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

Jens Hainmueller* Jonathan Mummolo† Yiqing Xu‡

February 11, 2016

[Word Count: 9,178]

Abstract

Regressions with multiplicative interaction terms are widely used in the social sciences to test whether the relationship between an outcome and an independent variable changes depending on a moderator. Despite much advice on how to use interaction models, two important problems are currently overlooked in empirical practice. First, multiplicative interaction models are based on the crucial assumption that the interaction effect is linear, which fails unless the effect of the independent variable changes at a constant rate with the moderator. Second, reliably estimating the marginal effect of the independent variable at a given value of the moderator requires sufficient common support. Replicating nearly 50 interaction effects recently published in five top political science journals, we find that these core assumptions fail in a majority of cases, suggesting that a large portion of published findings based on multiplicative interaction models are artifacts of misspecification or are at best highly model dependent. We propose straightforward diagnostic tests to assess the validity of these assumptions and offer simple flexible modeling strategies for estimating potentially nonlinear interaction effects.

*Associate Professor, Department of Political Science, Stanford University, jhain@stanford.edu.

†PhD Candidate, Department of Political Science, Stanford University, jmummolo@stanford.edu.

‡PhD Candidate, Department of Political Science, MIT, xyq@mit.edu. We thank all authors who generously provided their replication data.

1 Introduction

Linear regression models with multiplicative interaction terms of the form

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

are a workhorse model in the social sciences to examine whether the relationship between an outcome Y and a key independent variable D varies with levels of a moderator X. The motivation behind such models is that many hypotheses stipulate that the effect of the independent variable of interest varies depending on the context as captured by the moderator X. 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.

A large body of literature advises scholars how to test such conditional hypotheses using multiplicative interaction models. Perhaps most prominently, the pioneering article by Brambor, Clark and Golder (2006) provides a simple checklist of dos and don'ts. They recommend that scholars should (1) include all constitutive terms (D and X) alongside the interaction term ($D \cdot X$) in the model, (2) not interpret the coefficients on the constitutive terms (α and η) as unconditional marginal effects, and (3) compute substantively meaningful marginal effects and standard errors, 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 extremely widely cited[1] and are nowadays often considered the best practice. As our 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

[1]In fact, as of January 2006 Brambor, Clark and Golder (2006) has been cited over 3,000 times according to Google Scholar, which makes it one of the most cited political science articles of the last two decades. Braumoeller (2004) has been cited over 600 times.

(2006). In addition, scholars today rarley leave out constitutive terms or misinterpret the coefficients on the constitutive terms as unconditional marginal effects. Clearly, empirical practice improved dramatically with the publication of Brambor, Clark and Golder (2006) and related advice.[2]

Despite these important advances, we contend that important limitations remain in the current best practice of using multiplicative interaction models. In particular, we emphasize two important problems that are currently 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 often not even monotonic. In fact, replicating nearly 50 interaction effects published in the top five political science journals in the 2006-2015 period, we find that the effect of D on Y changes monotonically across levels of X in only 37 percent of cases and it changes linearly only in 17 percent of cases. This suggests that a large share of published work using multiplicative interaction models draws erroneous conclusions that rest on a modeling artifact that goes undetected even when applying the current best practice guidelines.

Second, another important problem that is currently often overlooked in empirical

[2]Brambor, Clark and Golder (2006) surveyed the top three political science journals between 1998-2002 and found that only 10% of articles that used multiplicative interaction models had followed their simple checklist. Similarly, Braumoeller (2004) showed that prominent articles in international relations routinely misinterpreted lower order terms.

practice is the issue of lack of common support. Scholars using multiplicative interaction models routinely report the conditional marginal 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 to reliably compute the conditional marginal effects. In particular, when computing the effect of D at a specific value of the moderator X_o there should be (1) a sufficient number of observations whose X values are close to X_o and (2) those observations should also exhibit sufficient variation on D. If either of these two conditions fails, the conditional marginal effect estimates are based on extrapolation or interpolation of the functional form to an area where there is no or only very sparse data and therefore the effect estimates are fragile and highly model dependent (King and Zeng 2006). In our replications we find that this type of extrapolation is very common in empirical practice. Many articles report conditional marginal effect estimates for values of the moderator where there are no or very few observations. Similarly, some articles report conditional marginal effect 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.

Our goal is not to point fingers at anyone but 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 estimate conditional marginal effects while relaxing the LIE assumption. Our diagnostics and estimation strategies are easy to implement using standard software packages. Finally, 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

the figures in `R` and `STATA`.[3].

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 conclude.

2 Multiplicative Interaction Models

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

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

In this model Y is the outcome variable, D is the key independent variable of interest or "treatment", X is the moderator, DX is the interaction term between D and X, Z is a vector of control variables, and μ and ϵ represent the constant and error terms, respectively.

We focus on the case where the treatment variable D is either binary or continuous and the moderator X is continuous.[4] Moreover, in the following discussion we focus on the interaction effect components of the model (D, X, and DX). When covariates Z are included in the model, we maintain the typical assumption of the standard multiplicative interaction that the model is correctly specified with respect to these covariates.

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

[3]ADD LINK HERE

[4]When D and X are both binary or discrete with few unique values one can 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.

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 reason to believe that the heterogeneity in the effect of D on Y takes exactly such a linear form. Instead, it might well be that the effect of D on Y is non-linear or 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 only hold if both functions are linear for all modeled contrasts of d_1 versus d_2.[5] In other words, linear interaction models are highly 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 resulting in outliers or bad influence points, etc. As our

[5]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.

empirical survey shows below, in practice this LIE assumption often fails because at least one of the two functions is not linear.[6]

The decomposition in Equation (3) points to another important aspect of linear interaction models which is the issue of common support. Since the conditional effect of D on Y is the difference between two linear functions, it is important that the two functions share a *common support* of X. In other words, to reliably compute the conditional effect of D on Y at a 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 sufficient variation on D. Otherwise, estimation of the marginal effect will essentially 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 the functional form results in fragile and highly model dependent estimates. Slight changes in the assumed functional form can lead to widely different answers (King and Zeng 2006). In our empirical survey below we show that such interpolation or extrapolation is common in empirical work using multiplicative interaction models.

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

[6] Although the linear regression framework is flexible enough to incorporate higher order terms of X and their interaction with D 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).

LIE assumption.

