Marginal Effects in Interaction Models: Determining and Controlling the False Positive Rate*

Justin Esarey† and Jane Lawrence Sumner‡

December 6, 2015

Abstract

When a researcher suspects that the marginal effect of x on y varies with z, a common approach is to plot $\partial y/\partial x$ at different values of z along with a pointwise confidence interval (generated using the procedure described in Brambor, Clark and Golder, 2006) in order to assess the magnitude and statistical significance of the relationship. In this paper, we demonstrate that this approach produces statistically significant findings under the null hypothesis at a rate that can be many times larger or smaller than the nominal false positive rate of the test. Conditioning inference on the statistical significance of the interaction term does not solve this problem. However, we demonstrate that the problem can be avoided by exercising qualitative caution in the interpretation of marginal effects and via methodological adjustments implemented in our `interactionTest` software package. (8946 words)

*Nathan Edwards provided research assistance while writing this paper, for which we are grateful.

†Assistant Professor, Department of Political Science, Rice University. Corresponding author: justin@justinesarey.com.

‡Department of Political Science, Emory University.

Introduction

Much of the recent empirical work in political science[1] has recognized that causal relationships between two variables x and y are often changed—strengthened or weakened—by contextual variables z. Such a relationship is commonly termed *interactive*. The substantive interest in these relationships has been coupled with an ongoing methodological conversation about the appropriate way to test hypotheses in the presence of interaction. The latest additions to this literature, particularly King, Tomz and Wittenberg (2000), Ai and Norton (2003), Braumoeller (2004), Brambor, Clark and Golder (2006), Kam and Franzese (2007), Berry, DeMeritt and Esarey (2010), and Berry, Golder and Milton (2012), emphasize visually depicting the marginal effect of x on y at different values of z (with a confidence interval around that marginal effect) in order to assess whether that marginal effect is statistically and substantively significant. The statistical significance of a multiplicative interaction term is seen as neither necessary nor sufficient for determining whether x has an important or statistically distinguishable relationship with y at a particular value of z.[2]

A paragraph from Brambor, Clark and Golder (2006) summarizes the current state of the art:

> The analyst cannot even infer whether x has a meaningful conditional effect on y from the magnitude and significance of the coefficient on the interaction term either. As we showed earlier, it is perfectly possible for the marginal effect of x on y to be significant for substantively relevant values of the modifying variable z even if the coefficient on the interaction term is insignificant. Note what this means. It means that one cannot determine whether a model should include an interaction term simply by looking at the significance of the coefficient on the interaction term. Numerous articles ignore this point and drop interaction terms

[1] Between 2000 and 2011, 338 articles in the *American Political Science Review*, the *American Journal of Political Science*, and the *Journal of Politics* tested some form of hypothesis involving interaction.

[2] More specifically, the statistical significance of the product term is sufficient (in an OLS regression) for concluding that $\partial y/\partial x$ is different at different values of z (Kam and Franzese, 2007, p. 50), but not whether $\partial y/\partial x$ is statistically distinguishable from zero at any particular value of z.

> if this coefficient is insignificant. In doing so, they potentially miss important conditional relationships between their variables (74).

In short, they recommend including a product term xz in linear models where interaction between x and z is suspected, then examining a plot of $\partial y/\partial x$ and its 95% confidence interval over the range of z in the sample.[3] If the confidence interval does not include zero for any value of z, one should conclude that x and y are statistically related (at that value of z), with the substantive significance of the relationship given by the direction and magnitude of the $\partial y/\partial x$ estimate. It is hard to exaggerate the impact that the methodological advice given in Brambor, Clark and Golder (2006) has had on the discipline: the article has been cited over 2800 times as of August 2015. Similar advice is given in Braumoeller (2004, pp. 815-818, esp. Figure 2), which has been cited over 600 times in the same time frame.

In this paper, we highlight a heretofore unrecognized hazard with this procedure: the reported α-level of confidence intervals and hypothesis tests constructed using the procedure can be inaccurate because of a multiple comparison problem (Sidak, 1967; Abdi, 2007). The source of the problem is that adding an interaction term z to a model like $y = \beta_0 + \beta_1 x$ is analogous to dividing a sample data set into subsamples defined by the value of z, each of which (under the null hypothesis) has a separate probability of a false positive (i.e., falsely rejecting the null hypothesis when the null is true). In contrast, the methods that are described in Brambor, Clark and Golder (2006) construct a pointwise confidence interval (typically using a two-tailed $\alpha = 0.05$); "pointwise" indicates that the confidence intervals are constructed without considering the joint coverage of the confidence interval for all values of z. That is, the confidence interval for each value of z assumes a *single* draw from the sampling distribution of the marginal effect of interest. As a result, these confidence intervals can either

[3]This advice is spelled out on pp. 75-76 of Brambor, Clark and Golder (2006), when they describe the application of their technique to a substantive example: "The solid sloping line in Fig. 3 indicates how the marginal effect of temporally-proximate presidential elections changes with the number of presidential candidates. Any particular point on this line is $\frac{\partial ElectoralParties}{\partial Proximity} = \beta_1 + \beta_3 PresidentialCandidates$. 95% confidence intervals around the line allow us to determine the conditions under which presidential elections have a statistically significant effect on the number of electoral parties—they have a statistically significant effect whenever the upper and lower bounds of the confidence interval are both above (or below) the zero line." We revisit this substantive example in our application section later in the paper.

be too wide or too narrow: plotting $\partial y/\partial x$ over values of z and reporting any statistically significant relationship tends to result in overconfident tests, while plotting $\partial y/\partial x$ over z and requiring statistically significant relationships at multiple values of z tends to result in underconfident tests.[4] The latter may be assessed when, for example, a theory predicts that $\partial y/\partial x > 0$ for $z = 0$ and $\partial y/\partial x < 0$ for $z = 1$ and we try to jointly confirm these predictions in a data set.

We believe that researchers can minimize the impact of this issue using a few simple measures. Our primary recommendation is for researchers to simply be aware that marginal effects plots generated under a given α could be over- or underconfident, and thus to take a closer look if results are at the margin of statistical significance. When overconfidence is an issue, researchers can control the *false discovery rate* (or FDR) in marginal effects plots by adapting the procedure of Benjamini and Hochberg (1995);[5] we provide code to accomplish this in R in the `interactionTest` package. Researchers can also control the *familywise error rate* (or FWER) of these plots using a simple F-test (Kam and Franzese, 2007, pp;. 43-51), although this procedure is more conservative and less powerful than controlling the FDR. We also rule out one possible solution for overconfidence: researchers cannot solve the problem by conditioning inference on the statistical significance of the interaction term (assessing $\partial y/\partial x$ for multiple z only when the product term indicates interaction in the DGP) because this procedure results in an excess of false positives. In situations of underconfident results, a bootstrapping procedure allows researchers to construct marginal effects plots with confidence intervals that have appropriate coverage; we provide R code for this procedure in the `interactionTest` package. Finally, we demonstrate the application of our recommendations by re-examining Clark and Golder (2006), one of the first published applications of the hypothesis testing procedures described in Brambor, Clark and Golder (2006).

[4]We thank an anonymous reviewer for suggesting this phraseology.

[5]For a variant of this procedure involving assigning differential weights to different kinds of hypotheses, see Spahn and Franco (2015).

