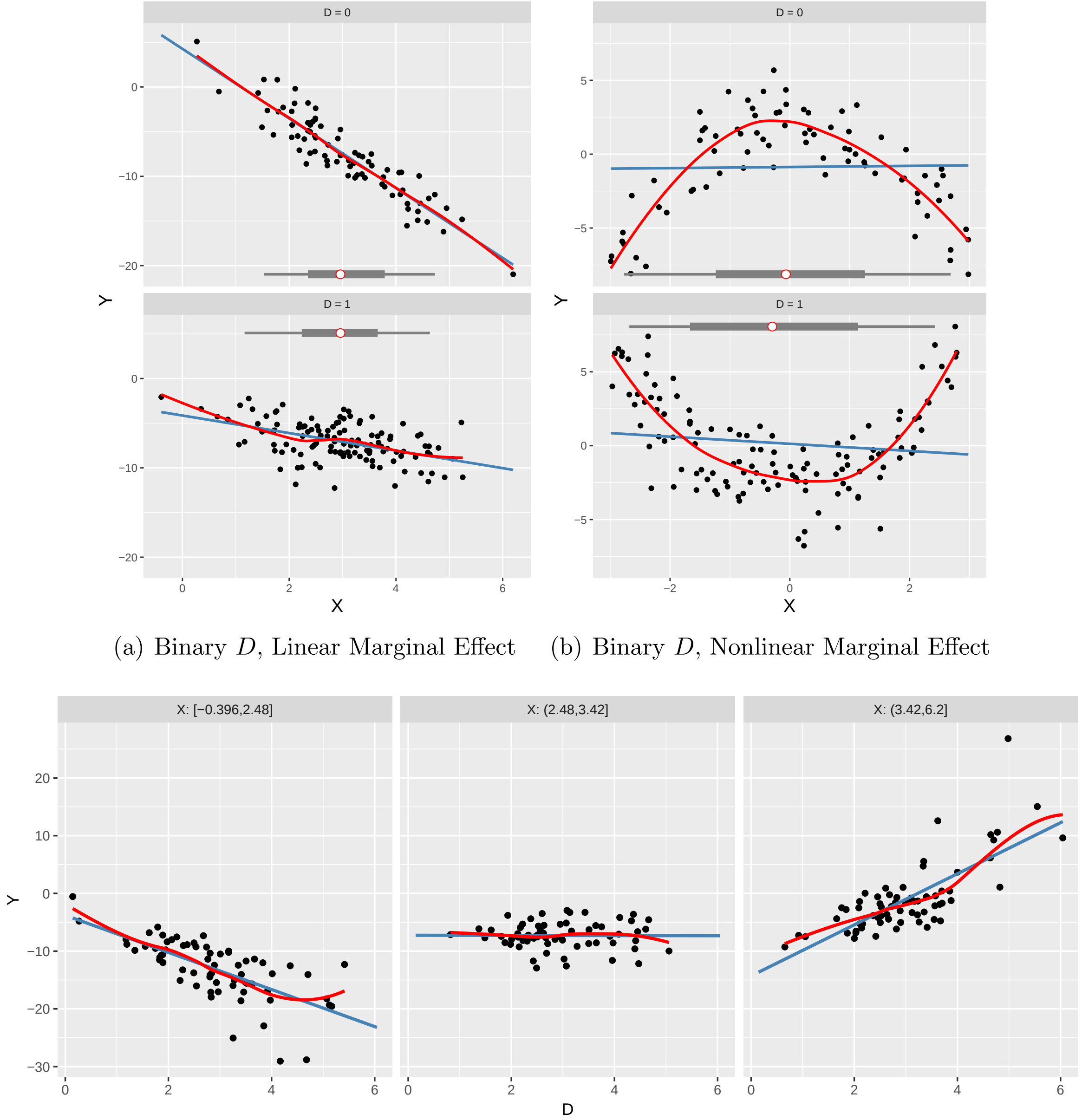
The first important issue to check is whether the relationship between Y and X is reasonably linear in both groups. For this we can simply check if the linear regression lines (blue) and the LOESS fits (red) diverge considerably across the range of X values. In Figure 1(a), where the true DGP contains a linear marginal effect, the two lines are very close to each other in both groups indicating that both conditional expectation functions are well approximated with a linear fit as required by the LIE assumption. However, as Figure 1(b) shows, LOESS and OLS will diverge considerably when the true marginal effect is nonlinear, thus alerting the researcher to a possible misspecification error.

We call these plots the *Linear Interaction Diagnostic* (LID) plots. In addition to shedding light on the validity of the LIE assumption, they provide other important insights as well. In Figure 1(a), the slope of Y on X in the treatment group is apparently larger (less negative) than that of the control group ($\hat{\eta} + \hat{\beta} > \hat{\eta}$), suggesting a possible positive interaction effect of D and X on Y. The LOESS fit in Figure 1(b) also gives evidence that the relationship between X and Y differs between the two groups, (in fact, the functions are near mirror opposites), a result that is masked by the OLS fit.

A final important issue to look out for is whether there is sufficient common support

[13] The same plots can be constructed after residualizing with respect to the covariates Z.

Figure 1. Linear Interation Diganostic Plots: Simulated Samples



(a) Binary D, Linear Marginal Effect (b) Binary D, Nonlinear Marginal Effect

(c) Continuous D, Linear Marginal Effect

Note: The above plots show the relationships among the treatment D, the outcome Y, and the moderator X using the raw data: (a) when D is binary and the true marginal effect is linear; (b) when D is binary and the true marginal effect is nonlinear (quadratic); and (c) when D is continuous and the true marginal effect is linear.

in the data. For this we can simply compare the distribution of X in both groups and examine the range of X values for which there are a sufficient number of data points for the estimation of marginal effects. The box plots near the center of the figures display quantiles of the moderator at each level of the treatment. The dot in the center denotes the median, the end points of the thick bars denote the 25th and 75th percentiles, and the end points of the thin bars denote the 5th and 95th percentiles. In Figure 1(a), we see that both groups share a common support of X for the range between about 1.5 to 4.5—whereas support exists across the entire range of X in Figure 1(b)—as we would expect given the simulation parameters.[14]

If the treatment and moderator are continuous, then visualizing the conditional relationship of Y and D across levels of X is more complicated, but in our experience a simple binning approach is sufficient to detect most problems in typical political science data. Accordingly, we recommend that researchers split the sample into three roughly equal sized groups based on the moderator: low X (first tercile), medium X (second tercile), and high X (third tercile). For each of the three groups we then plot Y against D while again overlaying both the linear and LOESS fits.

Panel (c) of Figure 1 presents an LID plot for the simulated data with the continuous treatment and linear marginal effect. The plot reveals that the conditional expectation function of Y given D is well approximated by a linear model in all three samples of observations with low, medium, or high values on the moderator X.

There is also clear evidence of an interaction as the slope of the line which captures the relationship between D on Y is negative at low levels of X, flat at medium levels of X, and positive at high levels of X. In this case of a continuous treatment and continuous moderator it is also useful to generate the LID plot in both directions to examine the conditional relationships of $D|X$ and $X|D$ as the standard linear interaction model assumes linearity in both directions. Moreover, it can be useful to visualize interactions using a three-dimensional surface plot generated by a generalized additive model (GAM, Hastie and Tibshirani 1986).[15]

[14] In addition, researchers can plot the estimated density of X in both groups in a single plot to further judge the range of common support.

[15] See Appendix for more information on this strategy.

4 Estimation Strategies

In this section we develop two simple estimation strategies to estimate the conditional marginal effect of D on Y across values of the moderator X. These approaches have the advantage that they remain in the regression framework familiar to applied researchers and at the same time relax the LIE assumption and flexibly allow for heterogeneity in how the conditional marginal effect changes across values of X. When the marginal effect is indeed linear in X, these strategies are less efficient than the linear interaction model, however, they offer protection against excessive model dependency and the lack of common support—a classic case of the bias-variance trade-off.

4.1 Binning Estimator

The first estimation approach is a binning estimator. Simply put, we break a continuous moderator into several bins represented by dummy variables and interact these dummy variables with the treatment indicator, with some adjustment to improve interpretability.[16] There are three steps to implement the estimator. First, we discretize the moderator variable X into three bins (respectively corresponding to the three terciles) as before and create a dummy variable for each bin. More formally, we define three dummy variables that indicate the interval X falls into:

$$G_1 = \begin{cases} 1 & X < \delta_{1/3} \\ 0 & otherwise \end{cases}, \quad G_2 = \begin{cases} 1 & X \in [\delta_{1/3}, \delta_{2/3}) \\ 0 & otherwise \end{cases}, \quad G_3 = \begin{cases} 1 & X \geq \delta_{2/3} \\ 0 & otherwise \end{cases},$$

in which $\delta_{1/3}$ and $\delta_{2/3}$ are respectively the first and second terciles of X. We can choose other numbers in the support of X to create the bins but the advantage of using terciles is that we obtain estimates of the effect at typical low, medium, and high values of X. While three bins tend to work well in practice for typical political science data

[16] This idea is analogous to breaking a continuous variable, such as age, into several bins in a linear regression model. We thank the Editor for highlighting this point.

that we encountered in replicating nearly 50 recently published interaction effects, the researcher can create more than three bins in order to get a finer resolution of the effect heterogeneity. Increasing the number of bins requires a sufficiently large number of observations.

Second, we pick an evaluation point within each bin, x_1, x_2, and x_3, where we want to estimate the conditional marginal effect of D on Y. Typically, we choose x_1, x_2, and x_3 to be the median of X in each bin, but researchers are free to choose other numbers within the bins (for example, the means).

Third, we estimate a model that includes interactions between the bin dummies G and the treatment indicator D, the bin dummies and the moderator X minus the evaluation points we pick (x_1, x_2, and x_3), as well as the triple interactions. The last two terms are to capture the changing effect of D on Y within each bin defined by G. Formally, we estimate the following model:

$$Y = \sum_{j=1}^{3} \Big\{ \mu_j + \alpha_j D + \eta_j (X - x_j) + \beta_j (X - x_j) D \Big\} G_j + Z\gamma + \epsilon \tag{4}$$

in which μ_j, α_j, η_j, and β_j ($j = 1, 2, 3$) are unknown coefficients.

The binning estimator has several key advantages over the standard multiplicative interaction model given in Model (1). First, the binning estimator is much more flexible as it jointly fits the interaction components of the standard model to each bin separately, thereby relaxing the LIE assumption.[17] Since $(X - x_j)$ equals zero at each evaluation point x_j, the conditional marginal effect of D on Y at the chosen evaluation points within each bin, x_1, x_2, and x_3, is simply given by α_1, α_2, and α_3, respectively. Here, the conditional marginal effects can vary freely across the three bins and therefore can take on any non-linear or non-monotonic pattern that might describe the heterogeneity in the effect of D on Y across low, medium, or high levels

[17] Note that given the usual assumption that the model is correctly specified with respect to the covariates Z, we do not let γ vary for each bin. If more flexibility is required the researcher can also include the interactions between the bin indicators and the covariates Z to let γ vary by bin.

of X.[18]

Second, since the bins are constructed based on the support of X, the binning ensures that the conditional marginal effects are estimated at typical values of the moderator and do not rely on excessive extrapolation or interpolation.[19]

Third, the binning estimator is easy to implement using any regression software and the standard errors for the conditional marginal effects are directly estimated by the regression so there is no need to compute linear combinations of coefficients to compute the conditional marginal effects.