3 Diagnostics

Before introducing the diagnostic tools, we provide two simulated samples for illustration (one with a continuous treatment and one with a binary treatment). The two samples, each of 200 observations, are generated with the following process:

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

Y_i is the outcome for unit i, the moderator is given by $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(3, 1)$, and the error term is given by $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 4)$. In the first sample, the treatment indicator is $D_i \overset{\text{i.i.d.}}{\sim} Bernoulli(0.5)$, 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 two samples share the same sets of X_i and ϵ_i. For simplicity, we do not include any control variables in both specifications.

We now present a simple visual diagnostic to help researchers to diagnose potential problems with the LIE assumption and the lack of common support.

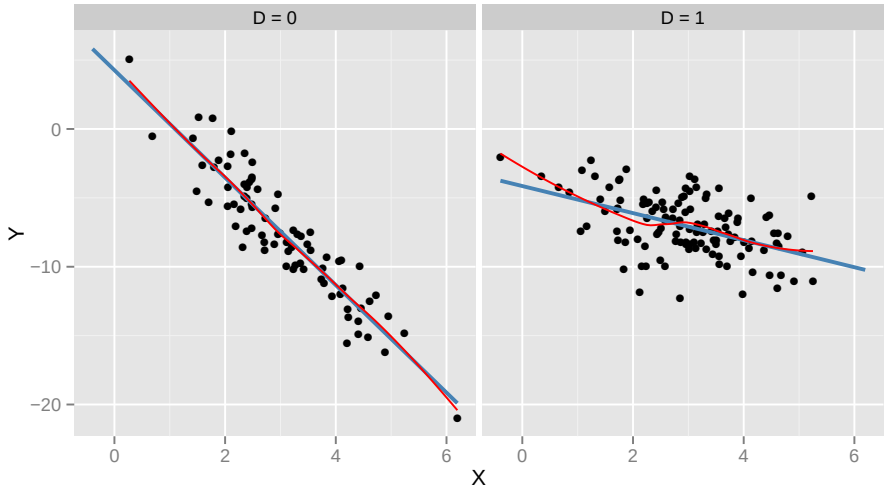
The diagnostic that we recommend is a scatterplot of raw data. This diagnostic is simple to implement and powerful in the sense that it readily reveals the main problems associated with the LIE assumption and lack of common support.

If the treatment D is binary, we recommend simply plotting the outcome variable 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 a LOESS lines in each group (Cleveland and Devlin 1988).[7] The upper panel of Figure 1 presents an example of such a plot for the simulated data with the binary treatment.

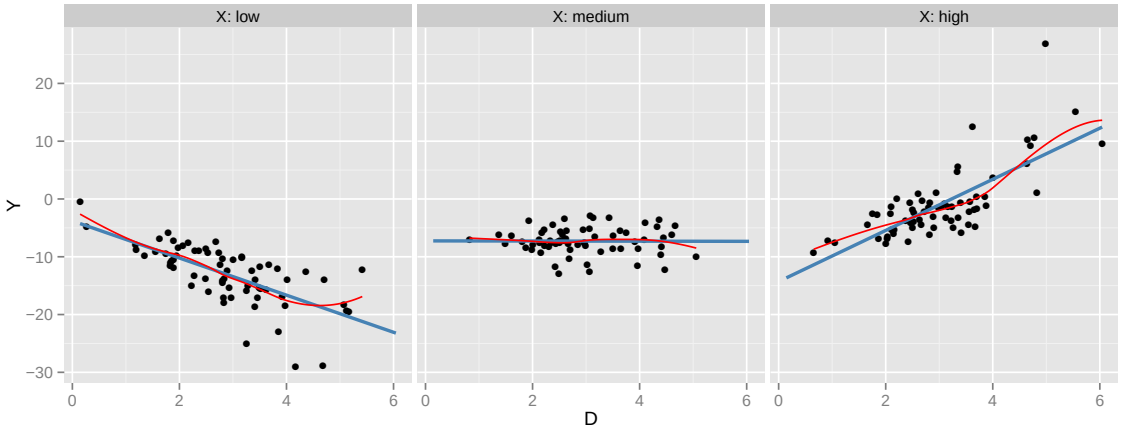
The first important issue to check is whether the relationship between Y and X is

[7] In addition, the same plots can be constructed after residualizing with respect to the covariates Z.

FIGURE 1. RAW PLOTS: SIMULATED SAMPLES



(a) Binary Treatment



(b) Continuous Treatment

reasonably linear in both groups. For this we can simply check if the linear regression lines (blue) and the LOESS lines (red) diverge considerably across the range of X values. In this case, 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. We also see that the slope of Y on X of 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 second important issue to look out for is whether there is sufficient common

support 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 is a sufficient number of data points in both groups. In our example, we see that we see that both groups share a common support of X for the range between about 1.5 to 5 as we would expect given the simulation parameters.[8]

If the moderator is 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 X: 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 adding again both the linear fit and the LOESS fit.

The lower panel of Figure 1 presents such a plot for the second simulated data with the continuous treatment. As before, this 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. Comparing across the three panels of the plot, we also see that there is again sufficient common support for X values ranging between 1.5 to 5. 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.[9]

[8] 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.

[9] In this case of a continuous treatment and continuous moderator it is often also useful to visualize interactions using a three-dimensional surface plot generated by a generalized additive model (GAM, Hastie and Tibshirani 1986). 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 relax the LIE assumption and flexibly allow for heterogeneity in how the conditional marginal effect changes across values of X. In addition, they offer protection against model dependency coming from excessive extrapolation or interpolation to areas where the data is very sparse.

4.1 Binning Estimator

The first estimation approach is a simple binning estimator. 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, 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 up a 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 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:

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

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 than the standard multiplicative interaction model. Essentially, the binning estimator jointly fits the interaction components of the standard model to each bin separately.[10] Since $(X - x_j)$ equals zero at each evaluation point x_j, the conditional marginal effect of D on Y at the chosen points within each bin, x_1, x_2, and x_3, is simply given by α_1, α_2, and α_3, respectively. The model does not impose the LIE assumption. Instead, 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 of X.

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.[11]

[10] 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.

[11] 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 in five top journals we found not a single case where this potential problem occurs (see below).