Interaction terms and the multiple comparison problem

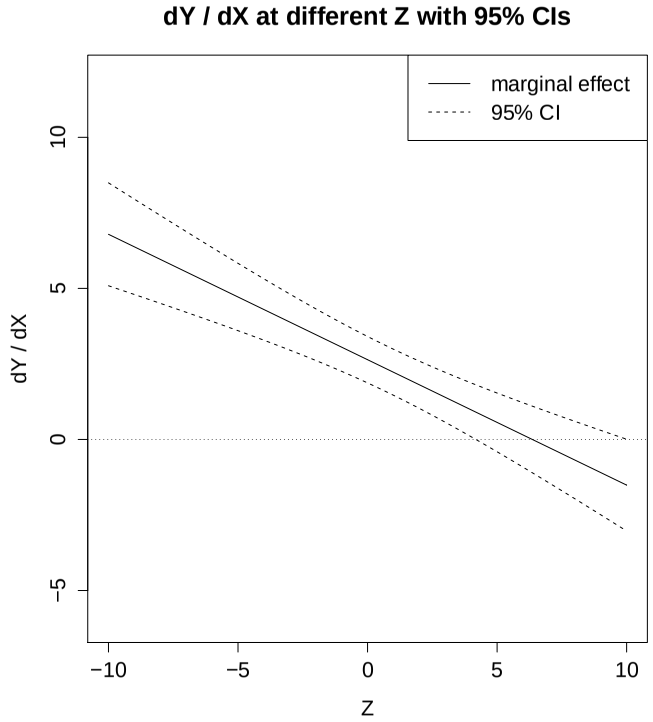
We begin by considering the following question: when we aim to assess the marginal effect of x on y ($\partial y/\partial x$) at different values of a conditioning variable z, how likely will at least one marginal effect come up statistically significant by chance alone? In the context of linear regression, Brambor, Clark and Golder (2006) recommend (i) estimating a model with x, z, and xz terms, then (ii) plotting the estimated $\partial y/\partial x$ from this model for different values of z along with 95% confidence intervals. If the CIs exclude zero at any z, they conclude that the evidence rejects the null hypothesis of no effect for this value of z (Brambor, Clark and Golder, 2006, pp. 75-76). Figure 1 depicts sample plots for continuous and dichotomous z variables; the 95% confidence interval excludes zero in both examples (for values of $z \lessapprox 4$ in the continuous case, and for both $z = 0$ and 1 in the dichotomous case), and so both samples can be interpreted as evidence for a statistical relationship between x and y.

Our goal is to assess the false positive rate of this test procedure—that is, the proportion of the time that this procedure detects a statistically significant $\partial y/\partial x$ for some value of z when in fact $\partial y/\partial x = 0$ for all z. If the false positive rate is greater than the nominal size of the test, α, then the procedure is overconfident: the confidence interval covers the true value less than α proportion of the time. If the false positive rate is less than α, then the procedure is underconfident: the confidence interval could be narrower while preserving its property of covering the true value α proportion of the time. In the case of the Brambor, Clark and Golder (2006) procedure, the question is whether the 95% CIs in Figure 1 exclude zero for at least one value of z more or less than 5% of the time under the null.

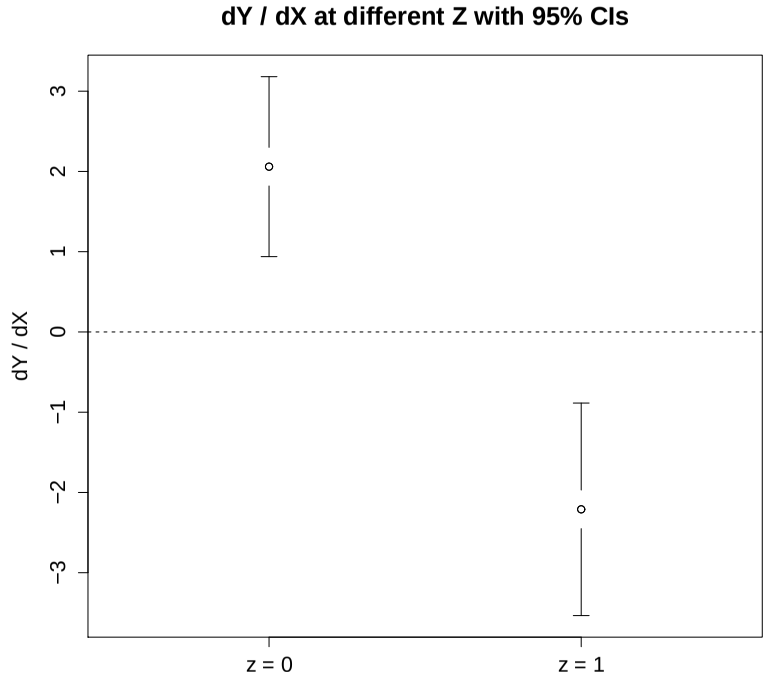
As most applied researchers know, when a t-test is conducted—e.g., for a coefficient or marginal effect in a linear regression model—the α level of that t-test (which is shorthand for the size of that test) is only valid for a single t-test conducted on a single coefficient or

Figure 1: Sample Marginal Effects Plots in the Style of Brambor, Clark and Golder (2006)*

(a) Continuous x and z



(b) Continuous x, Dichotomous z



*Continuous x and z: data were generated out of the model $y = 0.15 + 2.5 * x - 2.5 * z - 0.5 * xz + u$, $u \sim \Phi(0, 15)$, x and $z \sim U[-10, 10]$; model fitted on sample data set, $N = 50$. Dichotomous x and z: data were generated out of the model $y = 0.15 + 2.5 * x - 2.5 * z - 5 * xz + u$, $u \sim \Phi(0, 15)$, $x \sim U[-10, 10]$ and $z \in \{0, 1\}$with equal probability; model fitted on sample data set, $N = 50$.

marginal effect.[6] Consider the example of a simple linear model:

$$E[y|x_1, ..., x_k] = \hat{y} = \sum_{i=1}^{k} \hat{\beta}_i x_i$$

If a researcher conducts two t-tests on two different β coefficients, there is usually a greater than 5% chance that either or both of them comes up statistically significant by chance alone when $\alpha = 0.05$. In fact, if a researcher enters k statistically independent variables that have no relationship to the dependent variable into a regression, the probability in expectation that at least one of them comes up statistically significant is:

$$\begin{aligned} \Pr(\text{at least one false positive}) &= 1 - \Pr(\text{no false positives}) \\ &= 1 - \prod_{i=1}^{k} \left(1 - \Pr\left(\hat{\beta}_i \text{ is st. sig.} | \beta_i = 0\right)\right) \\ &= 1 - (1 - \alpha)^k \end{aligned}$$

so if the researcher tries five t-tests on five irrelevant variables, the probability that at least one of them will be statistically significant is $\approx 22.6\%$, not 5%. This is an instance of the *multiple comparison problem*; the problem is associated with a long literature in applied statistics (Lehmann, 1957*a*,*b*; Holm, 1979; Hochberg, 1988; Rom, 1990; Shaffer, 1995).

The same logic applies to testing one irrelevant variable in k different samples. Indeed, the canonical justification for frequentist hypothesis testing involves determining the sampling distribution of the test statistic, then calculating the probability that a particular value of the statistic will be generated by a sample of data produced under the null hypothesis. Thus, if a researcher takes a particular sample data set and randomly divides it into k subsamples, the probability of finding a statistically significant effect in at least one of these subsamples by chance is also $1 - (1 - \alpha)^k$.

[6] Incidentally, this statement is also true for a test for the statistical significance of the product term coefficient in a statistical model with interaction.