Fourth, the binning estimator provides a generalization that nests the standard multiplicative interaction model as a special case. It can therefore serve as a formal test on the validity of global LIE assumption imposed by the standard model. In particular, if the standard multiplicative interaction Model (1) is the true model, we have the following relationships:

$$\begin{aligned}
\mu &= \mu_j - \eta_j x_j \qquad & j = 1, 2, 3;\\
\eta &= \eta_j \qquad & j = 1, 2, 3;\\
\alpha &= \alpha_j - \beta_j x_j \qquad & j = 1, 2, 3;\\
\beta &= \beta_j \qquad & j = 1, 2, 3.
\end{aligned}$$

The marginal effect of D at $X = x_j$ $(j = 1, 2, 3)$, therefore, is:

$$ME(x_j) = \alpha_j = \alpha + \beta_j x_j = \alpha + \beta x_j.$$

[18] Note that in the context of a randomized experiment, a regression of the outcome on the treatment, the demeaned covariates, and the interaction between the treatment and the demeaned covariates provides a semi-parametric and asymptotically efficient estimator of the average treatment effect under the Neyman model for randomization inference (Lin 2013; Miratrix, Sekhon and Yu 2013). In this context, our binning estimator is similar except that it applies to subgroups of the sample defined by the bins of the moderator.

[19] Clearly, one could construct cases where the distribution of X within a bin is highly bimodal and therefore the bin median might involve interpolation, but this is not very common in typical political science studies. In fact, in our nearly 50 replications of recently published interaction effects we found not a single case where this potential problem occurs (see below).

In the appendix we formally show that when Model (1) is correct we have

$$\hat{\alpha}_j - (\hat{\alpha} + \hat{\beta} x_j) \xrightarrow{p} 0 \ , \qquad j = 1, 2, 3,$$

in which $\hat{\alpha}$ and $\hat{\beta}$ are estimated from Model (1) and $\hat{\alpha}_j$ ($j = 1, 2, 3$) are estimated using Model (4). As mentioned above, we face a bias-variance trade-off. In the special case when the standard multiplicative interaction model is correct and therefore the global LIE assumption holds, then—as the sample size grows—the marginal effects estimates from the binning estimator converge in probability on the unbiased marginal effects estimates from the standard multiplicative interaction model given by $ME(X) = \hat{\alpha} + \hat{\beta} X$. In this case, the standard estimator will be the most efficient estimator for the marginal effect at any given point in the range of the moderator and the estimates will be more precise than those from the binning estimator at the evaluation points simply because the linear model utilizes more information based on the modeling assumptions. However, when the linear interaction assumption does not hold, the standard estimator will be biased and inconsistent and researchers interested in minimizing bias are better off using the more flexible binning estimator that requires more degrees of freedom. Although the binning estimator may also have bias under the circumstance (the bias will disappear as the model becomes more and more flexible), disagreement between the binning estimates and estimates from the linear interaction model gives an indication the LIE assumption is invalid.

To illustrate the results from the binning estimator we apply it to our simulated datasets that cover the cases of a binary treatment with linear and nonlinear marginal effects. The results are shown in Figure 2. To clarify the correspondence between the binning estimator and the standard multiplicative interaction model we superimpose the three estimates of the conditional marginal effects of D on Y, $\hat{\alpha}_1$, $\hat{\alpha}_2$ and $\hat{\alpha}_3$, and their 95% confidence intervals from the binning estimator in their appropriate places (i.e., at $X = x_j$ in bin j) on the marginal effects plot generated from the standard multiplicative interaction model as recommended by Brambor, Clark and Golder (2006).

In the case of a binary treatment, we also recommend to display at the bottom of the figure a stacked histogram that shows the distribution of the moderator X. In this histogram the total height of the stacked bars refers to the distribution of the moderator in the pooled sample and the red and white shaded bars refer to the distribution of the moderator in the treatment and control groups, respectively. Adding such a histogram makes it easy to judge the degree to which there is common support in the data. In the case of a continuous treatment, we recommend a histogram at the bottom that simply shows the distribution of X in the entire sample.[20]

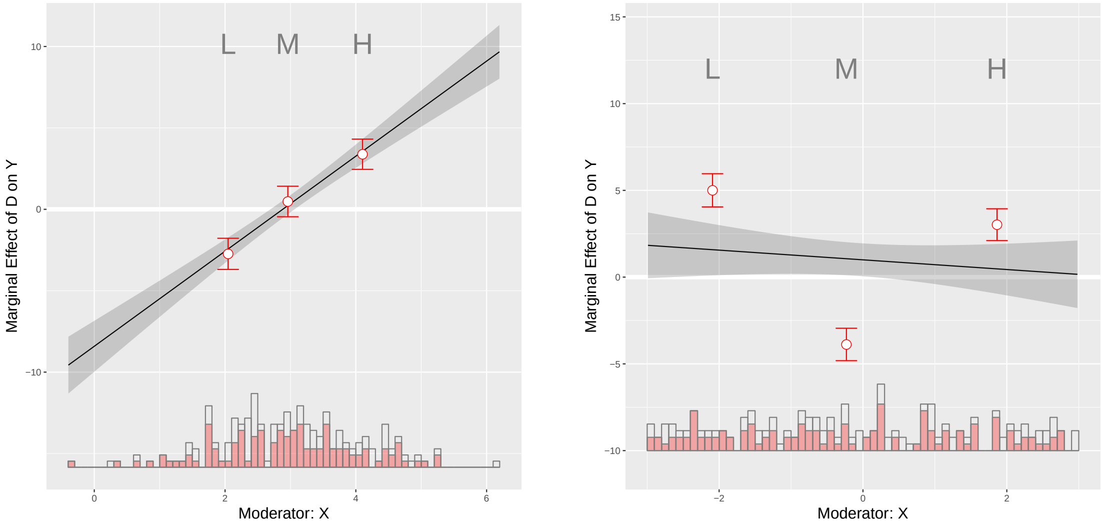
Figure 2(a) was generated using the DGP where the standard multiplicative interaction model is the correct model and therefore the LIE assumption holds. Hence, as Figure 2(a) shows, the conditional effect estimates from the binning estimator and the standard multiplicative interaction model are similar in both datasets. Even with a small sample size (i.e., $N = 200$), the three estimates from the binning estimator, labeled L, M, and H, sit almost right on the estimated linear marginal-effect line from the true standard multiplicative interaction model. Note that the estimates from the binning estimator are only slightly less precise than those from the true multiplicative interaction model, which demonstrates that there is at best a modest cost in terms of decreased efficiency from using this more flexible estimator. We also see from the histogram that the three estimates from the binning estimator are computed at typical low, medium, and high values of X with sufficient common support which is what we expect given the binning based on terciles.

Contrast these results with those in Figure 2(b), which were generated using our

[20] Berry, Golder and Milton (2012) also recommend adding a frequency distribution of the moderator to the marginal effects plots. We argue that in the case of a binary treatment it is advantageous to distinguish in the histogram between the two groups to get a better sense of the overlap across groups.

simulated data in which the true marginal effect of D is nonlinear. In this case, the standard linear model indicates a slightly negative, but overall very weak, interaction effect, whereas the binning estimates reveal that the effect of D is actually strongly conditioned by X: D exerts a positive effect in the low range of X, a negative effect in the midrange of X, and a positive effect again in the high range of X. In the event of such a nonlinear effect, the standard linear model delivers the wrong conclusion. When the estimates from the binning estimator are far off the line or when they are out of order, (for example, first increasing then decreasing), we have evidence that the LIE assumption does not hold.

Figure 2. Conditional Marginal Effects from Binning Estimator: Simulated Samples



(a) Linear Marginal Effect

(b) Nonlinear Marginal Effect

Note: The above plots show the estimated marginal effects using both the conventional linear interaction model and the binning estimator: (a) when the true marginal effect is linear; (b) when the true marginal effect is nonlinear (quadratic). In both cases, the treatment variable D is dichotomous.

4.2 Kernel Estimator

The second estimation strategy is a kernel smoothing estimator of the marginal effect, which is an application of semi-parametric smooth varying-coefficient models (Li and

Racine 2010). This approach provides a generalization that allows researchers to flexibly estimate the functional form of the marginal effect of D on Y across the values of X by estimating a series of local effects with a kernel reweighting scheme. While the kernel estimator requires more computation and its output is less easily summarized than that of the binning estimator, it is also fully automated (e.g. researchers do not need to select a number of bins) and characterizes the marginal effect across the full range of the moderator, rather than at just a few evaluation points.

Formally, the kernel smoothing method is based on the following semi-parametric model:

$$Y = f(X) + g(X)D + \gamma(X)Z + \epsilon, \tag{5}$$

in which $f(\cdot)$, $g(\cdot)$, and $\gamma(\cdot)$ are smooth functions of X, and $g(\cdot)$ captures the marginal effect of D on Y. It is easy to see that this kernel regression nests the standard interaction model given in Model (1) as a special case when $f(X) = \mu + \eta X$, $g(X) = \alpha + \beta X$ and $\gamma(X) = \gamma$. However, in the kernel regression the conditional effect of D on Y need not to be linear as required by the LIE assumption, but can vary freely across the range of X. In addition, if covariates Z are included in the model, the coefficients of those covariates are also allowed to vary freely across the range of X resulting in a very flexible estimator that also helps to guard against misspecification bias with respect to the covariates.

We use a kernel based method to estimate Model (5). Specially, for each given x_0 in the support of X, $\hat{f}(x_0)$, $\hat{g}(x_0)$, and $\hat{\gamma}(x_0)$ are estimated by minimizing the following weighted least squares objective function:

$$\left(\hat{\mu}(x_0), \hat{\alpha}(x_0), \hat{\eta}(x_0), \hat{\beta}(x_0), \hat{\gamma}(x_0)\right) = \underset{\tilde{\mu},\tilde{\alpha},\tilde{\eta},\tilde{\beta},\tilde{\gamma}}{\operatorname{argmin}}\ L(\tilde{\mu}, \tilde{\alpha}, \tilde{\eta}, \tilde{\beta}, \tilde{\gamma})$$

$$L = \sum_{i}^{N} \left\{ \left[Y_i - \tilde{\mu} - \tilde{\alpha} D_i - \tilde{\eta}(X_i - x_0) - \tilde{\beta} D_i (X_i - x_0) - \tilde{\gamma} Z_i \right]^2 K\left(\frac{X_i - x_0}{h} \right) \right\},$$

in which $K(\cdot)$ is a Gaussian kernel, h is a bandwidth parameter that we automatically select via least-squares cross-validation and $\hat{f}(x_0) = \hat{\mu}(x_0)$, $\hat{g}(x_0) = \hat{\alpha}(x_0)$. The two terms $\eta(X - x_0)$ and $\beta D(X - x_0)$ are included to capture the influence of the first partial derivative of Y with respect to X at each evaluation point of X, a common practice that reduces bias of the kernel estimator on the boundary of the support of X (e.g., Fan, Heckman and Wand 1995). As a result, we obtain three smooth functions $\hat{f}(\cdot)$, $\hat{g}(\cdot)$, and $\hat{\gamma}(\cdot)$, in which $\hat{g}(\cdot)$ represents the estimated marginal effect of D on Y with respect to X.[21] We implement this estimation procedure in both `R` and `STATA` and compute standard errors and confidence intervals using a bootstrap.