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 actually 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:

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

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). So in the special case when the standard multiplicative interaction model is correct and the global LIE assumption holds, then—as the sample size grows—the marginal effect estimates from the binning estimator converge in probability on the unbiased marginal effect estimates from the standard multiplicative interaction model given by $ME(X) = \hat{\alpha} + \hat{\beta} X$.

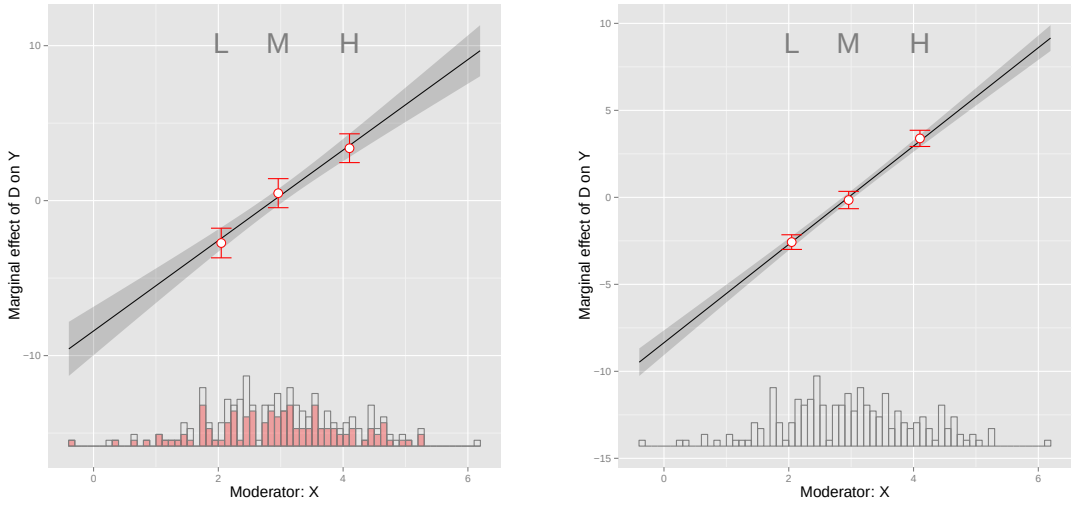
To illustrate the results from the binning estimator we apply it to both simulated datasets that cover the case of a binary and continuous treatment, respectively. The results are shown in Figure 2. To clarify the correspondence between the binning

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 original marginal effects plot generated from the standard multiplicative interaction model as recommended by Brambor, Clark and Golder (2006).

In the case of the binary treatment, we also 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 staked 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 group, 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, the histogram at the bottom simply shows the distribution of X in the entire sample.

Recall that in our simulated data the standard multiplicative interaction model is the correct model and the LIE assumption holds. Hence, as Figure 2 shows, the conditional effect estimates from the binning estimator and the standard standard multiplicative interaction model are extremely 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. When the estimates from the binning estimator are instead far off the line or when they are out of order, (for example, first increasing then decreasing), however, we have a critical piece of evidence that the LIE assumption does not hold. 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.

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



(a) Binary Treatment

(b) Continuous Treatment

Finally, to go further we can conduct three two-sided t-tests to test the following three hypotheses: (1) $\alpha_2 = \alpha_1$; (2) $\alpha_3 = \alpha_2$; and (3) $\alpha_1 = \alpha_3$. Ideally, we want reject these hypothesis when $|\beta| > 0$. In practice, we may lack statistical power with some of the tests, but we should at least be able to reject $\alpha_1 \neq \alpha_3$ if the marginal effect is linear in X and $|\beta| > 0$. In our simulated data, the estimate from the bin with "High" X is significantly different from that from the bin with "Low" X ($p < 0.000$ in both cases), which provides a clear sign of increasing treatment effects of D on Y.

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 is more complicated than the simple binning estimator, but 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. 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 $f(.)$, $g(.)$, and $\gamma(.)$ are smooth functions of X, and $g(.)$ 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) is 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 does not have to fall on a linear line 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.[12]

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:

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

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

in which $K(.)$ is a Gaussian kernel and h is a smooth parameter that is automatically selected 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 would reduce 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 $\hat{f}(.)$, $\hat{g}(.)$, and $\hat{\gamma}(.)$, in which $\hat{g}(.)$ represents the estimated marginal effect of D on Y with respect to X.[13] Note that this estimation procedure can be implemented

[12] If the model includes fixed effects as covariates those can be partialled out prior to the estimation of the kernel.

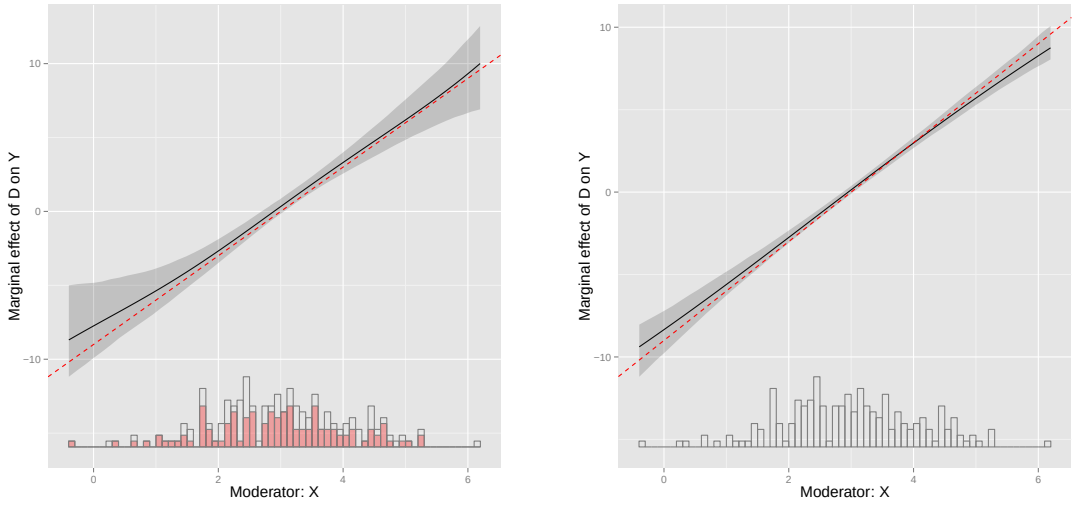
[13] For theoretical properties of the kernel smoothing estimator, see Li and Racine (2010).

in R using the `npscoef` function in the `np` package. Standard errors and confidence intervals can be computed using a bootstrap.