Interaction terms create a multiple comparison problem: the case of a dichotomous interaction variable between statistically independent regressors

It is not as commonly recognized that interacting two variables in a linear regression model effectively divides a sample into subsamples, thus creating the multiple comparison problem described above. The simplest and most straightforward example is a linear model with a continuous independent variable x interacted with a dichotomous independent variable $z \in \{0, 1\}$:

$$E[y|x, z] = \hat{y} = \hat{\beta}_0 + \hat{\beta}_x x + \hat{\beta}_z z + \hat{\beta}_{xz} xz \tag{1}$$

A researcher wants to know whether x has a statistically detectable relationship with y, as measured by the marginal effect of x on $E[y|x, z]$ from model (1): $\partial\hat{y}/\partial x$. Let $\widehat{ME}_x$ be shorthand notation for $\partial\hat{y}/\partial x$ and $\widehat{ME}_x^{z_0}$ be shorthand notation for $\partial\hat{y}/\partial x$ when $z = z_0$, where z_0 is any possible value of z. Because x is interacted with z, this means that the researcher needs to calculate confidence intervals for two quantities:

$$\left(\frac{\partial\hat{y}}{\partial x}|z = 0\right) = \widehat{ME}_x^0 = \hat{\beta}_x \tag{2}$$

$$\left(\frac{\partial\hat{y}}{\partial x}|z = 1\right) = \widehat{ME}_x^1 = \hat{\beta}_x + \hat{\beta}_{xz} \tag{3}$$

These (pointwise) confidence intervals can be created (i) by analytically calculating var $\left(\widehat{ME}_x^0\right)$ and var $\left(\widehat{ME}_x^1\right)$ using the asymptotically normal distribution of $\hat{\beta}$ and the variance-covariance matrix of the estimate, (ii) by simulating draws of $\hat{\beta}$ out of this distribution and constructing simulated confidence intervals of (2) and (3), or (iii) by bootstrapping estimates of $\hat{\beta}$ via repeated resampling of the data set and constructing confidence intervals using the resulting $\hat{\beta}$ estimates.

Common practice, and the practice recommended by Brambor, Clark and Golder (2006), is to report the estimated statistical and substantive significance of the relationship between

x and y at all values of the interaction variable z. Unfortunately, the practice inflates the probability of finding at least one statistically significant $\widehat{ME}_x^{z_0}$. A model with a dichotomous interaction term creates two significance tests in each of two subsamples, one for which $z = 0$ and one for which $z = 1$. This means that the probability that at least one statistically significant $\widehat{ME}_x^{z_0}$ will be found and reported under the null hypothesis that $ME_x^0 = ME_x^1 = 0$ is:

$$\begin{aligned} & \Pr(\text{false positive}) \\ = \; & \Pr\left(\widehat{ME}_x^0 \text{ is st. sig.} | ME_x^0 = 0\right) \vee \Pr\left(\widehat{ME}_x^1 \text{ is st. sig.} | ME_x^1 = 0\right) \\ = \; & 1 - \left(\Pr\left(\widehat{ME}_x^0 \text{ is not st. sig.} | ME_x^0 = 0\right) \wedge \Pr\left(\widehat{ME}_x^1 \text{ is not st. sig.} | ME_x^1 = 0\right)\right) \end{aligned}$$

If the two probabilities in the second term are unrelated, as when x and z are statistically independent and all β coefficients are fixed, then we can further reduce this term to:

$$\begin{aligned} & \Pr\left(\text{false positive}\right) \\ = \; & 1 - \left(\Pr\left(\widehat{ME}_x^0 \text{ is not st. sig.} | ME_x^0 = 0\right) * \Pr\left(\widehat{ME}_x^1 \text{ is not st. sig.} | ME_x^1 = 0\right)\right) \end{aligned}$$

where $ME_x^{z_0}$ is the true value of $\partial y / \partial x$ when $z = z_0$. If the test for each individual marginal effect has size α, this finally reduces to:

$$\Pr(\text{false positive}) = 1 - (1 - \alpha)^2 \tag{4}$$

The problem is immediately evident: the probability of accidentally finding at least one statistically significant $\widehat{ME}_x^{z_0}$ is no longer equal to α. For a conventional two-tailed $\alpha = 0.05$, this means there is a $1 - (1 - 0.05)^2 = 9.75\%$ chance of concluding that at least one of the marginal effects is statistically significant even when $ME_x^0 = ME_x^1 = 0$. Stated another way, the test is less conservative than indicated by α. The problem is even worse for a larger number of discrete interactions; if z has three categories, for example, there is a $1 - (1 - 0.05)^3 \approx 14.26\%$ chance of a false positive in this scenario.

To confirm this result, we conduct a simulation analysis to assess the false positive rate under the null for a linear regression model. For each of 10,000 simulations, 1,000 observations of a continuous dependent variable y are drawn from a linear model:

$$y = 0.2 + u$$

where $u \sim \Phi(0, 1)$. Covariates x and z are independently drawn from the uniform distribution between 0 and 1, with z dichotomized by rounding to the nearest integer. By construction, neither covariate has any relationship to y—that is, the null hypothesis is correct for both. We then estimate a linear regression of the form:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 z + \hat{\beta}_p xz$$

and calculate the predicted marginal effect $\widehat{ME}_x^{z_0}$ for the model when $z = 0$ and 1.

The statistical significance of the marginal effects $\widehat{ME}_x^{z_0}$ is assessed in three different ways. First, we use the appropriate analytic formula to calculate the variance of $\widehat{ME}_x^{z_0}$ using the variance-covariance matrix of the estimated regression; this is:

$$\text{var}\left(\widehat{ME}_x^{z_0}\right) = \text{var}\left(\hat{\beta}_x\right) + (z_0)^2 \,\text{var}\left(\hat{\beta}_{xz}\right) + 2z_0 \,\text{cov}\left(\hat{\beta}_x, \hat{\beta}_{xz}\right)$$

This enables us to calculate a pointwise 95% confidence interval using the critical t-statistic for a two-tailed $\alpha = 0.05$ test in the usual way. Second, we simulate 1000 draws out of the asymptotic (multivariate normal) distribution of $\hat{\beta}$ for the regression, calculate $\widehat{ME}_x^{z_0}$ at $z_0 = 0$ and 1 for each draw, and select the 2.5th and 97.5th percentiles of those calculations to form a 95% confidence interval. Finally, we construct 1000 bootstrap samples (with replacement) for each data set, then use the bootstrapped samples to construct simulated 95% confidence intervals.

The results for a model with a dichotomous z variable are shown in Table 1. The table

Table 1: Overconfidence in Interaction Effect Standard Errors of $ME_x = \partial y/\partial x$*

# of z categories	Calculation Method	Type I Error
2 categories	Simulated SE	9.86%
	Analytic SE	9.45%
	Bootstrap SE	10.33%
	Theoretical	9.75%
3 categories	Simulated SE	14.20%
	Analytic SE	13.93%
	Theoretical	14.26%
continuous	Simulated SE	14.51%
	Analytic SE	13.75%

*The reported number in the "Type I Error" column is the percentage of the time that a statistically significant (two-tailed, $\alpha = 0.05$) marginal effect $\partial y/\partial x$ for any z is detected in a model of the DGP from equation (1) under the null hypothesis where $\beta = 0$. Type I error rates calculated via simulated, analytic, or bootstrapped SEs using 10,000 simulated data sets with 1,000 observations each from the DGP $y = 0.2 + u$, $u \sim \Phi(0,1)$; $x \sim U[0,1]$, $z \in \{0,1\}$ with equal probability (2 categories), $z \in \{0,1,2\}$ with equal probability (3 categories), and $z \sim U[0,1]$ (continuous). For analytic SEs, se $\left(\widehat{ME}_x^{z_0}\right) = \sqrt{\text{var}\left(\hat{\beta}_x\right) + (z_0)^2 \text{var}\left(\hat{\beta}_{xz}\right) + 2z_0\text{cov}\left(\hat{\beta}_x, \hat{\beta}_{xz}\right)}$ and the 95% CI is $\left(\hat{\beta}_x + \hat{\beta}_{xz}z_0\right) \pm 1.96 * \text{se}\left(\widehat{ME}_x^{z_0}\right)$. Simulated SEs are created using 1000 draws out of the asymptotic (normal) distribution of $\hat{\beta}$ for the regression, calculating $\widehat{ME}_x^{z_0}$ for each draw, and selecting the 2.5th and 97.5th percentiles of those calculations to form a 95% confidence interval. Bootstrapped SEs are created using 1000 bootstrap samples (with replacement) for each data set, where the bootstrapped samples are used to construct simulated 95% confidence intervals. Theoretical false positive rates for discrete z are created using expected error rates from the nominal α value of the test as described in equation (4).

shows that, no matter how we calculate the standard error of the marginal effect, the probability of a false positive (Type I error) is considerably higher than the nominal $\alpha = 0.05$ and close to the theoretical expectation.