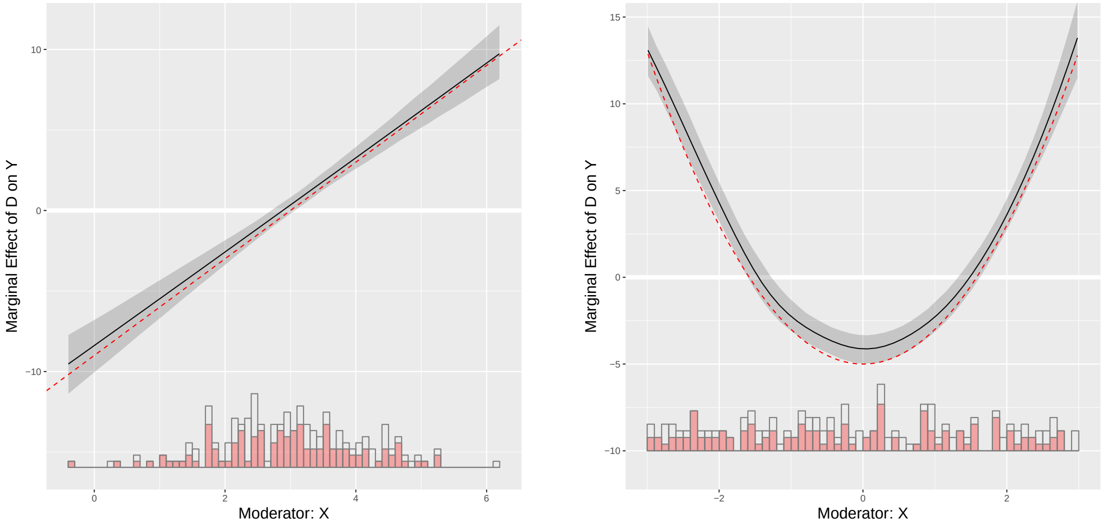
Figure 3 shows the results of our kernel estimator applied to the two simulated samples in which the true DGP contains either a linear or nonlinear marginal effect (the bandwidths are selected using a standard 5-fold cross-validation procedure). As in Figure 2, the x-axis is the moderator X and the y-axis is the estimated effect of D on Y. The confidence intervals are generated using 1,000 iterations of a non-parametric bootstrap where we resample the data with replacement. We again add our recommended (stacked) histograms at the bottom to judge the common support based on the distribution of the moderator.

Figure 3 shows that the kernel estimator is able to accurately uncover both linear and nonlinear marginal effects. Figure 3(a) shows a strong linear interaction where the conditional marginal effect of D on Y grows constantly and monotonically with X. The marginal effects estimates from the kernel estimator are close to those from the true multiplicative interaction model (red dashed line).[22] In addition, the kernel estimates in Figure 3(b) are a close approximation of the true quadratic marginal effect (red dashed line). In short, by utilizing a more flexible estimator, we are able to closely approximate the marginal effect whether the LIE assumption holds or not.

[21] For theoretical properties of the kernel smoothing estimator, see Li and Racine (2010).

[22] Compared with estimates from the conventional linear interaction model shown in Figure 2(a), we see that the kernel estimator does not result in a large increase in the uncertainty of the estimates when the linear interaction model is correct. This is mainly because when the LIE assumption is correct, the cross-validation scheme is likely to choose a large bandwidth.

Figure 3. Kernel Smoothed Estimates: Simulated Samples



(a) Linear Marginal Effect

(b) Nonlinear Marginal Effect

Note: The above plots show the estimated marginal effects using the kernel estimator: (a) when the true marginal effect is linear; (b) when the true marginal effect is nonlinear (quadratic). In both cases, the treatment variable D is dichotomous. The dotted line denotes the "true" marginal effect, while the solid line denotes the marginal effect estimate.

Also note that towards the boundaries in Figure 3(a), where there is limited common support on X, the conditional marginal effects estimates are increasingly imprecisely estimated as expected given that even in this simulated data there is less data to estimate the marginal effects at these points. The fact that the confidence intervals grow wider at those points is desirable because it makes clear the increasing lack of common support.

5 Data

We now apply our diagnostic and estimation strategies to published papers that used classical linear interaction models and claimed an interaction effect. To broadly assess the practical validity of the assumptions of the multiplicative interaction model, we canvassed studies published in five top political science journals, *The American Political Science Review* (APSR), *The American Journal of Political Science* (AJPS), *The*

Journal of Politics (JOP), *International Organization* (IO) and *Comparative Political Studies* (CPS).[23] Sampling occurred in two stages. First, for all five journals, we used Google Scholar to identify every study which cited Brambor, Clark and Golder (2006), roughly 170 articles. Within these studies, we subset to cases which: used plain OLS; had a substantive claim tied to an interaction effect; and interacted at least one continuous variable. We excluded methods and review articles, as well as triple interactions because those models impose even more demanding assumptions.

Second, we conducted additional searches to identify all studies published in the APSR and AJPS which included the terms "regression" *and* "interaction" published since Brambor, Clark and Golder (2006), roughly 550 articles. We then subset to articles which did *not* cite Brambor, Clark and Golder (2006). In order to identify studies within this second sample which featured interaction models prominently, we selected articles which included a marginal effect plot of the sort recommended by Brambor, Clark and Golder (2006) and then applied the same sampling filters as above. In the end, these two sampling strategies produced roughly 40 studies that met our sampling criteria.

After identifying these studies, we then sought out replication materials by emailing the authors and searching through the dataverses of the journals. (Again, we thank all authors who generously provided their replication data.) We excluded an additional 18 studies due to a lack of replication materials or an inability to replicate published findings, leaving a total of 22 studies from which we replicated 46 interaction effects. For studies that included multiple interaction effects, we focused on the most important ones which we identified as either: (1) those for which the authors generated a marginal effect plot of the sort Brambor, Clark and Golder (2006) recommends, or, (2) if no such plots were included, those which were most relied upon for substantive claims. We excluded interaction effects where the marginal effect was statistically insignificant

[23] Replication data can be found in Hainmueller, Mummolo and Xu (2018).

across the entire range of the moderator and/or where the authors did not claim to detect an interaction effect.[24]

While we cannot guarantee that we did not miss a relevant article, we are confident that our literature review has identified a large portion of recent high-profile political science studies employing this modeling strategy and claiming an interaction effect.[25] The articles cover a broad range of topics and are drawn from all empirical subfields of political science. Roughly 37% percent of the interaction effects are from the APSR, 20% are from the AJPS, 22% are from CPS, 15% are from IO, and 7% are from JOP, respectively.

There are at least three reasons why the conclusions from our sample might provide a lower bound for the estimated share of published studies where the assumptions of the standard multiplicative interaction model do not hold. The first one is that we only focus on top journals. Second, for three journals we focus exclusively on the studies that cite Brambor, Clark and Golder (2006) and therefore presumably took special care to employ and interpret these models correctly. Third, we restrict our sample to the subset of potentially more reliable studies where the authors made replication data available and where we were able to successfully replicate the results.[26]

It is important to emphasize that our replications and the conclusions that we draw from them are limited to reanalyzing the main models that underlie the interaction effect plots and tables presented in the original studies. Given the methodological focus of our article, we do not consider any additional evidence that the authors might have presented in their original studies to corroborate their substantive claims. Readers should keep this caveat in mind and consult the original studies and replication data to

[24] We cap the number of replicated interactions at four per study. In the rare cases with more than four interaction plots we chose the four most important ones based on our reading of the article.

[25] According to Google scholar our replicated studies have been cited nearly 1,900 times as of December 2016. The mean number of citations per article was roughly 86.

[26] In addition, given the problems arising from a lack of common support, demonstrated below, the decision to exclude triple interactions from our sample—which, all else equal, are more susceptible to the common support problem—likely removed several problematic cases from our analysis.

judge the credibility of the original claims in light of our replication results. Moreover, we emphasize that our results should not be interpreted as accusing any scholars of malpractice or incompetence. We remind readers that the authors of the studies that we replicate below were employing the accepted best practices at the time of publication, but following these existing guidelines for interaction models did not alert them to the problems that we describe below.

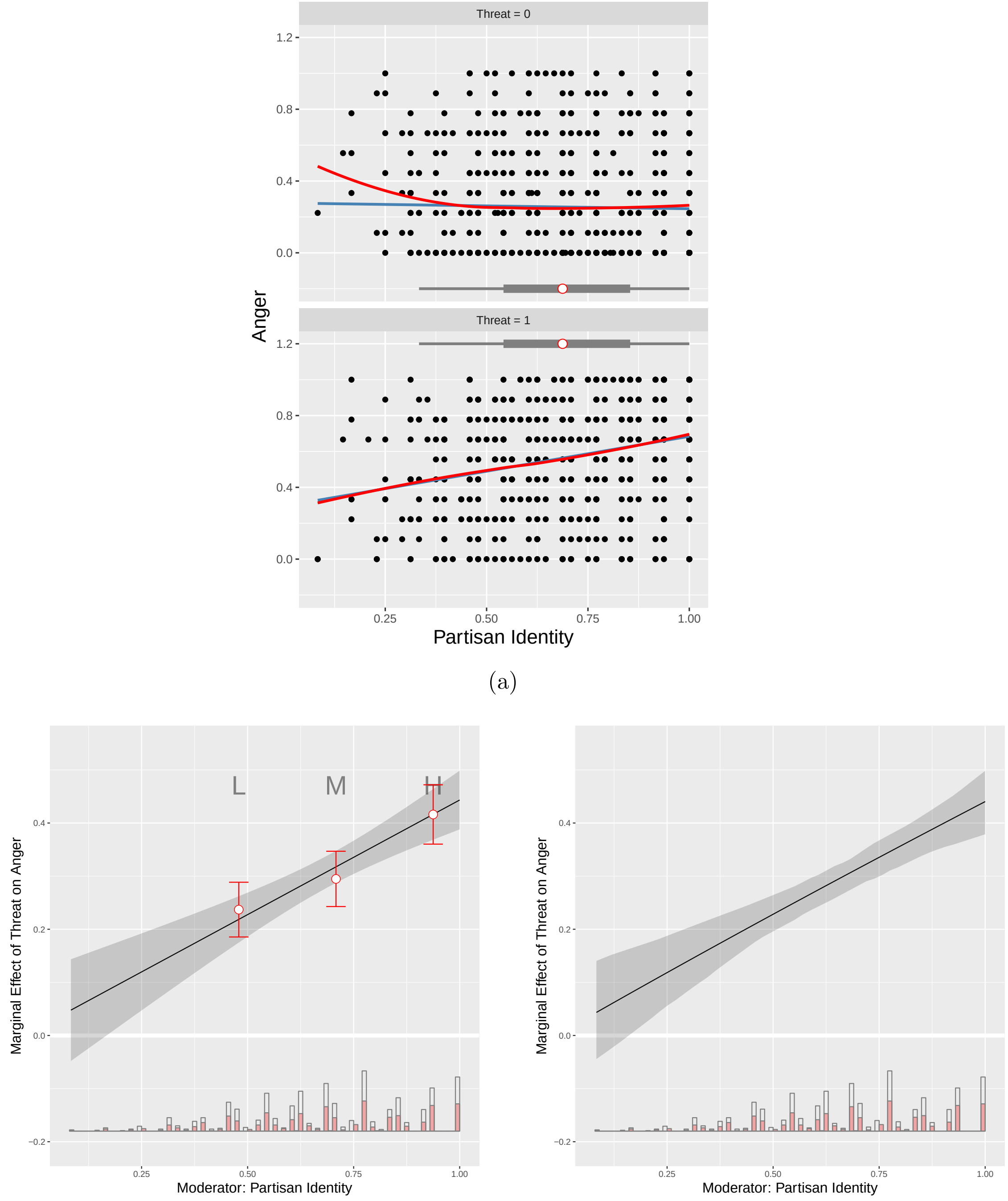
6 Results

Case 1: Linear Marginal Effects

We begin our discussion with a replication of Huddy, Mason and Aarøe (2015), an example of a study in which the assumptions of the multiplicative interaction model appear to hold well. This study uses a survey experiment and a multiplicative interaction model to test the hypothesis that a threat of electoral loss has a larger effect on anger if respondents are stronger partisan identifiers. The outcome is anger, the treatment is the threat of electoral loss (binary yes/no), and the moderator is the partisan identity of the respondent (continuous scale, 0 to 1). The key finding is that "Strongly identified partisans feel angrier than weaker partisans when threatened with electoral loss" (Huddy, Mason and Aarøe 2015, pg. 1).