Figure 3 shows the results of the kernel estimator applied to the two simulated samples. 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 (staked) histograms at the bottom to judge the common support based on the distribution of the moderator.

We see that the results are very similar to the binning estimator as both plots suggest a strong linear interaction where the conditional marginal effect of D on Y grows linearly with X. The marginal estimates from the kernel estimator are very close to those from the true multiplicative interaction model (red dashed line).

FIGURE 3. KERNEL SMOOTHED ESTIMATES: SIMULATED SAMPLES



(a) Binary Treatment

(b) Continuous Treatment

Also note that on the boundaries where there is limited common support on X, the local estimates are imprecisely estimated as expected given that there is little data to estimate the marginal effects at these points. The fact that the confidence intervals get very wide at those points is desirable because it makes clear the lack of

common support.

5 Data

We now apply our diagnostic and estimation strategies to published papers that used classical linear interaction models. In order to 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). 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 model; and interacted at least one continuous variable. We excluded methods and review articles, as well as triple interactions.

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. In order to identify studies within this second sample which featured interaction models prominently, we subset to 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 39 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 across the entire range of the moderator and/or where the authors did not claim to detect a marginal effect.[14]

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. 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.[15]

[14] We cap the number of replicated interactions at 4 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.

[15] 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.

6 Results

Case 1: Linear Marginal Effects

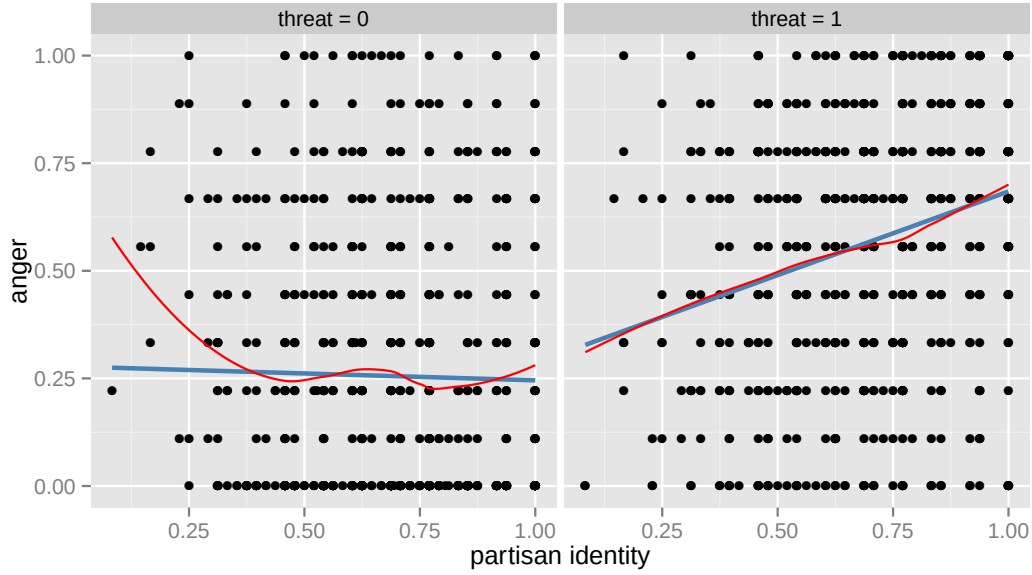
We begin our discussion with a replication of Huddy, Mason and Aarøe (2015), an example of a study in which the assumptions of the multiplicative interaction model appear to hold well. This study uses a survey experiment and a multiplicative interaction model to test the hypothesis that a threat of electoral 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 (scale 0/1). 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, there is sufficient common support for the range of partisan identity between about .25 to 1.

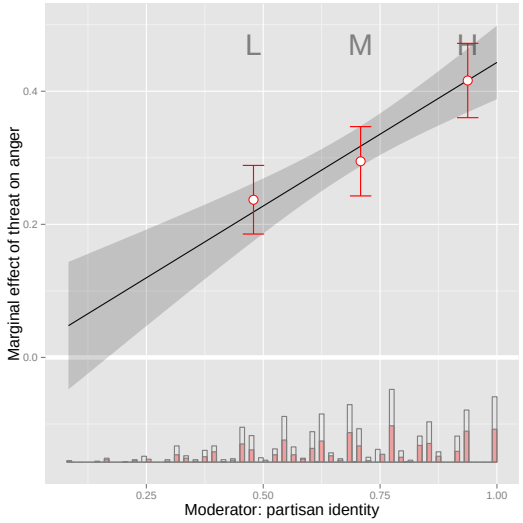
The middle panel in Figure 4 displays the conditional marginal effect estimates of the binning estimator superimposed on the estimates from the multiplicative interaction model used by the authors. As expected given the scatterplot, the conditional marginal effect estimates of the binning estimator for 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 threat effect at medium levels

falls about right in the middle of the low and high estimates. In addition, the staked histogram at the bottom suggests that there is good sufficient common support with both treated and control observations across a wide range of values of the moderator.

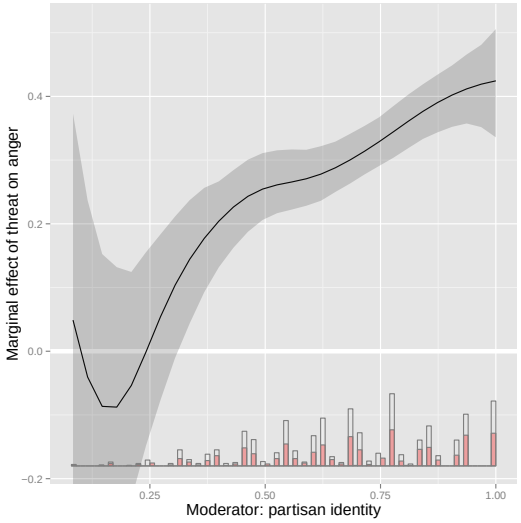
FIGURE 4. LINEAR INTERACTION EFFECT:
REPLICATION OF HUDDY, MASON AND AARØE (2015)



(a) Scatterplot



(b) Marginal Effects from Replicated Model (black line) and from Binning Estimator (red dots)



(c) Marginal Effects from Kernel Estimator

The lower panel in Figure 4 presents the conditional marginal effect estimates from the kernel estimator which again confirms the support for the LIE assumption. The magnitude of the threat effect increases at an approximately constant rate with higher partisan identity. The confidence bands start to get rather wide below values of .25 for partisan identity which is consistent with the fact that there are very few data-points with such low levels of partisan identity.