Continuous interaction variables

The multiple comparison problem and resulting overconfidence in hypothesis tests for marginal effects can be worsened when a linear model interacts a continuous independent variable x with a z variable that has more than two categories. For example, an interaction term between x and a continuous variable z implicitly cuts a given sample into many small subsamples for each value of z in the range of the sample. By subdividing the sample further, we create a larger number of chances for a false positive.

To illustrate the potential problem with overconfidence in models with more categories of z, we repeat our Monte Carlo simulation with statistically independent x and z variables using a three-category $z \in \{0, 1, 2\}$ (where each value is equally probable) and a continuous $z \in [0, 1]$ (drawn from the uniform distribution) instead of a discrete z. Bootstrapping is computationally intensive and yields no different results than the other two processes when z is dichotomous; we therefore only assess simulated and analytic standard errors for the 3 category and continuous z cases. The results are shown in Table 1.

As before, the observed probability of a Type I error is far from the nominal α probability of the test. A continuous z tends to have a higher false positive rate than a dichotomous z ($\approx 14\%$ compared to $\approx 10\%$ under equivalent conditions), and roughly equivalent to a three-category z.

Statistical interdependence between marginal effects estimates

In the above, we assumed that marginal effects estimates at different values of z are uncorrelated. But if $\widehat{ME}_x^0$ is related to $\widehat{ME}_x^1$ when z is dichotomous, then the probability of a

false positive result is:

$$\begin{aligned} & \Pr\left(\text{false positive}\right) \\ = \; & 1 - \left(\Pr\left(\widehat{ME}_x^0 \text{ is not st. sig.} | ME_x^0 = 0\right) \wedge \Pr\left(\widehat{ME}_x^1 \text{ is not st. sig.} | ME_x^1 = 0\right)\right) \end{aligned}$$

In this case, the probability of a false positive will be greater than or equal to the α value for each individual test. We would expect correlation between the statistical significance of marginal effects estimates when (for example) x and z are themselves correlated, or when β_x and β_{xz} are stochastic and correlated. If the two individual probabilities are perfectly correlated, then we expect their joint probability to be equal to either individual probability (α). In that case, the individual tests have correct size. As their correlation falls, the joint probability rises above α as the proportion of the time that one occurs without the other rises. When the correlation reaches zero, we have the result in Table 1.[7]

To illustrate the effect of correlated x and z on marginal effects estimates, Table 2 shows the result of repeating the simulations of Table 1 with varying correlation between the x and z variables. When z is dichotomous,[8] it appears that correlation between x and z is not influential on the false positive rate for ME_x; the false positive rate is near 9.8% (our theoretical expectation from Table 1) for all values of ρ_{xz}. This may be because the dichotomous nature of z creates the equivalent of a split sample regression, wherein $\widehat{ME}_x^1$ is quasi-independent from $\widehat{ME}_x^0$ despite the correlation between x and z. This interpretation is supported by the observed correlation between t-statistics for $\widehat{ME}_x^0$ and $\widehat{ME}_x^1$ in our simulation, which never exceeds 0.015 even when $|\rho_{xz}| \geq 0.9$. We conclude that it may be possible for $\widehat{ME}_x^0$ and $\widehat{ME}_x^1$ to be correlated in a way that brings the false positive rate

[7] In the event that the statistical significance of one marginal effect were negatively associated with the other—that is, if $\widehat{ME}_x^0$ were less likely to be significant when $\widehat{ME}_x^1$ is significant and vice versa—then the probability of a false positive could be even higher than that reported in Table 1. We believe that this is unlikely to occur in cases when β is fixed, as our results in Table 2 indicate that a wide range of positive and negative correlation between x and z does not produce false positive rates that exceed those of Table 1.

[8] Correlation between the continuous x and dichotomous z was created by first drawing x and a continuous $z^\star$ from a multivariate normal with mean zero and VCV $= \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, then choosing $z = 1$ with probability $\Phi(z^\star | \mu = 0, \sigma = 0.5)$.

Table 2: Overconfidence in Interaction Effect Standard Errors of $ME_x = \partial y/\partial x$*

	Type I Error (Analytic SE)		
		continuous z	
ρ_{xz}	binary z	uniform	normal
0.99	9.91%	7.29%	5.28%
0.9	9.26%	11.80%	6.42%
0.5	9.81%	14.06%	8.42%
0.2	9.78%	13.82%	8.87%
0	9.83%	13.69%	8.68%
-0.2	10.0%	13.60%	8.39%
-0.5	10.0%	13.81%	8.22%
-0.9	9.75%	11.57%	6.52%
-0.99	9.73%	7.61%	5.01%

*The reported number in the "Type I Error" column is the percentage of the time that a statistically significant (two-tailed, $\alpha = 0.05$) marginal effect $\partial y/\partial x$ for any z is detected in a model of the DGP from equation (1) under the null hypothesis where $\beta = 0$. Type I error rates are determined using 10,000 simulated data sets with 1,000 observations each from the DGP $y = 0.2 + u$, $u \sim \Phi(0, 1)$. When z is continuous, x and z are either (a) drawn from a multivariate distribution with uniform marginals and a multivariate normal copula with mean zero and VCV $= \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ (column "uniform"), or (b) drawn from the bivariate normal distribution with mean zero and VCV $= \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ (column "normal"). When z is binary, x and $z^\star$ are drawn from the bivariate normal with mean zero and VCV $= \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ and $\Pr(z = 1) = \Phi(z^\star | \mu = 0, \sigma = 0.5)$. Analytic SEs are used to determine statistical significance: $\text{se}\left(\widehat{ME}_x^{z_0}\right) = \sqrt{\text{var}\left(\hat{\beta}_x\right) + (z_0)^2 \,\text{var}\left(\hat{\beta}_{xz}\right) + 2z_0 \text{cov}\left(\hat{\beta}_x, \hat{\beta}_{xz}\right)}$ and the 95% CI is $\left(\hat{\beta}_x + \hat{\beta}_{xz} z_0\right) \pm 1.96 * \text{se}\left(\widehat{ME}_x^{z_0}\right)$.

closer to α, but that simple collinearity between x and a dichotomous z will not produce this outcome.