The upper panel in Figure 4 displays our diagnostic scatterplot applied to this data. We see that the relationship between anger and partisan identity is well approximated by a linear fit in both groups with and without threat, as the linear and LOESS lines are close to each other. This provides good support for the validity of the LIE assumption in this example. There seems to be a linear interaction, with the effect of threat on anger increasing with higher levels of partisan identity. In addition, the boxplots suggests that there is sufficient common support for the range of partisan

Figure 4. Linear Interaction Effect: Replication of Huddy, Mason and Aarøe (2015)



(a)

(b)

(c)

Note: The above plots examine the marginal effects plot in Huddy, Mason and Aarøe (2015): (a) linear interaction digonostic plot; (b) marginal effects estimates from the replicated model (black line) and the binning estimator (red dots); (c) marginal effects estimates from the kernel estimator.

identity between about .3 to 1.

The middle panel in Figure 4 displays the conditional marginal effects estimates of our binning estimator superimposed on the estimates from the multiplicative interaction model used by the authors. As expected given the scatterplot, the conditional marginal effect estimates of the binning estimator for the threat effect at low, medium, and high levels of partisan identity line up very closely with the linear interaction effects from the original model. The threat effect is almost twice as large at high compared to low levels of partisan identity and the difference between these two effects is statistically significant ($p < 0.0001$). The threat effect at medium levels falls about right in between the low and high estimates. In addition, the stacked histogram at the bottom again corroborates that there is sufficient common support with both treated and control observations across a wide range of values of the moderator. The lower right panel in Figure 4 presents the conditional marginal effects estimates from the kernel estimator. The optimal bandwidth selected by cross-validation is relatively large. The result from the kernel estimation shows that the LIE assumption is supported by the data. The magnitude of the threat effect increases at an approximately constant rate with higher partisan identity.

Case 2: Lack of Common Support

The next example illustrates how the linear interaction model can mask a lack of common support in the data, which can occur when the treatment fails to vary across a wide range of values of the moderator. Chapman (2009) examines the effect of authorizations granted by the U.N. Security Council on public opinion of U.S. foreign policy, positing that this effect is conditional on public perceptions of member states' interests. The outcome is the number of "rallies" (short term boosts in public opinion), the treatment is the granting of a U.N. authorization (binary yes/no) and the moderator is the preference distance between the U.S. and the Security Council (continuous scale, -1 to 0). In Figure 2 in the study, the author plots the marginal effect of U.N. authorization, and states, "[c]learly, the effect of authorization on rallies decreases as similarity increases," (p. 756).

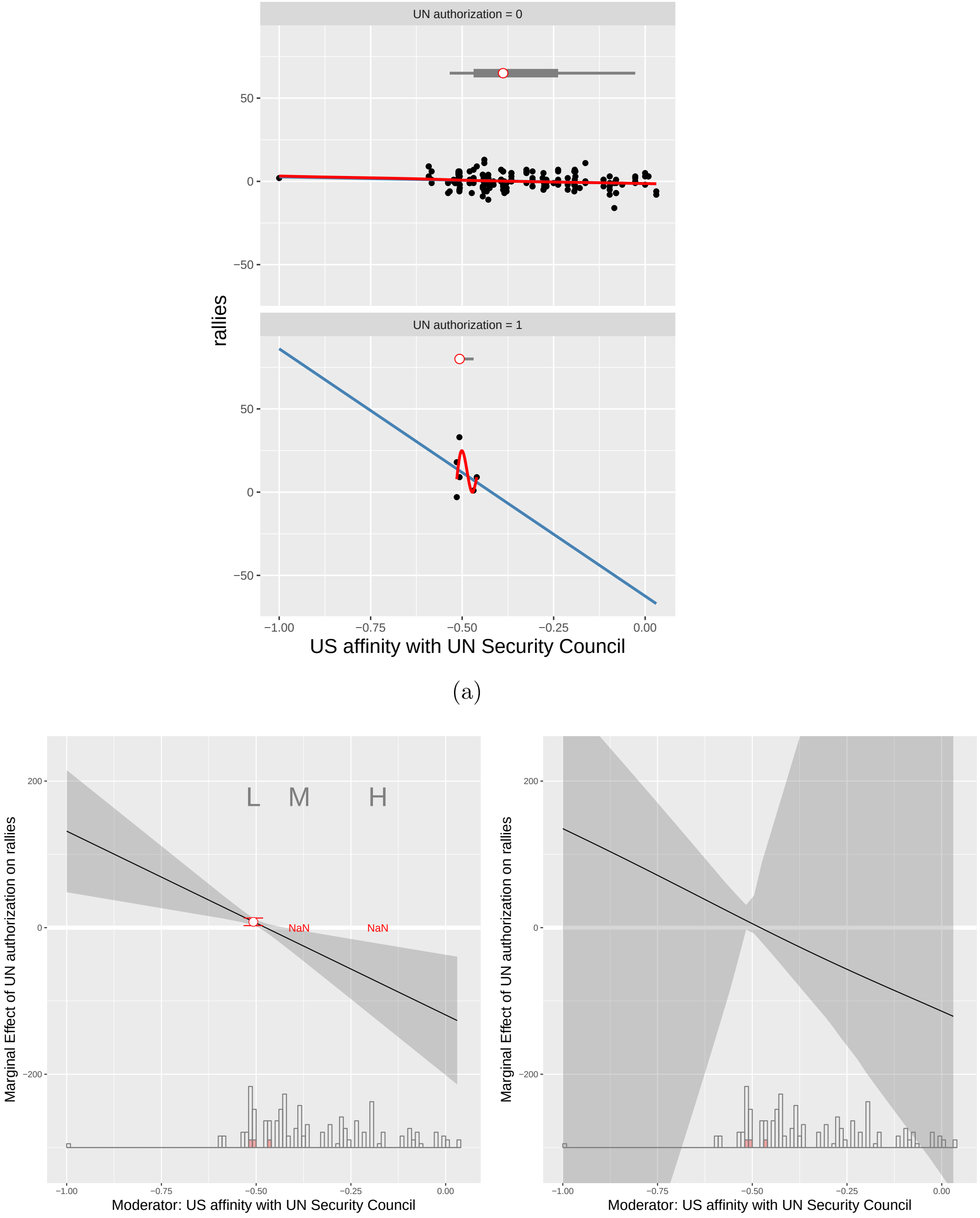
The upper panel in Figure 5 shows our diagnostic scatterplot for this model and the lower left panel in Figure 5 reproduces the original plot displayed in the study (Figure 2) but overlays the estimates from our binning estimator for low, medium, and high values of the moderator. Again, in the latter plot the stacked histogram at the bottom shows the distribution of the moderator in the treatment and control group with and without U.N. authorization, respectively.

As the plots show, there is a dramatic lack of common support. There are very few observations with a U.N. authorization and those observations are all clustered in a narrow range of moderator values of around -.5. In fact, as can be seen in the histogram at the bottom of the plot in the lower panel, or in the boxplots in the upper panel of Figure 5, all the observations with a U.N. authorization fall into the lowest tercile of the moderator and the estimated marginal effect in this lowest bin is close to zero. In the medium and high bin, the effect of the U.N. authorizations cannot be estimated using the binning estimator because there is zero variation on the treatment variable for values of the moderator above about -.45.

The common practice of simply fitting the standard multiplicative interaction model and computing the conditional marginal effects from this model will not alert the researcher to this problem. Here the effect estimates from the standard multiplicative interaction model for values of the moderator above -.45 or below -.55 are based purely on extrapolation that relies on the specified functional form, and are therefore highly model dependent and fragile.[27] This model and data cannot reliably answer the re-

[27] In footnote 87, the author writes: "Note that the graph suggests rallies of greater than 100 percent change in approval with authorization and an S score close to -1. However, authorization occurs in the sample when the S score is between -.6 and -.4, meaning that predictions outside this interval

FIGURE 5. LACK OF COMMON SUPPORT: CHAPMAN (2009)



(a)

(b)

(c)

Note: The above plots examine the marginal effects plot in Chapman (2009): (a) linear interaction diagnostic plot; (b) marginal effects estimates from the replicated model (black line) and the binning estimator (red dots); (c) marginal effects estimates from the kernel estimator.

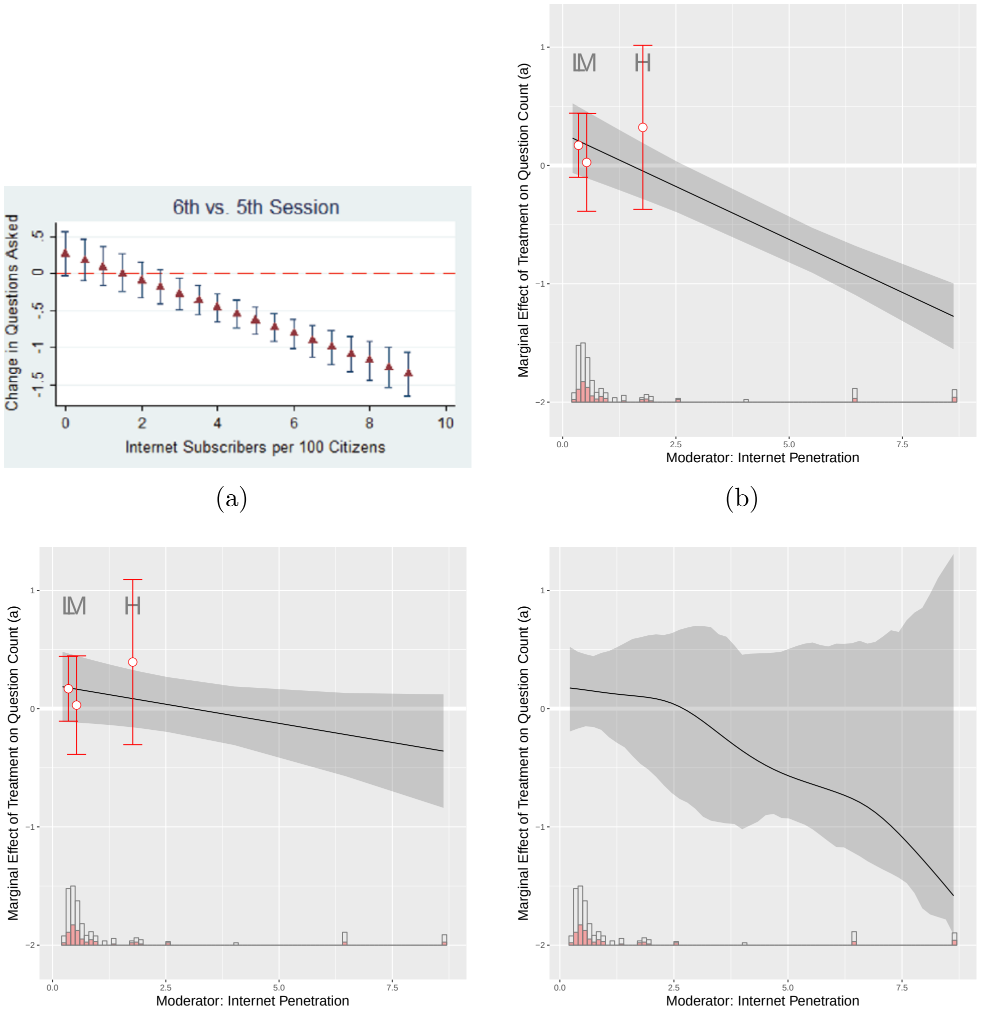
are made with less confidence. This is a drawback of generating predictions based on the small number of authorizations. A more realistic interpretation would suggest that authorizations should exhibit decreasing marginal returns at extreme values of S," (Chapman 2009, p. 756).

search question without heroic assumptions as to how the effect of U.N. authorizations varies across the preference distance between the U.S. and the Security Council because the very few cases with and without authorizations are all concentrated in the narrow range of the moderator around -.5, while for other moderator values there is no variation in the treatment. This becomes yet again clear in the marginal effects estimates from the kernel estimator (with a relatively large bandwidth chosen via cross-validation) displayed in the lower right panel of Figure 5. Once we move outside the narrow range where there is variation on the treatment, the confidence intervals from the marginal effects estimates blow up indicating that the effect simply cannot be estimated given the lack of common support. This shows the desired behavior of the kernel estimator in alerting researchers to the problem of lack of common support.