Case 2: Lack of Common Support

The next example illustrates how fitting an interaction model with a continuous variable without additional diagnostic checks can mask the fact that the data lack common support since the treatment is not varying 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 on 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 (scale -1/0). In Figure 2 in the study, the authors plot the marginal effect of U.N. authorization, and state, "[c]learly, the effect of authorization on rallies decreases as similarity increases," (p. 756).

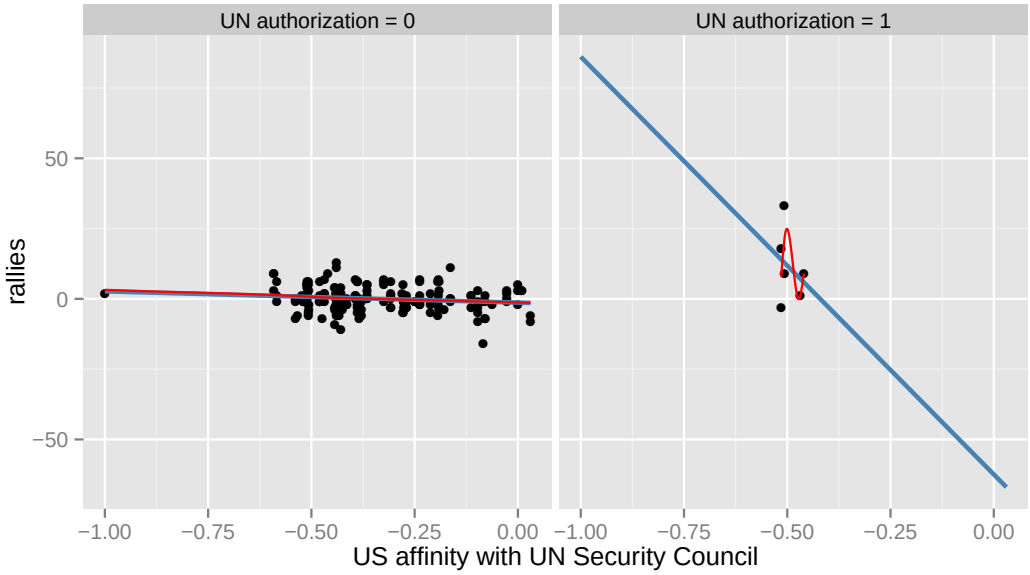
The upper panel in Figure 5 shows our diagnostic scatterplot for this data and the lower left panel in Figure 5 reproduces the original plot displayed in the study but overlays the estimates from the binning estimator for low, medium, and high values of the moderator. Again, in the latter plot the staked 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 an U.N. authorization and those observations are all clustered

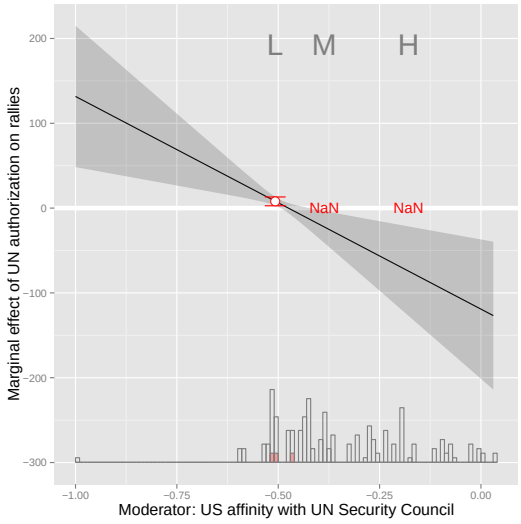
in a narrow range of moderator values of around -.5. In fact, as can be seen in the histogram at the bottom of the plot in the lower panel or the raw plot in the upper panel of Figure 5, all the observations with an U.N. authorization fall into the lowest tercile of the moderator and the estimated marginal 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 purely based on extrapolation based on the functional form and therefore highly model dependent and fragile. The data here simply cannot reliably answer the question as to how the effect of U.N. authorizations varies across the preference distance between the U.S. and the Security Council because there the very few cases with and without authorizations are all concentrated in the narrow range of the moderator around .45, while for other moderator values there is no variation whatsoever in the treatment. This becomes yet again clear in the marginal effect estimates from the kernel estimator 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 effect estimates blow up indicating that the effect cannot be reliably 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.

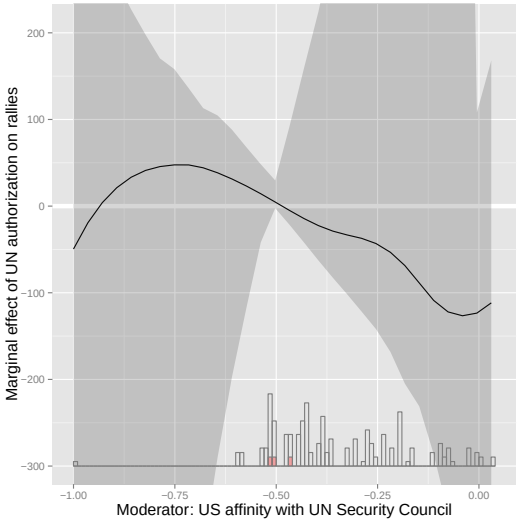
FIGURE 5. LACK OF COMMON SUPPORT: CHAPMAN (2009)



(a) Raw plot



(b) Marginal Effects from Replicated Model (black line) and from Binning Estimator (red dots)



(c) Marginal Effects from Kernel Estimator

Case 3: Severe Interpolation

Similar to the lack of common support issue, the next published example illustrates how sparsity of data in various regions of the moderator (as opposed to no variation at all in the treatment) can lead to severe misspecifications. Malesky, Schuler and Tran

(2012) 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. The authors argue that the response of delegates to this transparency intervention is conditional on the level of internet penetration in their district. 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 district, and the interaction between the two.

The upper panel in Figure 6 reprints the marginal effect plots presented by the authors in Figure 1 of their APSR 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" (p. 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).