The results with a continuous z are more interesting. We look at two cases: one where x and z are drawn from a multivariate distribution with uniform marginal densities and a normal copula[9] (in the column labeled "uniform"), and one where x and z are drawn from a multivariate normal[10] distribution (in the column labeled "normal"). We see that the false positive rate indeed approaches the nominal $\alpha = 5\%$ for extreme correlations between x and z. Furthermore, we *also* see that the false positive rate when $\rho_{xz} = 0$ is about 8.7%; this is lower than the 13.69% false positive rate that we see in the uniformly distributed case (which is comparable to the 14.51% false positive rate that we observed in Table 1). It therefore appears that the false positive rate for marginal effects can depend on the distribution of x and z.[11]

Underconfidence is possible for conjoint tests of theoretical predictions

The analysis in the prior section asks how often we expect to see $\partial y/\partial x$ turn up statistically significant by chance when our analysis allows this marginal effect to vary with a conditioning variable z. Although we believe this is typically the right criterion against which to judge a significance testing regime, there are situations where it is a poor fit. For example, a theory with interaction relationships often makes multiple predictions; it may predict that $\partial y/\partial x < 0$ when $z = 0$ and $\partial y/\partial x > 0$ when $z = 1$. Such a theory is falsified if either prediction is not confirmed. This situation creates a different kind of multiple comparison

[9]This is accomplished using `rCopula` in the R package `copula`. The normal copula function has mean zero and VCV = $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$.

[10]The multivariate normal density has mean zero, VCV = $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$.

[11]Of course, when the correlation between x and z gets very large ($|\rho| > 0.9$), the problems that accompany severe multicollinearity may also appear (e.g., inefficiency); we do not study these problems in detail.

problem: if we use a significance test with size α on each subsample (one where $z = 0$ and one where $z = 1$), the joint probability that both predictions are simultaneously confirmed due to chance is much smaller than α and the resulting confidence intervals of the Brambor, Clark and Golder (2006) procedure are too wide. In this case, a researcher can achieve greater power to detect true positives without losing control over size by reducing the α of the individual tests.

Dichotomous interaction variable

Consider the model of equation (1), where a continuous independent variable x is interacted with a dichotomous independent variable $z \in \{0, 1\}$. A researcher might hypothesize that x has a statistically significant and positive relationship with y when $z = 0$, but no statistically significant relationship when $z = 1$. That researcher will probably go on to plot the marginal effects of equations (2) and (3). If the researcher's theory is correct, then (2) should be statistically significant and (3) should not.[12] If the null hypothesis is correct, so that both marginal effects equal zero, what is the probability that the researcher will find a positive, statistically significant marginal effect for equation (2) and no statistically significant effect for equation (3)? When $\widehat{ME}_x^0$ and $\widehat{ME}_x^1$ are statistically independent and $\alpha = 0.05$ for a

[12]This procedure raises an interesting and (to our knowledge) still debatable question: how does one test for the absence of a (meaningful) relationship between x and y at a particular value of z? We have phrased our examples in terms of expecting statistically significant relationships (or not), but a researcher will likely find zero in a 95% CI considerably more than 5% of the time even when the marginal effect $\neq 0$ (i.e., the size of the test will be larger than α). Moreover, a small but non-zero marginal effect could still qualify as the absence of a *meaningful* relationship. Alternative procedures have been proposed, but are not yet common practice (e.g., Rainey, 2014). We speculate that a researcher should properly test these hypotheses by specifying a range of ME_x^z consistent with "no meaningful relationship" and then determining whether the 95% CI intersects this range; this is the proposal of Rainey (2014). We assess the (somewhat unsatisfactory) status quo of checking whether 0 is contained in the 95% CI; the major consequence is that hypothesizing ME_x^z is not substantively meaningful for some z will not boost the power of a hypothesis-testing procedure as much as it might. The size of is already too small for conjoint hypothesis tests of this type, and so overconfidence is not a concern despite the excessive size of the individual test. In our corrected procedure, the size of the test is numerically controlled and therefore correctly set at α. See Suggestion 3 in the next section for more details of our corrected procedure.

one-tailed test, this probability must be:

$$
\begin{aligned}
& \Pr(\text{false positive}) \\
= \; & \Pr\left(\widehat{ME}_x^0 \text{ is stat. sig. and } > 0 | ME_x^0 = 0\right) \wedge \Pr\left(\widehat{ME}_x^1 \text{ is not stat. sig.} | ME_x^1 = 0\right) \\
= \; & \alpha\,(1 - 2\alpha) \\
= \; & 0.05 * 0.90 \\
= \; & 0.045
\end{aligned}
$$

That is, the probability of finding results that match the researcher's suite of predictions under the null hypothesis is 4.5%, a slightly *smaller* probability than that implied by α. In short, the α level is too conservative. Setting $\alpha \approx 0.0564$ yields a 5% false positive rate.

The situation is even better if a researcher hypothesizes that $ME_x^0 > 0$ and $ME_x^1 < 0$; in this case, when $\widehat{ME}_x^0$ and $\widehat{ME}_x^1$ are statistically independent and for a one-tailed test where $\alpha = 0.05$,

$$
\begin{aligned}
& \Pr(\text{false positive}) \\
= \; & \Pr\left(\widehat{ME}_x^0 \text{ is stat. sig. and } > 0 | ME_x^0 = 0\right) \wedge \Pr\left(\widehat{ME}_x^1 \text{ is stat. sig. and } < 0 | ME_x^1 = 0\right) \\
= \; & \alpha^2 = 0.05^2 = 0.0025
\end{aligned}
$$

That is, the probability of a false positive for this theory is one-quarter of one percent (0.25%), an extremely conservative test! Setting a one-tailed $\alpha = \sqrt{0.05} \approx .224$ corresponds to a false positive rate of 5%.

Perhaps the most important finding is that the underconfidence of the test—the degree to which the nominal α is larger than the actual probability of a false positive—is a function of the pattern of predictions being tested. This means that some theories are harder to "confirm" with evidence than others under a fixed α, and therefore our typical method for assessing how compatible a theory is with empirical evidence does not treat all theories equally.

Continuous interaction variable

The underconfidence problem can be more *or* less severe (compared to the dichotomous case) when z is continuous, depending on the pattern of predictions being tested. To determine the false positive rate when z is continuous, we ran the Monte Carlo simulation from Table 1 under the null ($\beta_x = \beta_z = \beta_{xz} = 0$) and checked for statistically significant marginal effects that matched a specified pattern of theoretical predictions using a two-tailed test, $\alpha = 0.05$. These results (along with simulations for binary z for comparison) are shown in Table 3. All the simulated false positive rates are smaller than the 5% nominal α, and all but one are smaller than the 2.5% one-tailed α to which a directional prediction corresponds. The degree of the test's underconfidence varies according to the pattern of predictions.

Thorough testing of possible hypotheses: underconfidence or overconfidence?

The tension between over- and underconfidence of empirical results is illustrated in a recent paper by Berry, Golder and Milton (2012) in the *Journal of Politics*. In that paper, Berry, Golder and Milton (2012) (hereafter BGM) recommend thoroughly testing all of the possible marginal effects implied by a statistical model. For a model like equation (1), that means looking not only at $\partial y/\partial x$ at different values of z, but also at $\partial y/\partial z$ at different values of x. Their reasoning is that ignoring the interaction between $\partial y/\partial z$ and x allows researchers to ignore implications of a theory that may be falsified by evidence:

> ...the failure of scholars to provide a second hypothesis about how the marginal effect of Z is conditional on the value of X, together with the corresponding marginal effect plot, means that scholars often subject their conditional theories to substantially weaker empirical tests than their data allow (653).