Case 3: Severe Interpolation

The next published example illustrates how sparsity of data in various regions of a skewed moderator (as opposed to no variation at all in the treatment) can lead to misspecification bias. Malesky, Schuler and Tran (2012) use a field experiment to examine whether legislative transparency interventions that have been found to have positive effects on legislator performance in democratic contexts produce the same benefits when exported to countries with authoritarian regimes. To this end the researchers randomly selected a subgroup of Vietnamese legislators for a transparency intervention which consisted of an online newspaper publishing a profile about each legislator that featured transcripts and scorecards to document that legislator's performance in terms of asking questions, critical questions in particular, in parliament. While the transparency intervention had no effect on average, the authors argue that the response of delegates to this transparency intervention is conditional on the level of internet penetration in their province. To test this they regress the outcome, measured as the change in the number of questions asked by the legislator, on the treatment, a binary dummy for whether legislators were exposed to the transparency intervention or not, the moderator, measured as the number of internet subscribers per 100 citizens in the province, and the interaction between the two (Table 5 in the original study).

FIGURE 6. SEVERE INTERPOLATION: MALESKY, SCHULER AND TRAN (2012)



(a) (b)

(c) (d)

Note: The above plots examine the marginal effects plot in Malesky, Schuler and Tran (2012): (a) the authors' original plot; (b) marginal effects estimates from the replicated model (black line) and the binning estimator (red dots); (d) marginal effects estimates from the binning estimator after dropping 4 influential observations; (d) marginal effects estimates from the kernel estimator.

Figure 6(a) reprints the marginal effect plots presented by the authors in Figure 1 of their article which is based on plotting the conditional marginal effects from the standard multiplicative interaction model that they fit to the data. They write: "[t]he graphs show clearly that at low levels of Internet penetration, the treatment has no impact on delegate behavior, but at high levels of Internet penetration, the treatment effect is large and significant" (pg. 17). Based on this negative effect at higher levels of Internet penetration the authors conclude that, "delegates subjected to high treatment intensity demonstrate robust evidence of curtailed participation [...]. These results make us cautious about the export of transparency without electoral sanctioning," (Malesky, Schuler and Tran 2012, pg. 1).

Figure 6(b) displays the marginal effect estimates from our replication of the original model and the binning estimator. Our replication plots show two critical concerns. First, the effect of the transparency intervention appears non-monotonic and non-linear in the moderator. In fact, the point estimates from the binning estimator grow smaller between typical low and typical medium levels of Internet penetration, but then larger between typical medium and typical high levels of Internet penetration. None of the three estimates are significant suggesting that the transparency intervention had no significant effect at either typical low, medium, or high levels of Internet penetration as measured by the median values in the low, medium, and high terciles.[28] This suggests that the LIE assumption employed in the original model does not hold and when relaxed by the binning estimator there is no compelling evidence of a negative interaction effect.

Second, as illustrated by the stacked histogram and the placement of the binned estimates (which lie at the median of Internet penetration in each bin), there are very few observations which exhibit levels of Internet penetration higher than about 2.5, which is the point above which the effect of the transparency intervention starts to become significant according to the original model.[29] In fact, for the range between 2.5

[28] The medians of the three terciles, 0.35, 0.53, 1.77, refer to the 17th, 50th, and 83rd percentile of the moderator, respectively.

[29] There are 4 observations at an Internet penetration of 4.07, 22 at 6.47, and 20 at 8.63. Together

and 9, where the original model suggests a negative effect, there is very little data and the results are based on severe interpolation of the likely incorrect linear functional form to an area far outside the bulk of the data (see Anderson (2013) for a similar critique).[30] The linear downward trend that underlies the claim of a negative interaction is entirely driven by the outliers with extremely high levels of Internet penetration that occur in two metropolitan areas. This severe interpolation suggests that the estimates from this model are highly model dependent and fragile. To diagnose the robustness of the estimates we investigated how many leverage points need to be dropped before the findings change considerably when using the original misspecified model. The result of this robustness check is shown in Figure 6(c), where we see that once only four extreme leverage points——which make up less than 0.9% of all observations—are removed from the data the effect estimates from the original interaction model flatten, indicating no effect of the intervention at any level of Internet penetration.[31]

Figure 6(d) shows the marginal effects estimates from our kernel estimator. In Figure 6(d), we use block bootstrap to obtain the uncertainty estimates. The confidence intervals are much wider than those in Figures 6(b) and (c), which are based on cluster-robust standard errors. This is because when the number of clusters is relatively small (in this case, 64 in total, and much fewer in the right tail), cluster-robust standard errors can severely underestimate the uncertainty (Cameron and Miller 2015).

these make up less than 10% of all observations.

[30] Our replications below show that the same problem applies to all the other three outcomes used by Malesky, Schuler and Tran (2012) in their Figure 1.

[31] Note that the unit of analysis in the original analysis is the delegate, who is exposed or not exposed to the intervention, while the moderator Internet penetration is measured at the level of the province. In our robustness check we drop only four delegates, but keep all provinces in the data including all the metropolitan areas with extreme values of Internet penetration. Also note that we have no theoretical rationale to drop these data points. This is merely a robustness check to demonstrate to readers the fragility of estimates from a linear interaction model that relies on severe interpolation.

Case 4: Nonlinearity

Our next example underscores how fitting linear interaction models can mask nonlinearities in interaction effects and therefore result in severe misspecification bias. Clark and Golder (2006) argue that the temporal proximity of presidential elections affects the number of parties that compete in an election, but that this effect is conditional on the number of presidential candidates. After estimating a linear interaction model, the authors plot the marginal effect in Figure 2 in their paper, which we replicated in the left plot of Figure 7, again superimposing the estimates from our binning estimator where we use four bins to discretize the moderator. The authors interpret their linear interaction effect estimates by writing that, "[i]t should be clear that temporally proximate presidential elections have a strong reductive effect on the number of parties when there are few presidential candidates. As predicted, this reductive effect declines as the number of candidates increases. Once the number of presidential candidates becomes sufficiently large, presidential elections stop having a significant effect on the number of parties," (Clark and Golder 2006, pg. 702).[32]

But as the estimates from the binning estimator in Figure 7 show, the story is more complicated. In fact, the moderator is highly skewed and for 59% of observations takes on the lowest value of zero. Moreover, as in the Chapman (2009) example above, there is no variation at all on the treatment variable in this first bin where the moderator takes on the value of zero, such that the treatment effect at this point is not even identified given the absence of common support. This directly contradicts the original claim of a strong negative effect when there is a low effective number of candidates. And rather than evidencing a positive interaction, as the study claims, the effect is insignificant in the second bin, but then rapidly drops to be negative and significant at the third bin, only to increase again back to zero in the last bin.[33] The LIE assumption

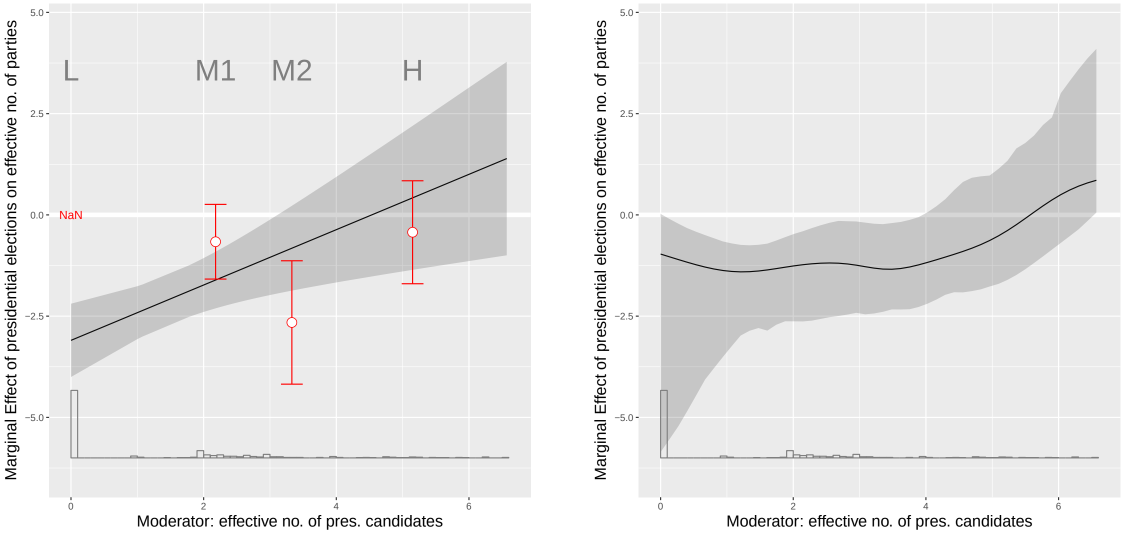
[32] Note that this marginal effect plot also appears in Brambor, Clark and Golder (2006).

[33] We split the remaining observations where the moderator is not zero into bins defined by $(0, 3)$,

does not hold and accordingly the linear interaction model is misspecified and exhibits a lack of common support for the majority of the data.

This is confirmed by the effect estimates from the kernel estimator which are shown in the right plot in Figure 7.[34] Consistent with the binning estimates, the marginal effect appears nonlinear and the confidence intervals blow up as the moderator approaches zero given that there is no variation in the treatment variable at this point. Contrary to the authors' claims, the number of candidates in an election does not appear to moderate the effect of proximate elections in a consistent manner and the effect is not identified for a majority of the data due to the lack of common support.

Figure 7. Nonlinearity: Clark and Golder (2006)



(a)

(b)

Note: The above plots examine the marginal effects plot in Clark and Golder (2006): (a) marginal effects estimates from the replicated model (black line) and the binning estimator (red dots); (b) marginal effects estimates from the kernel estimator.

$[3, 4)$ and $[4, 7]$ such that the binning estimates well represent the entire range.

[34] Because of the extreme skew in the distribution of the moderator which only sparsely overlaps with the treatment, we manually chose a bandwidth of 1 when employing the kernel estimator in this example.

Summary of Replications

The previous cases highlight stark examples of some of the issues that can go undiagnosed if the standard linear interaction model is estimated and key assumptions go unchecked. But how common are such problems? How much should we trust published estimates from multiplicative interaction models in political science? To investigate this question we replicated 46 interaction effects from our sample of published work in the top five political science journals. To rank these cases, we constructed a simple additive scoring system whereby cases were allocated single points for exhibiting (1) no statistically different treatment effects at typical low and typical high levels of the moderator, (2) severe extrapolation, and (3) nonlinear interaction effects.