How trustworthy is this result which carries significant policy implications? Given that the treatment was randomized, the results might be taken to be highly credible. The lower left panel in Figure 6 displays the marginal effect estimates from our

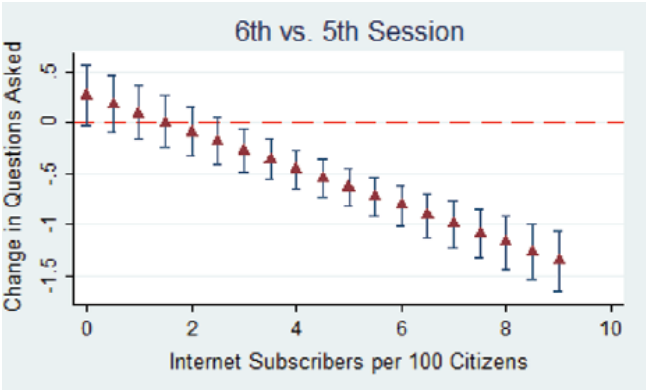
replication of the original model and the binning estimator. Our replication plots shows two critical concerns. First, the effect of the transparency intervention appears non-monotonic and non-linear in the moderator. In fact, the point estimates grow smaller between typical low and typical medium levels of internet penetration, but then larger between typical medium and typical high levels of internet penetration. This suggests that the LIE assumption does not hold and when relaxed by the binning estimator there is no compelling evidence of a negative interaction effect. In fact, in direct contrast to the central claim by the authors that the treatment had a large negative effect at high levels of internet penetration, the effect of the transparency intervention is, if anything, positive and largest at the highest tercile of internet penetration indicating that transparency, if anything, made legislators from typical high penetration districts more likely to ask questions, (although the estimate is not significantly different from zero).

Second, as illustrated by the staked 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 higher levels of internet penetration 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. In fact, for the range between 2.5 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 incorrect functional form to an area far outside the bulk of the data. In this case, the estimates also appear biased given that the functional form is clearly not linear for the bulk of the data.[16] The linear downward trend is entirely driven by the severe outliers with unusually high levels of internet penetration and once these outliers (which make up less than 5% of the data) are removed the trend utterly flattens, indicating no effect of the intervention at any level of internet penetration (see the lower right panel in Figure

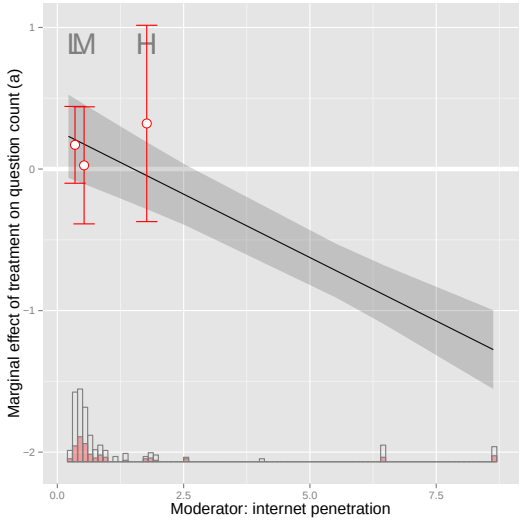
[16] In the appendix we show that the same problem applies to all the other four outcomes used by Malesky, Schuler and Tran (2012) in their Figure 1.

6 below).

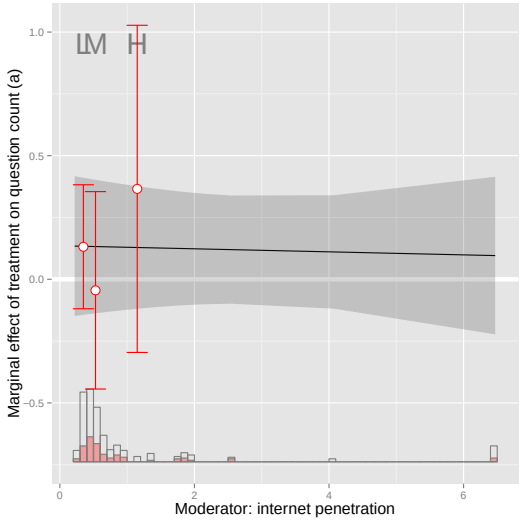
FIGURE 6. SEVERE INTERPOLATION: MALESKY, SCHULER AND TRAN (2012)



(a) Original plot



(b) Marginal Effects from Replicated Model (black line) and from Binning Estimator (red dots)



(c) Marginal Effects from Replicated Model (black line) and from Binning Estimator (red dots) with outliers (less than 5% of data) removed

Case 4: Nonlinearity

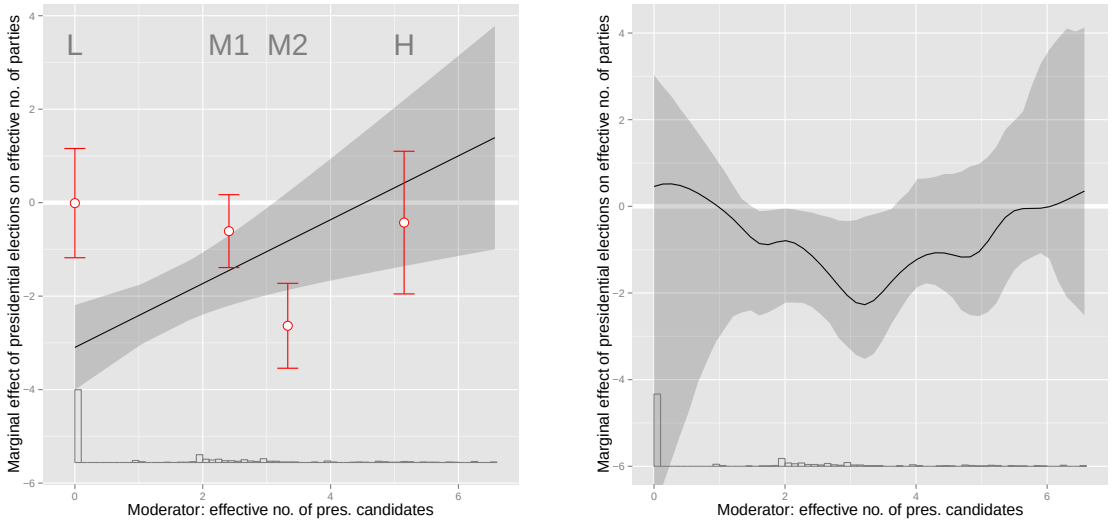
Our next example underscores how fitting linear interaction models can mask nonlinearities in interaction effects and therefore result in severe misspecification bias. Clark and Golder (2006) argue that the temporal proximity of presidential elections affects the number of parties that compete in an 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 and again superimposed the estimates from the binning estimator where we use four bins to discretize the moderator.[17] The authors interpret the plot 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).