If BGM are describing holistic testing of a particular theory with a large number of predictions, then we believe that our analysis tends to support their argument. As we show above,

Table 3: Underconfidence in Confirmation of Multiple Predictions with Interaction Effects*

Predictions assessed	z type	Monte Carlo Type I Error
ME_x^z st. insig. $\mid z = 0$, $ME_x^z < 0 \mid z = 1$	binary	2.25%
$ME_x^z > 0 \mid z = 0$, $ME_x^z < 0 \mid z = 1$	binary	0.07%
ME_x^z st. insig. $\mid z < 0.5$, $ME_x^z < 0 \mid z \geq 0.5$	continuous	2.81%
$ME_x^z > 0 \mid z < 0.5$, $ME_x^z < 0 \mid z \geq 0.5$	continuous	0.49%
$ME_x^z > 0 \mid z < 0.5$, $ME_x^z < 0 \mid z \geq 0.5$, $ME_z^x > 0 \mid x < 0.5$, $ME_z^x < 0 \mid x \geq 0.5$	continuous	0.34%
$ME_x^z > 0 \mid z < 0.5$, $ME_x^z < 0 \mid z \geq 0.5$, $ME_z^x < 0 \mid x \in (-\infty, \infty)$	continuous	0.40%

*The "predictions assessed" column indicates how many distinct theoretical predictions must be matched by statistically significant findings in a sample data set in order to consider the null hypothesis rejected. The "z type" column indicates whether z is binary (1 or 0) or continuous ($\in [0, 1]$). The "Type I Error" column indicates the proportion of the time that the assessed predictions are matched and statistically significant (two-tailed, $\alpha = 0.05$, equivalent to a one-tailed test with $\alpha = 0.025$ for directional predictions) in a model of the DGP from equation (1) under the null hypothesis where $\beta_x = \beta_z = \beta_{xz} = 0$. Monte Carlo Type I errors are calculated using 10,000 simulated data sets with 1,000 observations each from the DGP $y = 0.2 + u$, $u \sim \Phi(0, 1)$. z and x are independently drawn from $U[0, 1]$ when z is continuous; when z is binary, it is drawn from $\{0, 1\}$ with equal probability and independently of x. Standard errors are calculated analytically: $\text{se}\left(\widehat{ME}_x^{z_0}\right) = \sqrt{\text{var}\left(\hat{\beta}_x\right) + (z_0)^2 \,\text{var}\left(\hat{\beta}_{xz}\right) + 2z_0 \text{cov}\left(\hat{\beta}_x, \hat{\beta}_{xz}\right)}$.

making multiple predictions about $\partial y/\partial x$ at different values of z lowers the chance of a false positive under the standard hypothesis testing regime. The false positive rate is even lower if we holistically test a theory using multiple predictions about both $\partial y/\partial x$ and $\partial y/\partial z$.

However, it is vital to note that following BGM's suggestion will *also* make it more likely that at least one marginal effect will appear as statistically significant by chance alone. The reason for this is relatively straightforward: testing a larger number of hypotheses means multiplying the risk of a single false discovery under the null hypothesis. In short, we contend that BGM are correct when testing a single theory by examining its multiple predictions as a whole, but caution that analyses that report any findings separately could be made more susceptible to false positives by this procedure.

What now? Determining and controlling the false positive rate for tests of interaction

The goal of this paper is evolutionary, not revolutionary. We do not argue for a fundamental change in the way that we test hypotheses about marginal effects estimated in an interaction model—viz., by calculating estimates and confidence intervals, and graphically assessing them—but we do believe that there is room to improve the interpretation of these tests. Specifically, we believe that the confidence intervals that researchers report should reflect an intentional choice. We suggest three best practices to help political scientists achieve this goal.

Suggestion 1: do not condition inference on the interaction term, as it does not solve the multiple comparison problem

A researcher's first inclination might be to fight the possibility of overconfidence by conditioning inference on the statistical significance of the interaction term. That is, for the case

when z is binary:

1. If $\hat{\beta}_{xz}$ is statistically significant: calculate $\widehat{ME}_x^0 = \hat{\beta}_x$ and $\widehat{ME}_x^1 = \hat{\beta}_x + \hat{\beta}_{xz}$ and interpret the statistical significance of each effect using the relevant 95% CI.

2. If $\hat{\beta}_{xz}$ is not statistically significant: drop xz from the model, re-estimate the model, calculate $\widehat{ME}_x^0 = \widehat{ME}_x^1 = \hat{\beta}'_x$, and base acceptance or rejection of the null on the statistical significance of $\hat{\beta}'_x$

However, this procedure results in an excess of false positives for $\widehat{ME}_x$. The reason is that a multiple comparison problem remains: the procedure allows two chances to conclude that $\partial y/\partial x \neq 0$, one for a model that includes xz and one for a model that does not.

Monte Carlo simulations reveal that the overconfidence problem with this procedure is substantively meaningful. Repeating the analysis of Table 1 with a binary $z \in \{0, 1\}$ under the null hypothesis ($\partial y/\partial x = 0$), conditioning inference on the statistical significance of the interaction term results in a 8.17% false positive rate when $\alpha = 0.05$ (two-tailed); the false positive rate is 9.60% under the continuous z.[13] This is less overconfident than the Brambor, Clark and Golder (2006) procedure using $\widehat{ME}_x$ only, which resulted in $\approx 10\%$ false positive rates, but still larger than the advertised α value. Therefore, we cannot recommend this practice as a way of correcting the overconfidence problem.

Suggestion 2: use tests designed to minimize false discoveries and false null findings

In cases where a researcher believes that the over- or underconfidence of traditional hypothesis test procedures may be decisive to a result (i.e., when results are at the margin of some threshold for statistical significance), s/he can use an alternative test procedure in order to minimize the probability of a false positive (when overconfidence is a potential problem) or a false null finding (when underconfidence is the relevant threat).

[13] These numbers are calculated using simulation-based standard errors.

Overconfidence corrections for estimated marginal effects

When a multiple comparison problem creates the danger of excess false rejections of the null, the literature supports two broad approaches to the problem. The first approach involves controlling the *false discovery rate* (FDR), or the number of rejected null hypotheses that are false as a proportion of the total number of statistically significant results (Benjamini and Hochberg, 1995, pp. 291-292). In the context of testing the statistical significance of $\widehat{ME}_x^z$ at multiple values of z, the FDR is the proportion of statistically significant values of $\widehat{ME}_x^z$ for which the null is actually true ($ME_x^z = 0$) in repeated tests. The second approach involves controlling the *familywise error rate* (FWER), or the proportion of the time that a set of multiple comparisons (a "family" of hypothesis tests) will produce at least one false rejection of the null hypothesis (Abdi, 2007, pp. 2-4). For testing $\widehat{ME}_x^z$ at multiple values of z, the FWER is the proportion of the time (in repeated tests) in which at least one $\widehat{ME}_x^z$ is statistically significant when the true $ME_x^z = 0$. In general, a test that sets the FWER at some value is a more conservative procedure than a test that limits the FDR to the same value: a single rejection of any hypothesis where the null is true in a set of multiple comparisons raises the FWER, whereas the FDR allows a fixed level of false positive hypothesis tests as a proportion of all statistically significant results. Consequently, procedures that control the FWER tend to be less powerful than those which control the FDR (Benjamini and Hochberg, 1995, p. 290).