We determined the first criterion by testing whether the marginal effect estimate from the binning estimator at the median value in the low tercile of the moderator was statistically different from the effect estimate at the median of the high tercile of the moderator ($p < .05$, two-tailed). This criterion provides a formal test of the extent to which the data actually contains evidence of a significant interaction effect once we relax the stringent LIE assumption that underlies the claim of a significant interaction in the original study.

We determined the second criterion of severe extrapolation by examining whether the L-Kurtosis of the moderator exceeds a threshold that indicates severe extrapolation. The L-Kurtosis is a robust and efficient measure of the degree to which the shape of the distribution is characterized by outliers[35] and therefore captures to what extent the estimates reported in the marginal effect plots are based on extrapolation to moderator values where there is little or no data.[36]

[35] The L-Kurtosis is based on linear combination of order statistics and is therefore less sensitive to outliers and has better asymptotic properties than the classical kurtosis (Hosking 1990).

[36] For example, in the case of Huddy, Mason and Aarøe (2015) the moderator has an L-Kurtosis of .065 which is half way between a normal distribution (L-Kurtosis=.12) and a uniform distribution (L-Kurtosis=0) and therefore indicates good support across the range of the moderator. 80% of the density is concentrated in about 53% of the interval reported in the marginal effects plot. In

Finally, to determine whether the interaction effect is indeed linear as claimed in the original study, we re-parameterize Model (4) as

$$\begin{aligned} Y =& \mu + \alpha D + \eta X + \beta DX + G_2(\mu_{2'} + \alpha_{2'} D + \eta_{2'} X + \beta_{2'} DX) \\ &+ G_3(\mu_{3'} + \alpha_{3'} D + \eta_{3'} X + \beta_{3'} DX) + Z\gamma + \epsilon \end{aligned}$$

such that the new model nests Model (1). We then test the null that the eight additional parameters are jointly equal to zero (i.e., $\mu_{2'} = \alpha_{2'} = \eta_{2'} = \beta_{2'} = \mu_{3'} = \alpha_{3'} = \eta_{3'} = \beta_{3'} = 0$) using a standard Wald test. This criterion provides a formal test of whether the linear interaction model used in the original study can be rejected in favor of the more flexible binning estimator model that relaxes the LIE assumption. If we rejected the null, we obtained a piece of evidence against the linear interaction model. Hence, we allocated one point to the case for a nonlinear interaction effect. However, it is worth noting that failing the reject the null does *not* necessarily mean that the LIE assumption holds, especially when the sample size is small and the test is underpowered. We therefore regard this coding decision as lenient. Taken together, failing these three tests indicates that marginal effect estimates based on a linear interaction model are likely to produce misleading results.

In addition to our scoring system we also display more complete analyses of each case in the Online Appendix B so that readers may examine them in more detail and come to their own conclusions.

Table 1 provides a numerical summary of the results and Figure 8 displays the marginal effects from the binning estimator superimposed on the original marginal effect estimates from the replicated multiplicative interaction models used in the original studies. In all, only 4 of the 41 cases where the data were sufficient to conduct all three

stark contrast, in the case of Malesky, Schuler and Tran (2012) the moderator has an L-Kurtosis of .43 which indicates severe extrapolation. In fact, about 80% of the density of the moderator is concentrated in a narrow interval that only makes up 11% of the range of the moderator over which the marginal effects are plotted in the study. We code studies where the L-Kurtosis exceeds .16 as exhibiting severe extrapolation. This cut-point roughly corresponds to the L-Kurtosis of an exponential or logistic distribution.

tests[37] (9.8%) received a perfect score of zero indicating that the reported marginal effects meet all three criteria of differential treatment effects across the low and high levels of the moderator, no severe extrapolation, and linearity. This is an unnervingly low fraction for a sample that consists only of top journal publications. Twelve cases (29.3%) received a score of 1, while 18 cases (43.9%) received a score of 2. Seven cases (17.1%) received a score of three, failing to pass a single one of the three tests.[38] We also find that there is considerable heterogeneity in the scores for interactions that are reported in the same article suggesting that checks for the linear interaction assumption and common support are not consistently applied.[39]

Table 1. Replication Results by Journal

Journal	N	Not Rejecting Same Effect at Low vs. High	Severe Extra-polation	Rejecting Linear Model	Mean Score
AJPS	9	0.78	0.78	0.22	1.8
APSR	17	0.71	0.41	0.65	1.8
CPS	10	0.67	0.3	0.5	1.3
IO	7	0.83	0.71	0.67	2.3
JOP	3	0.33	0.33	0.67	1.3
All Journals	46	0.71	0.50	0.52	1.7

The table displays the mean for each criterion for each journal, as well as the mean additive score for each journal. The unit of analysis is the interaction, not the article. Note that only 44 cases and 42 cases are used for the low vs. high and linearity tests, respectively, due to data limitations that prevented these tests.

Once we break out the results by journal, we find that the issues raised by our review are not unique to any one subfield or journal in political science. The Journal of Politics (JOP) and Comparative Political Studies (CPS) received the lowest (best)

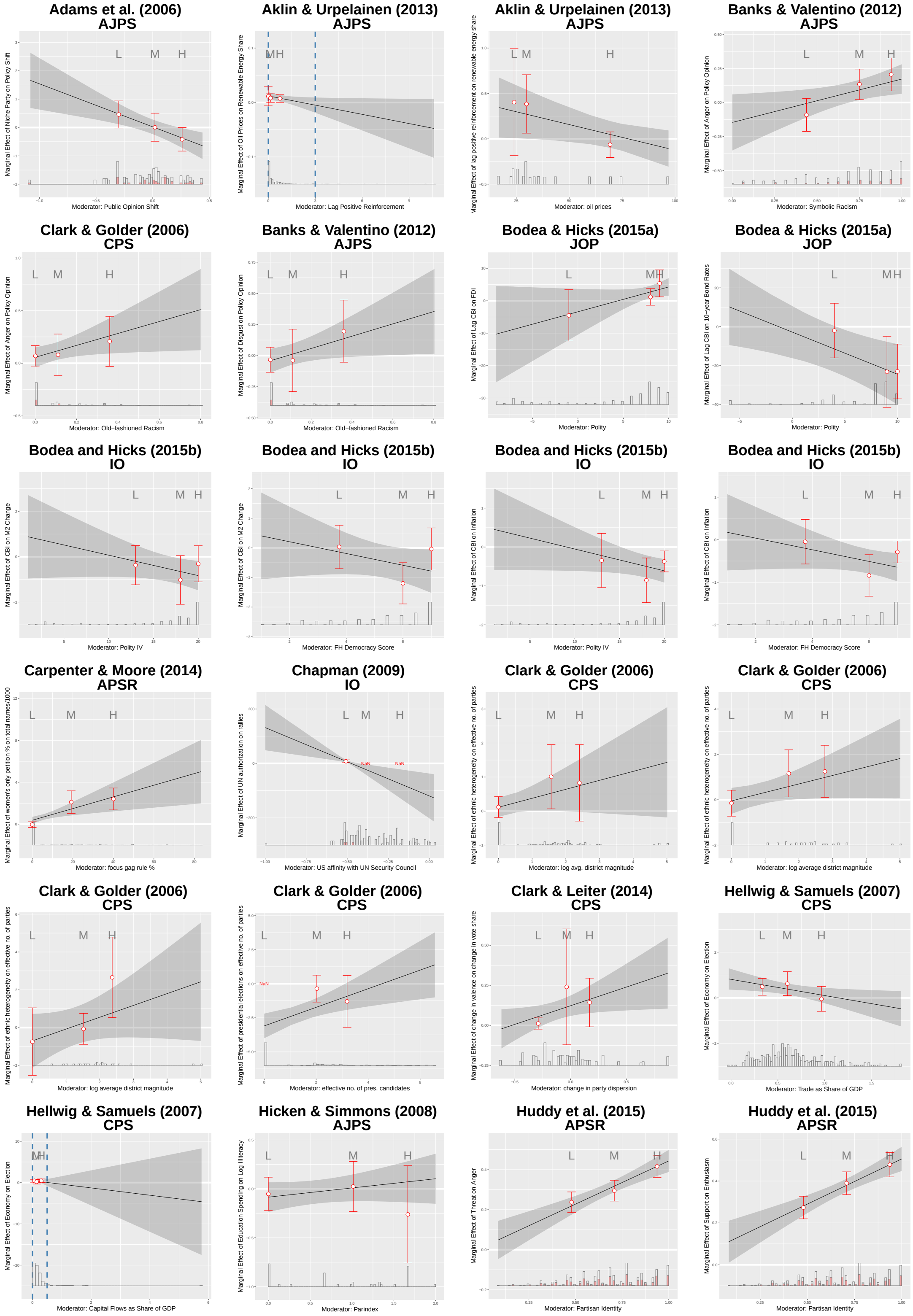
[37] In two cases we could not test for equality in the marginal effects at low and high levels of the moderator due to a lack of common support. In other cases a singular variance-covariance matrix precluded Wald tests for linearity. Dropping these cases rather than scoring them as failing the test likely improved these aggregate scores.

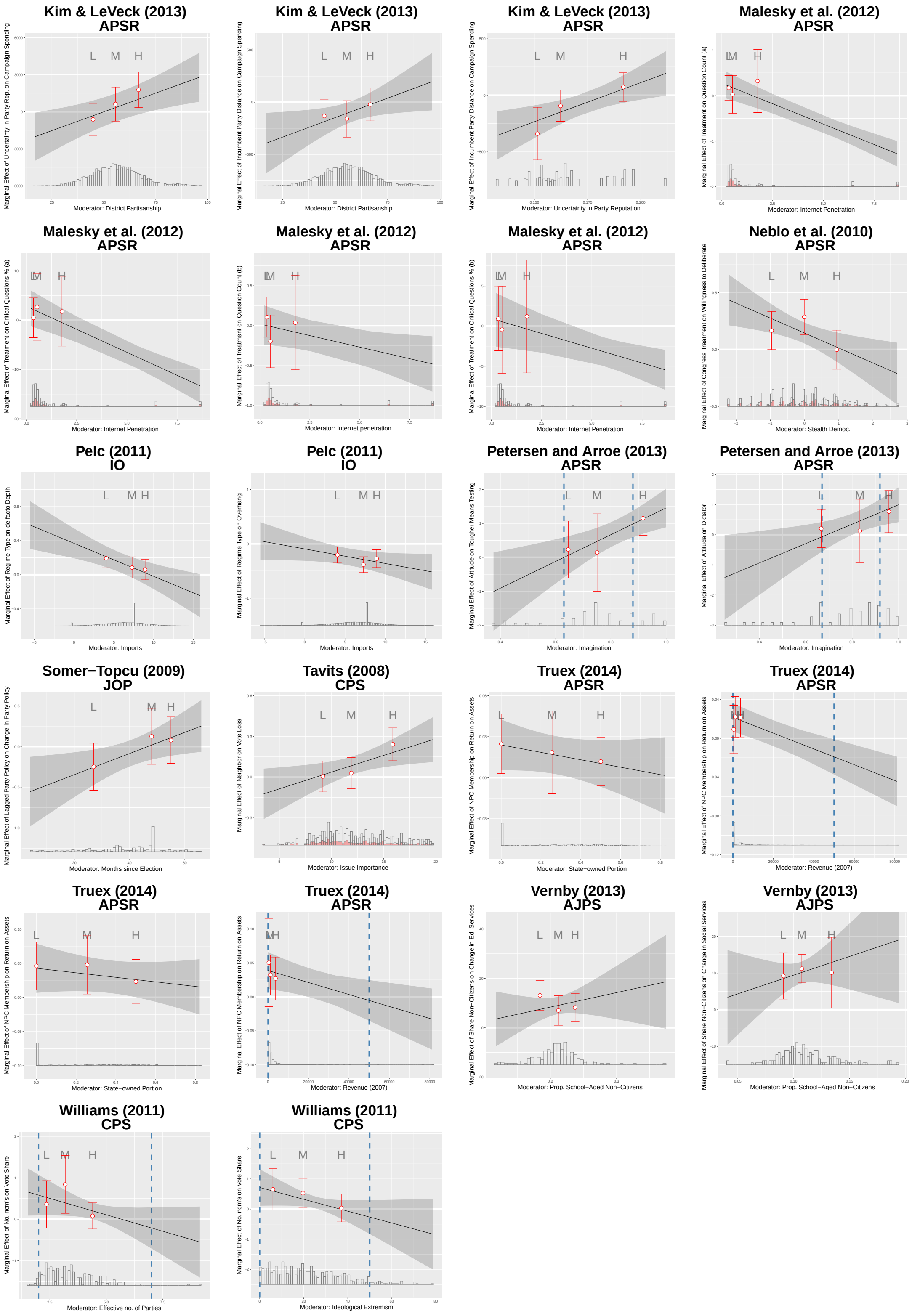
[38] For details, see Table A1 in the Appendix.