But as the estimates from the binning estimator in this figure show, the story is much more complicated. In the first bin which contains the 59% of observations where the moderator takes on the value of zero, the treatment effect is essentially zero much to the contrary of the claim of a strong negative effect when there is a low effective number of candidates. Again contrary to the synopsis in Clark and Golder (2006), the effect slightly *declines* in the second bin, and then rapidly drops to be negative and significant at the third bin, only to increase again back to zero in the last bin. Clearly, the LIE assumption does not hold and accordingly the linear interaction model is misspecified. The linearly increasing marginal effect estimates in this case are a modeling artifact. This is confirmed by the marginal effect estimates from the kernel estimator which are shown in the right plot of in Figure 7. Consistent with the binning estimates, the marginal effect appears highly nonlinear. Contrary to the authors' claims, the number of candidates in an election does not appear to condition the effect of proximate elections in a consistent manner.

[17] Among the 487 observations, 59% of them have 0 value for the moderator; we split the range of the moderator into $[0, 2)$, $[2, 3)$, $[3, 4)$ and $[4, 7]$ such that the binned estimates well represent the entire range. Note that this plot is also used in Brambor, Clark and Golder (2006).

FIGURE 7. NONLINEARITY: CLARK AND GOLDER (2006)



(a) Marginal Effects from Replicated Model (black line) and from Binning Estimator (red dots)

(b) Marginal Effects from Kernel Estimator

Summary of Replications

The previous cases highlight extreme 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 in published work? How much should we trust published estimates from multiplicative interaction models? 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 awarded single points for exhibiting (1) no severe extrapolation, (2) monotonic marginal effects, and (3) linearity.

We determined the first criterion by examining whether the L-Kurtosis of the moderator (Hosking 1990) 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[18] and therefore captures to what extent

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

the estimates reported in the marginal effect plots are based on extrapolation to moderator values where there is little or no data.[19] We determined the second criterion of monotonic effects by checking whether the estimates from the binning estimator changed monotonically across the three tercile bins of the moderating variable. Finally, if the estimates from the binning estimator fell close to the original marginal effect line and if the kernel plot appeared more or less linear, we coded the case as being a linear interaction. While there is admittedly some subjectivity in the application of our scoring system, we feel it provides a useful framework for summarizing the current state of the literature employing these models. We also display more complete analyses of each case in the Online Appendix 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 5 of the 46 cases (10.9%) received a perfect score of three indicating that the reported marginal effects meet all three criteria of no severe extrapolation, monotonicity, and linearity. This is an unnervingly low fraction given the loose scoring system we imposed, which did not even demand that the binned marginal effects be statistically distinguishable. Nine cases (19.6%) received a score of 2, while 16 cases (34.8%) received a score of 1. Sixteen cases (34.9%) received a score of zero, failing to meet a single one of the three criteria.[20]

[19]For example, 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. In contrast, in the case of Huddy, Mason and Aarøe (2015) the L-Kurtosis is .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. In fact, in this case 80% of the density is concentrated in about 53% of the interval reported in the marginal effects plot. 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.

[20]For details on the scores for each case, see Table A1 in the Appendix.

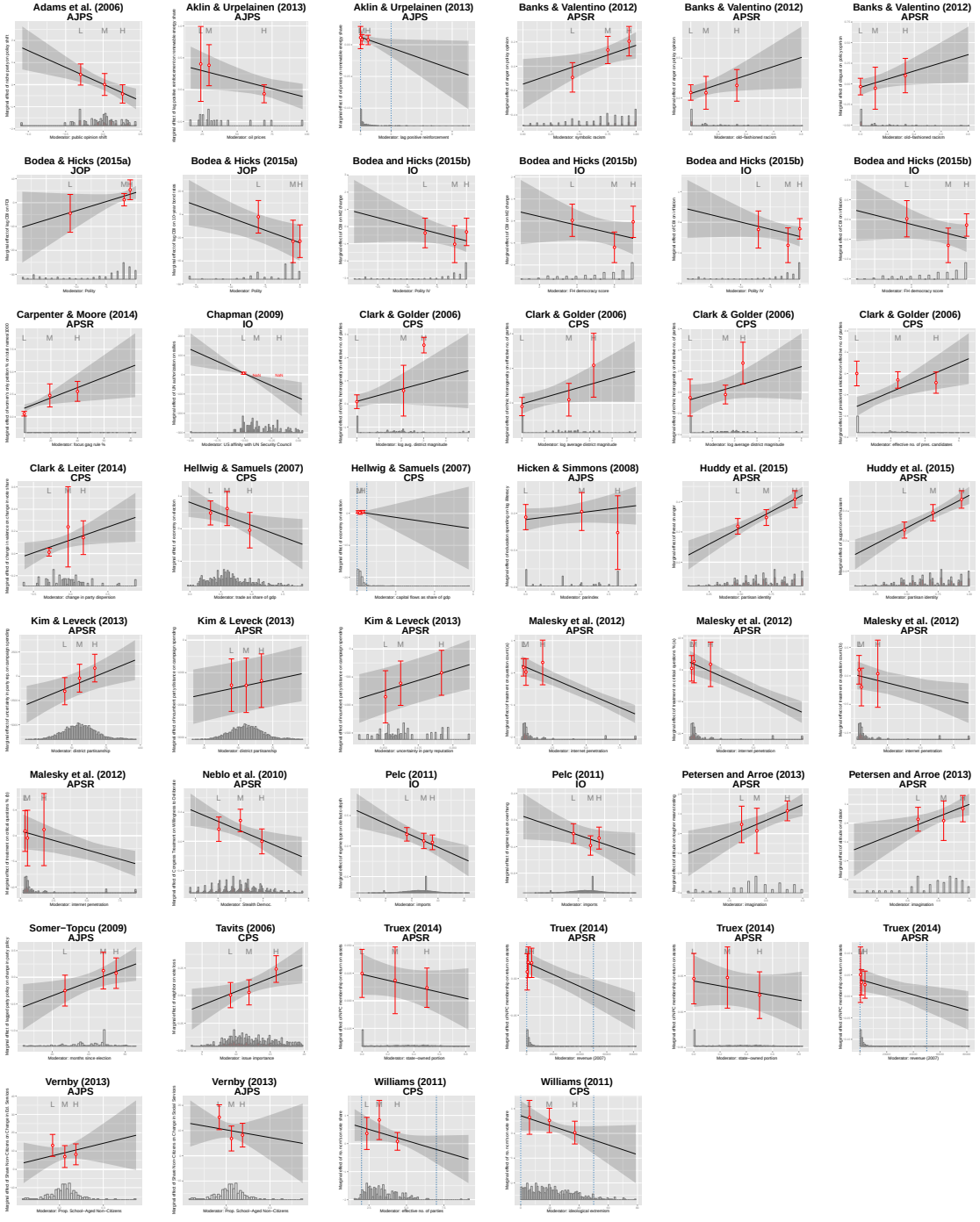
Once we break out the results by journal, we find that the issues raised by our review are not unique to any one subfield or journal in political science. Comparative Political Studies (CPS) received the highest overall mean score, 1.4 on our 0-to-3 scale, while the APSR ranked a close second with a score of 1.2. The lowest score was 0.57 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 an urgent need for improved practices when employing multiplicative interaction models.