A researcher can control the FDR by adapting the procedure of Benjamini and Hochberg (1995, p. 293-294; see also Spahn and Franco, 2015). For a categorical interaction variable z with m categories, the procedure suggests that the researcher should order each of the m values of $\widehat{ME}_x^{z_i}$ from $i = 1...m$ according to the magnitude of their p-values, $p_1, p_2, ..., p_m$ then find the largest k that satisfies $p_k < \alpha\frac{k}{m}$. The researcher then rejects the null hypothesis for all $\widehat{ME}_x^{z_j}$ from $j = 1...k$ at level α; this procedure ensures that the FDR is no larger than α, though it can (in some cases) be smaller (see Theorem 1 in Benjamini and Hochberg,

1995).[14] To visually depict which marginal effects are statistically significant, a researcher can use the critical t statistic $t^\star$ corresponding to $\alpha\frac{k}{m}$ when constructing a 95% CI using $\hat{\beta} \pm t^\star * \text{se}\left(\widehat{ME}_x^z\right)$ at all values of z. Note that this procedure also imposes a weak limit on the FWER: when all null hypotheses are true, or $(\partial y/\partial x \,|\, z = z_i) = 0$ for all z_i, the FDR is equivalent to the FWER (Benjamini and Hochberg, 1995, p. 291). The procedure to find an appropriate FDR-controlling $t^\star$ is included as a part of our `interactionTest` R library.

For controlling the FWER, Kam and Franzese (2007, pp. 43-51) recommend conducting a joint F-test to determine whether $\widehat{ME}_x^z \neq 0$ for any value of z when interaction between x and z (or other variables) is suspected. For a simple linear DGP with two variables of interest, this means running two models:

1. $\hat{y} = \hat{\beta}_0 + \hat{\beta}_x x + \hat{\beta}_z z + \hat{\beta}_{xz} xz$
2. $\hat{y} = \hat{\beta}_0 + \hat{\beta}_z z$

Then, the researcher can use an F-test to see whether the restrictions of model (2) can be rejected by the data. If so, the researcher can proceed to construct, plot, and interpret $\widehat{ME}_x^z$ using the procedure described in Brambor, Clark and Golder (2006).[15]

We used both of these procedures on the simulated data from Table 2; in each case, we set the target α value of the procedure to $\alpha = 0.05$, two-tailed. The results are shown in Table 4. Because all the null hypotheses are true in the simulated data set (that is, $\widehat{ME}_x^{z_i} = 0$ for all z_i), both the procedures should yield roughly equivalent results (because the FDR in

[14] Our explanation of the Benjamini and Hochberg (1995) procedure borrows from the surprisingly good description available on Wikipedia (as of 8/21/2015), which also contains an excellent summary of the false discovery rate and its relationship to the multiple comparison problem. This page is available at `https://en.wikipedia.org/wiki/False_discovery_rate`.

[15] A joint F-test of coefficients is a direct test for the statistical significance of $\partial\hat{y}/\partial x = \hat{\beta}_x + \hat{\beta}_{xz} z$ against the null that they all equal 0. For a generalized linear model with a non-linear link, this relationship between coefficients and marginal effects is not direct. Therefore, an F-test for restriction in these models may not correspond to a test for the statistical significance of marginal effects for the same reason that the statistical significance of coefficients in non-interaction relationships in a GLM does not necessarily indicate the statistical significance of marginal effects (Berry, DeMeritt and Esarey, 2010). In the case of underconfident tests, the null is that at least one marginal effect does not match the theoretical prediction; this does not correspond to the null of the F-test, that all marginal effects equal zero. Thus the F-test is inappropriate in this scenario. In this case, the bootstrapping procedure described in the next subsection can be adapted to limit the FWER to 5% for a single marginal effect.

Table 4: FDR and FWER control results for $ME_x = \partial y/\partial x$*

	FDR			FWER (F-test)		
		continuous z			continuous z	
ρ_{xz}	binary z	uniform	normal	binary z	uniform	normal
0.99	0.0498	0.0294	0.0432	0.0487	0.0343	0.0277
0.9	0.0478	0.0319	0.0359	0.0468	0.0470	0.0296
0.5	0.0495	0.0365	0.0322	0.0448	0.0538	0.0376
0.2	0.0513	0.0323	0.0290	0.0476	0.0480	0.0375
0	0.0525	0.0345	0.0339	0.0488	0.0517	0.0396
-0.2	0.0509	0.0320	0.0309	0.0478	0.0494	0.0378
-0.5	0.0504	0.0353	0.0318	0.0493	0.0531	0.0366
-0.9	0.0502	0.0313	0.0344	0.0481	0.0462	0.0286
-0.99	0.0503	0.0324	0.0413	0.0482	0.0339	0.0226

*The reported number in the "FDR" column is the percentage of the time that a statistically significant (two-tailed, $\alpha = 0.05$) marginal effect $\partial y/\partial x$ for any z is detected in a model of the DGP from equation (1) under the null hypothesis where $\beta = 0$ using the procedure of Benjamini and Hochberg (1995). The reported number in the "FWER" column is the percentage of the time that a statistically significant (two-tailed, $\alpha = 0.05$) marginal effect $\partial y/\partial x$ for any z is detected in a model of the DGP from equation (1) under the null hypothesis where $\beta = 0$ *and simultaneously* where an F-test for the joint significance of β_x and β_{xz} has been passed (two-tailed, $\alpha = 0.05$); this procedure is recommended by Kam and Franzese (2007). Figures are determined using 10,000 simulated data sets with 1,000 observations each from the DGP $y = 0.2 + u$, $u \sim \Phi(0, 1)$. When z is continuous, x and z are either (a) drawn from a multivariate distribution with uniform marginals and a multivariate normal copula with mean zero and VCV $= \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ (column "uniform"), or (b) drawn from the bivariate normal distribution with mean zero and VCV $= \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ (column "normal"). When z is binary, x and $z^\star$ are drawn from the bivariate normal with mean zero and VCV $= \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ and $\Pr(z = 1) = \Phi(z^\star|\mu = 0, \sigma = 0.5)$. Analytic SEs are used to determine statistical significance: $\text{se}\left(\widehat{ME}_x^{z_0}\right) = \sqrt{\text{var}\left(\hat{\beta}_x\right) + (z_0)^2 \text{var}\left(\hat{\beta}_{xz}\right) + 2z_0\text{cov}\left(\hat{\beta}_x, \hat{\beta}_{xz}\right)}$ and the 95% CI is $\left(\hat{\beta}_x + \hat{\beta}_{xz}z_0\right) \pm t_{FDR} * \text{se}\left(\widehat{ME}_x^{z_0}\right)$ for the FDR and $\left(\hat{\beta}_x + \hat{\beta}_{xz}z_0\right) \pm 1.96 * \text{se}\left(\widehat{ME}_x^{z_0}\right)$ for the FWER. The value of t_{FDR} is determined by following the Benjamini and Hochberg (1995) procedure for controlling the false discovery rate (as described in the text), then setting $t_{FDR} = \alpha\frac{k}{m}$ for the appropriate value of k; for continuous values of z, the number of points z_i at which $\partial y/\partial x|z_i$ is used for m (we use 11 points in our simulations).

this case is equivalent to the FWER). Indeed, as the table indicates, both procedures are effective at limiting false rejections of the null to a probability of $\lessapprox \alpha$.

Underconfidence corrections for estimated marginal effects

As noted above, the Brambor, Clark and Golder (2006) procedure is underconfident whenever a researcher is trying to conduct a conjoint test of multiple interaction relationships predicted by a pre-existing theory. Consequently, the appropriate critical t value to set a 5% probability of falsely rejecting the null of this conjoint test when examining confidence intervals is not the typical $t = 1.96$ (for $n \to \infty$). Instead, we suggest a nonparametric bootstrapping approach to hypothesis testing that chooses the appropriate critical t.

The procedure is simple:

1. For a particular data set, run a model $\hat{y} = G\left(\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 z + \hat{\beta}_2 xz + \textbf{controls}\right)$ with link function G. Calculate $\widehat{ME}_x^{z_0}$, $\widehat{ME}_z^{x_0}$, and their standard errors for multiple values of z_0 and x_0 using the fitted model.