[39] In all, only 30% of the interaction effects we examine allow us to reject the null of identical marginal effects in the first and third terciles of the moderator (i.e. the low vs. the high bins) at the 5% significance level. Lowering the significance threshold to the 10% and 25% levels leads us to reject the null in 34% and 55% of cases, respectively. Note that two cases where a lack of data prevented us from conducting this t-test were dropped and are not included in these calculations. See Online Appendix for a full list of p-values from these tests.

overall mean scores, 1.3 on our 0-to-3 scale, while APSR and AJPS tied for second with scores of 1.8. The highest (worst) score was 2.3 for IO. The mean scores here are computed using a small number of cases, and so their precision could rightly be questioned. Still, given that our sample is restricted to work published only in top political science journals, these results indicate that many of the most substantively important findings in the discipline involving interaction effects in recent years may be modeling artifacts, and highlight a need for improved practices when employing multiplicative interaction models.

Figure 8. The assumptions of the linear interaction model rarely hold in published political science work





Note: The blue dashed vertical lines indicate the range of the moderators displayed in the original manuscripts.

Conclusion

Multiplicative interaction models are widely used in the social sciences to test conditional hypotheses. While empirical practice has improved following the publication of Brambor, Clark and Golder (2006) and related advice, this study demonstrates that there remain serious problems that are overlooked by scholars using the existing best practice guidelines. In particular, the multiplicative interaction model implies the key assumption that the interaction effects are linear, but our replications of published work in five top political science journals suggests that this assumption often does not hold in practice. In addition, as our replications also show, scholars often compute marginal effects in areas where there is no or only very limited common support, which results in fragile and model dependent estimates.

To improve empirical practice we develop a simple diagnostic that allows researchers to detect problems with the linear interaction effects assumption and/or lack of common support. In addition, we propose more flexible estimation strategies based on a simple binning estimator and a kernel estimator that allow researchers to estimate marginal effects without imposing the stringent linear interaction assumption while safeguarding against extrapolation to areas without (or with limited) common support. When applying these methods to our replications, we find that the key findings often change substantially. Given that our sample of replications only includes top journal articles, our findings here most likely understate the true extent of the problem in published work in political science. Overall, our replications suggest that a large portion of published findings employing multiplicative interaction models with at least one continuous variable are based on modeling artifacts, or are at best highly model dependent, and suggest a need to augment the current best practice guidelines.[40]

[40] Consistent with this pattern, Collaboration et al. (2015) who replicated 100 studies published in three psychology journals found that the replication success rate for significant effects was much lower for studies that tested interaction effects (22% replicated) compared to studies that tested main or simple effects (47% replicated).

We recommend that researchers engaged in modeling interaction effects and testing conditional hypotheses should engage in the following:

1. **Checking the raw data**. Generate the *Linear Interaction Diagnostic* plot using the raw data to check whether the conditional relationships between the outcome, treatment, and moderator are well approximated by a linear fit and check whether there is sufficient common support to compute the treatment effect across the values of the moderator. If additional covariates are involved in the model, the same diagnostic plots can be constructed after residualizing with respect to those covariates. If both the treatment and the moderator are continuous, a GAM plot can be used to further assist with these checks (see Appendix for details on GAM plots). Given the symmetry of interaction models, we also recommend that the diagnostic plots are constructed two ways to examine the marginal effects of $D|X$ and of $X|D$, as linearity is implied for both in the standard model. If the distribution of the variables are highly skewed and/or asymmetric we recommend that researchers use appropriate power and or root transformations to reduce skewness, increase symmetry, and aid with linearizing the relationships between the variables (Mosteller and Tukey 1977).

2. **Applying the binning estimator**. Compute the conditional marginal effects using the binning estimator. In our experience, three equal sized bins for each tercile with the evaluation points set to the bin medians provide a reasonable default to get a good sense of the effect heterogeneity. More bins should be used if more detail is required and more data is available. The number of bins could be pre-specified in a pre-analysis plan to reduce subjectivity. Close attention should be paid to not compute marginal effects in areas where the data is too sparse either because there are no observations for those values of the moderator or there is no variation in the treatment. To aid with this we recommend to

add a (stacked) histogram at the bottom of the marginal effect plot to show the distribution of the moderator and detect problems with lack of common support.

3. **Applying the kernel estimator**. In addition, generating the marginal effects estimates using the kernel estimator is recommended to further evaluate the effect heterogeneity and relax the linearity assumption on the covariates. Researchers may also use other machine learning methods to gauge how treatment effects vary across different subgroups or levels of a moderator,[41] but we believe that the kernel methods strike a good balance between model complexity and interpretability, as well as accessibility to applied researchers.

4. **Be cautious when applying the linear interaction model.** The standard linear interaction model and marginal effects plots should only be used if the estimates from the binning and or kernel estimator suggest that the interaction is really linear, and marginal effects should only be computed for areas with sufficient common support. If a standard linear interaction model is used in this case, the researchers should follow the existing guidelines as described in Brambor, Clark and Golder (2006) and related advice.

Following these revised guidelines would have solved the problems we discussed in the set of published studies that we replicated. Accordingly, we hope that applying these guidelines will lead to a further improvement in empirical practice.[42] That said, it is important to emphasize that following these revised guidelines does not guarantee that the model will be correctly specified. When other covariates are included in the model, it is important for researchers to apply all the usual regression diagnostics with respect to these covariates[43] in addition to the checks we proposed here. Moreover, it

[41] See, for examples, Imai and Ratkovic 2013; Grimmer, Messing and Westwood 2014; Athey and Wager 2016.

[42] We provide software routine `interflex` in both `R` (`https://goo.gl/cG8uwA`) and `STATA` (`https://goo.gl/OuYvLb`) to implement these diagnostic tests.

[43] Such as tests for linearity and the existence of outliers.

is important to recognize that the checks cannot help with other common problems such as endogeneity or omitted variables that often plague inferences from regression models and can often only be solved through better research designs.

References

Adams, James, Michael Clark, Lawrence Ezrow and Garrett Glasgow. 2006. "Are Niche Parties Fundamentally Different from Mainstream Parties? The Causes and the Electoral Consequences of Western European Parties' Policy Shifts, 1976–1998." *American Journal of Political Science* 50(3):513–529.

Aiken, Leona S., Stephen G. West and Raymond R. Reno. 1991. *Multiple regression: Testing and interpreting interactions.* Sage.

Aklin, Michaël and Johannes Urpelainen. 2013. "Political Competition, Path Dependence, and The Strategy of Sustainable Energy Transitions." *American Journal of Political Science* 57(3):643–658.

Anderson, James H. 2013. "Sunshine Works–Comment on'The Adverse Effects of Sunshine: A Field Experiment on Legislative Transparency in an Authoritarian Assembly'." *World Bank Policy Research Working Paper* (6602).

Athey, Susan and Stefan Wager. 2016. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association* .

Banks, Antoine J. and Nicholas A. Valentino. 2012. "Emotional Substrates of White Racial Attitudes." *American Journal of Political Science* 56(2):286–297.

Beck, Nathaniel, Gary King and Langche Zeng. 2000. "Improving quantitative studies of international conflict: A conjecture." *American Political Science Review* 94(01):21–35.

Berry, William D., Matt Golder and Daniel Milton. 2012. "Improving tests of theories positing interaction." *Journal of Politics* 74(3):653–671.

Bodea, Cristina and Raymond Hicks. 2015*a*. "International Finance and Central Bank Independence: Institutional Diffusion and the Flow and Cost of Capital." *The Journal of Politics* 77(1):268–284.

Bodea, Cristina and Raymond Hicks. 2015*b*. "Price Stability and Central Bank Independence: Discipline, Credibility, and Democratic Institutions." *International Organization* 69(1):35–61.

Brambor, Thomas, William Roberts Clark and Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14:63–82.

Braumoeller, Bear F. 2004. "Hypothesis Testing and Multiplicative Interaction Terms." *International organization* 58(04):807–820.

Cameron, A. Colin and Douglass L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *The Journal of Human Resources* 50(2):317–372.

Carpenter, Daniel and Colin D. Moore. 2014. “When Canvassers Became Activists: Antislavery Petitioning and the Political Mobilization of American Women.” *American Political Science Review* 108(3):479–498.

Chapman, Terrence L. 2009. “Audience Beliefs and International Organization Legitimacy.” *International Organization* 63(04):733–764.

Clark, Michael and Debra Leiter. 2014. “Does the Ideological Dispersion of Parties Mediate the Electoral Impact of Valence? A Cross-national Study of Party Support in Nine Western European Democracies.” *Comparative Political Studies* 47(2):171–202.

Clark, William Roberts and Matt Golder. 2006. “Rehabilitating Duverger’s Theory Testing the Mechanical and Strategic Modifying Effects of Electoral Laws.” *Comparative Political Studies* 39(6):679–708.

Cleveland, William S. and Susan J. Devlin. 1988. “Locally weiGhted Regression: An Approach to Regression Analysis by Local Fitting.” *Journal of the American Statistical Association* 83(403):596–610.

Collaboration, Open Science et al. 2015. “Estimating the reproducibility of psychological science.” *Science* 349(6251):aac4716.

Fan, Jianqing, Nancy E. Heckman and Matt P. Wand. 1995. “Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions.” *Journal of the American Statistical Association* 90(429):141.

Friedrich, Robert J. 1982. “In defense of multiplicative terms in multiple regression equations.” *American Journal of Political Science* pp. 797–833.

Grimmer, Justin, Solomon Messing and Sean J. Westwood. 2014. “Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods.”.

Hainmueller, Jens and Chad Hazlett. 2013. “Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach.” *Political Analysis* p. mpt019.

Hainmueller, Jens, Jonathan Mummolo and Yiqing Xu. 2018. “Replication Data for: How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice.” *Harvard Dataverse* . doi:10.7910/DVN/Q1VOOG.

Hastie, Trevor and Robert Tibshirani. 1986. “Generalized Additive Models.” *Statistical Science* 1(3):297–318.

Hellwig, Timothy and David Samuels. 2007. “Voting in Open Economies The Electoral Consequences of Globalization.” *Comparative Political Studies* 40(3):283–306.