Table 1. Replication Results by Journal

Journal	Cases	No severe extrapolation	Monotonic effects	Linearity	Score
APSR	17	0.59	0.41	0.18	1.20
AJPS	9	0.22	0.44	0.22	0.89
JOP	3	0.67	0.00	0.33	1.00
CPS	10	0.80	0.50	0.10	1.40
IO	7	0.29	0.14	0.14	0.57
Total	46	0.52	0.37	0.17	1.10

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.

FIGURE 8. THE ASSUMPTIONS OF THE LINEAR INTERACTION MODEL RARELY HOLD IN PUBLISHED WORK



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 important problems that are currently 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 journals, our findings here most likely understate the true extent of the problem in published work. Overall, our replications suggest that a large portion of findings in published work employing multiplicative interaction models are based on modeling artifacts, or are at best highly model dependent, and suggest a need to augment the current best practice guidelines.

We recommend that researchers engaged in testing conditional hypotheses should engage in the following:

1. Generate the diagnostic scatterplot of 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 diagnostics 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 assists with these checks (see Appendix for details on GAM plots).

2. 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 are sufficient to get a good sense of the effect heterogeneity. More bins can be used if more precision is required and more data is available. In addition, generating the marginal effects estimates using the kernel estimator can be helpful to further evaluate the effect heterogeneity. In any case, 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 highly recommend to always add a (staked) 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. 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).

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

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.

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

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.

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.

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

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.

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.

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.

Neblo, Michael A., Kevin M. Esterling, David M.J. Kennedy, Ryan P. 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.

A Supplementary Information

Appendix: Table of Contents

A.1 Proofs

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 :

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

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:

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

Since $\hat{\underline{\alpha}}_j = \hat{\alpha}_j - \hat{\beta}_j x_j$, 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	Monotonic effects	No severe extrapolation	Linearity	Score
Adams and Glasgow (2006)	AJPS	1	0	1	2
Aklin and Urpelainen (2013)	AJPS	1	0	0	1
Aklin and Urpelainen (2013)	AJPS	1	0	0	1
Banks and Valentino (2012)	AJPS	1	1	1	3
Banks and Valentino (2012)	AJPS	0	0	0	0
Banks and Valentino (2012)	AJPS	0	0	0	0
Bodea and Hicks (2015*a*)	JOP	0	1	0	1
Bodea and Hicks (2015*a*)	JOP	0	0	0	0
Bodea and Hicks (2015*b*)	IO	0	0	0	0
Bodea and Hicks (2015*b*)	IO	0	1	0	1
Bodea and Hicks (2015*b*)	IO	0	0	0	0
Bodea and Hicks (2015*b*)	IO	0	1	0	1
Carpenter and Moore (2014)	APSR	1	0	0	1
Chapman (2009)	IO	0	0	0	0
Clark and Golder (2006)	CPS	1	1	0	2
Clark and Golder (2006)	CPS	1	1	0	2
Clark and Golder (2006)	CPS	1	1	0	2
Clark and Golder (2006)	CPS	0	1	0	1
Clark and Leiter (2014)	CPS	0	0	0	0
Hellwig and Samuels (2007)	CPS	0	1	0	1
Hellwig and Samuels (2007)	CPS	0	0	0	0
Hicken and Simmons (2008)	AJPS	0	1	0	1
Huddy, Mason and Aarøe (2015)	APSR	1	1	1	3
Huddy, Mason and Aarøe (2015)	APSR	1	1	1	3
Kim and LeVeck (2013)	APSR	1	1	0	2
Kim and LeVeck (2013)	APSR	0	1	0	1
Kim and LeVeck (2013)	APSR	1	1	0	2
Malesky, Schuler and Tran (2012)	APSR	0	0	0	0
Malesky, Schuler and Tran (2012)	APSR	0	0	0	0
Malesky, Schuler and Tran (2012)	APSR	0	0	0	0
Malesky, Schuler and Tran (2012)	APSR	0	0	0	0
Neblo et al. (2010)	APSR	0	1	0	1
Pelc (2011)	IO	1	0	1	2
Pelc (2011)	IO	0	0	0	0
Petersen and Aarøe (2013)	APSR	0	1	0	1
Petersen and Aarøe (2013)	APSR	0	1	0	1
Somer-Topcu (2009)	JOP	0	1	1	2
Tavits (2008)	CPS	1	1	0	2
Truex (2014)	APSR	1	1	1	3
Truex (2014)	APSR	0	0	0	0
Truex (2014)	APSR	0	1	0	1
Truex (2014)	APSR	1	0	0	1
Vernby (2013)	AJPS	0	0	0	0
Vernby (2013)	AJPS	0	0	0	0
Williams (2011)	CPS	0	1	0	1
Williams (2011)	CPS	1	1	1	3

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 absent 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 differently from Figure A1.

FIGURE A1. GAM PLOT: SIMULATED SAMPLE WITH CONTINUOUS TREATMENT