Hicken, Allen and Joel W. Simmons. 2008. "The Personal Vote and the Efficacy of Education Spending." *American Journal of Political Science* 52(1):109–124.

Hosking, J. R. M. 1990. "L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics." *Journal of the Royal Statistical Society. Series B (Methodological)* 52(1):105–124.

Huddy, Leonie, Lilliana Mason and Lene Aarøe. 2015. "Expressive Partisanship: Campaign Involvement, Political Emotion, and Partisan Identity." *American Political Science Review* 109(01):1–17.

Imai, Kosuke, Luke Keele and Teppei Yamamoto. 2010. "Identification, inference and sensitivity analysis for causal mediation effects." *Statistical science* pp. 51–71.

Imai, Kosuke and Marc Ratkovic. 2013. "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation." *Annals of Applied Statistics* 7(1):443–470.

Jaccard, James and Robert Turrisi. 2003. *Interaction effects in multiple regression.* Number 72 Sage.

Kam, Cindy D. and Robert J. Franzese Jr. 2007. *Modeling and Interpreting Interactive Hypotheses in Regression Analysis.* Ann Arbor: The University of Michigan Press.

Kim, Henry A. and Brad L. LeVeck. 2013. "Money, Reputation, and Incumbency in US House Elections, or Why Marginals Have Become More Expensive." *American Political Science Review* 107(3):492–504.

King, Gary and Langche Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14(2):131–159.

Li, Qi and Jeffrey S. Racine. 2010. "Smooth Varying-Coefficient Estimation and Inference for Qualitative and Quantitative Data." *Econometric Theory* 26:1–31.

Lin, Winston. 2013. "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman?s critique." *The Annals of Applied Statistics* 7(1):295–318.

Malesky, Edmund, Paul Schuler and Anh Tran. 2012. "The Adverse Effects of Sunshine: A Field Experiment on Legislative Transparency in An Authoritarian Assembly." *American Political Science Review* 106(04):762–786.

Miratrix, Luke W, Jasjeet S Sekhon and Bin Yu. 2013. "Adjusting treatment effect estimates by post-stratification in randomized experiments." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75(2):369–396.

Mosteller, Frederick and John Wilder Tukey. 1977. "Data analysis and regression: a second course in statistics." *Addison-Wesley Series in Behavioral Science: Quantitative Methods* .

Neblo, Michael A., Kevin M. Esterling, Ryan P. Kennedy, David M.J. Lazer and Anand E. Sokhey. 2010. "Who Wants to Deliberate—and Why?" *American Political Science Review* 104(3):566–583.

Pelc, Krzysztof J. 2011. "Why do Some Countries Get Better WTO Accession Terms Than Others?" *International Organization* 65(4):639–672.

Petersen, Michael Bang and Lene Aarøe. 2013. "Politics in the Mind's Eye: Imagination as a Link between Social and Political Cognition." *American Political Science Review* 107(2):275–293.

Somer-Topcu, Zeynep. 2009. "Timely Decisions: The Effects of Past National Elections on Party Policy Change." *The Journal of Politics* 71(1):238–248.

Tavits, Margit. 2008. "Policy Positions, Issue Importance, and Party Competition in New Democracies." *Comparative Political Studies* 41(1):48–72.

Truex, Rory. 2014. "The Returns to Office in a 'Rubber Stamp' Parliament." *American Political Science Review* 108(2):235–251.

Vernby, Kåre. 2013. "Inclusion and Public Policy: Evidence from Sweden's Introduction of Noncitizen Suffrage." *American Journal of Political Science* 57(1):15–29.

Williams, Laron K. 2011. "Unsuccessful Success? Failed No-confidence Motions, Competence Signals, and Electoral Support." *Comparative Political Studies* 44(11):1474–1499.

Wood, Simon N. 2003. "Thin plate regression splines." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65(1):95–114.

A Supplementary Information

Appendix: Table of Contents

A.1 Proof

Model (1) and Model (4) in the main text are re-stated as follows:

$$Y = \mu + \eta X + \alpha D + \beta DX + \gamma Z + \epsilon; \qquad (1)$$

$$Y = \sum_{j=1}^{3} \{\mu_j + \alpha_j D + \eta_j (X - x_j) + \beta_j (X - x_j) D\} G_j + \gamma Z + \epsilon. \qquad (4)$$

It is to be proved that, if Model (1) is correct :

$$\hat{\alpha}_j - (\hat{\alpha} + \hat{\beta} x_j) \xrightarrow{p} 0, \qquad j = 1, 2, 3,$$

in which $\hat{\alpha}$ and $\hat{\beta}$ are estimated from Model (1) and $\hat{\alpha}_j$ are estimated from Model (4).

Proof: First, rewrite Model (4) as:

$$Y = \sum_{j=1}^{3} \{(\mu_j - \eta x_j) + \eta_j X + (\alpha_j - \beta_j x_j) D + \beta_j DX\} G_j + \gamma Z + \epsilon \qquad (6)$$

and define $\underline{\alpha}_j = \alpha_j - \beta_j x_j$. When Model (1) is correct, if we regress Y on G_j, XG_j, DG_j, XDG_j $(j = 1, 2, 3)$ and Z, we have:

$$\underline{\hat{\alpha}}_j \xrightarrow{p} \alpha \text{ and } \hat{\beta}_j \xrightarrow{p} \beta, \quad j = 1, 2, 3.$$

Since $\underline{\hat{\alpha}}_j = \hat{\alpha}_j - \hat{\beta}_j x_j$, we have: $\hat{\alpha}_j \xrightarrow{p} \alpha - \beta x_j$. Because

$$\hat{\alpha} \xrightarrow{p} \alpha \text{ and } \hat{\beta} \xrightarrow{p} \beta$$

when Model (1) is correct, we have:

$$\hat{\alpha}_j - (\hat{\alpha} + \hat{\beta} x_j) \xrightarrow{p} 0 \quad j = 1, 2, 3.$$

Q.E.D.

A.2 Additional information on replication files

Table A1. Replication Results

Study	Journal	Not Rejecting Same Effect at Low vs. High	Severe Extrapolation	Rejecting Linear Model	Overall Score
Adams et al. (2006)	AJPS	0	1	0	1
Aklin and Urpelainen (2013)	AJPS	1	1	1	3
Aklin and Urpelainen (2013)	AJPS	1	1	1	3
Banks and Valentino (2012)	AJPS	0	0	0	0
Banks and Valentino (2012)	AJPS	1	1	0	2
Banks and Valentino (2012)	AJPS	1	1	0	2
Bodea and Hicks (2015*a*)	JOP	0	0	1	1
Bodea and Hicks (2015*a*)	JOP	0	1	1	2
Bodea and Hicks (2015*b*)	IO	1	1	0	2
Bodea and Hicks (2015*b*)	IO	1	0	0	1
Bodea and Hicks (2015*b*)	IO	1	1	0	2
Bodea and Hicks (2015*b*)	IO	1	0	1	2
Carpenter and Moore (2014)	APSR	0	1	1	2
Chapman (2009)	IO	n.a.	1	0	n.a.
Clark and Golder (2006)	CPS	1	0	1	2
Clark and Golder (2006)	CPS	0	0	1	1
Clark and Golder (2006)	CPS	0	0	0	1
Clark and Golder (2006)	CPS	n.a.	1	1	n.a.
Clark and Leiter (2014)	CPS	1	1	0	2
Hellwig and Samuels (2007)	CPS	1	0	0	1
Hellwig and Samuels (2007)	CPS	1	1	0	2
Hicken and Simmons (2008)	AJPS	1	0	0	1
Huddy, Mason and Aarøe (2015)	APSR	0	0	0	0
Huddy, Mason and Aarøe (2015)	APSR	0	0	0	0
Kim and LeVeck (2013)	APSR	0	0	1	1
Kim and LeVeck (2013)	APSR	1	0	1	2
Kim and LeVeck (2013)	APSR	0	0	1	1
Malesky, Schuler and Tran (2012)	APSR	1	1	1	3
Malesky, Schuler and Tran (2012)	APSR	1	1	1	3
Malesky, Schuler and Tran (2012)	APSR	1	1	0	2
Malesky, Schuler and Tran (2012)	APSR	1	1	1	3
Neblo et al. (2010)	APSR	1	0	1	2
Pelc (2011)	IO	0	1	1	2
Pelc (2011)	IO	1	1	1	3
Petersen and Aarøe (2013)	APSR	1	0	0	1
Petersen and Aarøe (2013)	APSR	1	0	0	1
Somer-Topcu (2009)	JOP	1	0	0	1
Tavits (2008)	CPS	0	0	0	0
Truex (2014)	APSR	1	0	0	1
Truex (2014)	APSR	1	1	0	2
Truex (2014)	APSR	1	0	0	1
Truex (2014)	APSR	1	1	0	2
Vernby (2013)	AJPS	1	1	0	2
Vernby (2013)	AJPS	1	1	0	2
Williams (2011)	CPS	1	0	0	1
Williams (2011)	CPS	1	0	1	2

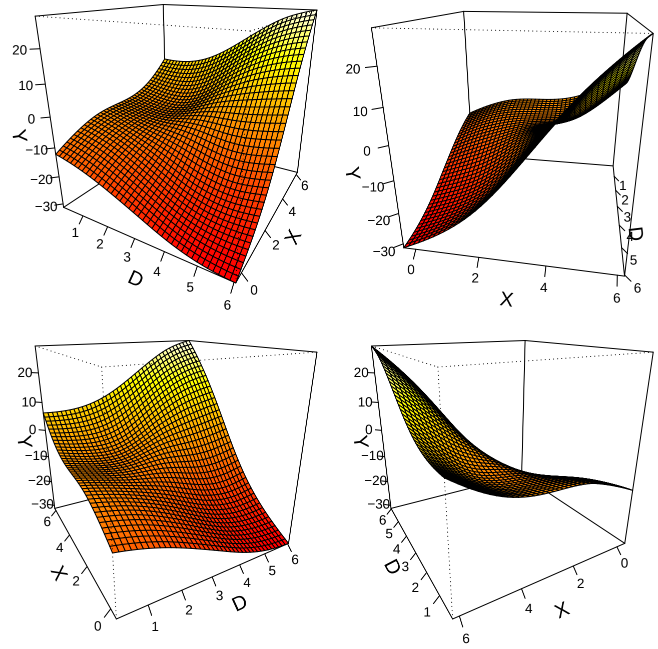
Note: that missing values are due to restrictions in the data, such as lack of common support, that prevented the test from being conducted. In such cases an aggregate score was not computed.

A.3 GAM Plot

In cases where both D and X are continuous, an alternative to the scatterplot is to use a generalized additive model (GAM) to plot the surface that describes how the average Y changes across D and X. While the statistical theory underlying GAMs is a bit more involved (Hastie and Tibshirani 1986), the plots of the GAM surface can be easily constructed using canned routines in R. Figure A1 shows such a GAM plot for the simulated data from the second sample looking at the surface from four distinctive directions. Lighter color on the surface represents a higher value of Y.

Figure A1 has several features. First, it is obvious that holding X constant, Y is increasing in D and holding D constant, Y is increasing in X. Second, the slope of Y on D is larger with higher X than with lower X. Third, the surface of Y over D and X is fairly smooth, with a gentle curvature in the middle but devoid of drastic humps, wrinkles, or holes. In the Online Appendix, we will see that the GAM plots of examples that likely violate the linearity assumption look quite different from Figure A1.

Figure A1. GAM Plot: Simulated Sample with Continuous Treatment



Y
20
10
0
−10
−20
−30
D
1
2
3
4
5
6
X
0
2
4
6