2. Draw (with replacement) a random sample of data from the data set.

3. Run the model $\hat{y} = G\left(\tilde{\beta}_0 + \tilde{\beta}_1 x + \tilde{\beta}_2 z + \tilde{\beta}_2 xz + \textbf{controls}\right)$ on the bootstrap sample from step 2. Calculate $ME_x^{z_0}$, $\widetilde{ME}_z^{x_0}$, and their standard errors using the model. (The tilde distinguishes the bootstrap replicates from the hat used for estimates on the original sample.)

4. Calculate $\tilde{t}_x^{z_0} = \frac{\widetilde{ME}_x^{z_0} - \widehat{ME}_x^{z_0}}{se(\widetilde{ME}_x^{z_0})}$ and $\tilde{t}_z^{x_0} = \frac{\widetilde{ME}_z^{x_0} - \widehat{ME}_z^{x_0}}{se(\widetilde{ME}_z^{x_0})}$ for all values of z_0 and x_0. (Subtracting $\widehat{ME}_x^{z_0}$ or $\widehat{ME}_z^{x0}$ allows us to determine the distribution of t under the null hypothesis.)

5. Repeat steps 2-4 many times.

6. Using the bootstrapped values of t_x and t_z, find a critical t statistic $t^\star$ such that the theoretical predictions are met α proportion of the time when the null hypothesis is

true. For example, if a theory predicts that $ME_x > 0 | z > z_0$ and $ME_z < 0 | x > x_0$, $t^\star$ would satisfy $\Pr\left[\left(\exists z > z_0 : \tilde{t}_x^z > t^\star\right) \wedge \left(\exists x > x_0 : \tilde{t}_z^x < -t^\star\right)\right] = \alpha$.

7. Use the $t^\star$ to construct plots of $\widehat{ME}_x$ and/or $\widehat{ME}_z$ with confidence intervals; for $\widehat{ME}_x$, these confidence intervals are given by $\widehat{ME}_x^{z_0} \pm t^\star * \text{se}\left(\widehat{ME}_x^{z_0}\right)$.

We provide R code to implement this procedure for generalized linear models as a part of our `interactionTest` R library.

We tested the effectiveness of the nonparametric bootstrapping procedure in 1,000 simulated data sets with $N = 1000$ observations each under the null hypothesis for four different patterns of theoretical predictions; these theoretical predictions, the rejection rate of the bootstrapping procedure, and the median critical t found by the bootstrapping procedure are shown in Table 5. We also show the proportion of the time that using the critical t statistic generated from the bootstrapping procedure results in a rejection of the null hypothesis (note that the null is true in all of our simulated data sets). The table shows that different patterns of predictions have a different probability of appearing by chance, which in turn necessitate a different critical t statistic; furthermore, this critical t changes according to the correlation between x and z. Indeed, some patterns are so unlikely under some conditions that *any* estimates matching the pattern are not ascribable to chance, regardless of their uncertainty. The procedure results in false positive rates that match the nominal 5% rate targeted by the test.

Suggestion 3: maximize empirical power by specifying strong theories in advance

Correcting for the overconfidence of conventional 95% confidence intervals when performing interaction tests does come at a price: when the null hypothesis is *false*, the sensitivity of the corrected test is necessarily less than that of an uncorrected test. This tradeoff is fundamental to all hypothesis tests and not specific to the analysis of interaction: lowering the size of

Table 5: Median bootstrapped t-statistics for holistic testing of theoretical predictions, $\alpha = 0.05$*

Predictions assessed	statistic	ρ 0	-0.2	-0.5	-0.9	-0.99
one insignificant, one directional	median critical t	1.13	1.13	1.1	1.07	0.58
e.g.: ME_x^z st. insig. $\mid z < 0.5$, $ME_x^z < 0 \mid z \geq 0.5$	null rejection rate	0.04	0.05	0.04	0.04	0.04
opposite-sign directional predictions	median critical t	1.35	1.34	1.24	0.77	0
$ME_x^z > 0 \mid z < 0.5$, $ME_x^z < 0 \mid z \geq 0.5$	null rejection rate	0.06	0.03	0.06	0.04	0.08
opposite-sign directional predictions for both ME_x^z and ME_z^x	median critical t	1.24	1.24	1.17	0.68	0
e,g,: $ME_x^z > 0 \mid z < 0.5$, $ME_x^z < 0 \mid z \geq 0.5$, $ME_z^x > 0 \mid x < 0.5$, $ME_z^x < 0 \mid x \geq 0.5$	null rejection rate	0.06	0.03	0.05	0.04	0.07
opposite-sign directional predictions for one variable, constant directional prediction for other variable; e.g.: $ME_z^x < 0$, $ME_x^z > 0 \mid z < 0.5$, $ME_x^z < 0 \mid z \geq 0.5$	median critical t	1.30	1.29	1.22	0.74	0
	null rejection rate	0.07	0.03	0.05	0.04	0.08

*The "predictions assessed" column indicates how many distinct theoretical predictions must be matched by statistically significant findings in a sample data set in order to consider the null hypothesis rejected. The critical t row indicates the median nonparametrically bootstrapped t-statistic found to yield a 5% statistical significance rate for the predictions assessed. The rejection rate row gives the proportion of the time that the null hypothesis is rejected in the 1000 simulated data sets when using the bootstrapped t statistic. The DGP is $y = \varepsilon$, with $\varepsilon \sim \Phi(0, 1)$; in each data set, a model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_x x + \hat{\beta}_z z + \hat{\beta}_{xz} xz$ is fitted to the data. The value of ρ in the column indicates the correlation between x and z, which are drawn from the multivariate normal distribution with mean = 0 and variance = 1; results for values of $\rho > 0$ were similar to those for values of $\rho < 0$ with the same absolute magnitude.

the test, as we do by setting the FDR or FWER to equal 0.05, weakens the power of a test to detect relationships when they are actually there. On the other hand, correcting for underconfidence when simultaneously testing multiple theoretical predictions makes (jointly) confirming these predictions easier. As a result, we suggest that researchers generate and simultaneously test multiple empirical predictions whenever possible to maximize the power of their empirical test. For interaction terms, this means:

1. predicting the existence and direction of a marginal effect for multiple values of the intervening variable, and/or

2. predicting the existence and direction of the marginal effect of both constituent variables in an interaction.

These suggestions are subject to one important caveat: the predictions must be made before consulting sample data in order for the lowered confidence thresholds to apply. The lowered significance thresholds are predicated on the likelihood of simultaneous appearance of a particular combination of results under the null hypothesis, not on the joint likelihood of many possible combinations of results.

Application: Rehabilitating "Rehabilitating Duverger's Law" (Clark and Golder, 2006)

After publishing their recommendations for the proper hypothesis test of a marginal effect in the linear model with interaction terms, Clark and Golder (2006) went on to apply this advice in a study of the relationship between the number of political parties in a polity and the electoral institutions of that polity. Their reassessment of Duverger's Law applies the spirit behind the simple relationship between seats and parties predicted by Duverger to specify a microfoundational mechanism by which institutions and sociological factors are linked to political party viability. Based on a reanalysis of their results with the methods

that we propose, we believe that some of the authors' conclusions are more uncertain than originally believed.

Clark and Golder (2006) expect that ethnic heterogeneity (one social pressure for political fragmentation) will have a positive relationship with the number of parties that gets larger as average district magnitude increases. Specifically, they propose:

> "Hypothesis 4: Social heterogeneity increases the number of electoral parties only when the district magnitude is sufficiently large" (Clark and Golder, 2006, p. 694).

We interpret their hypothesis to mean that the marginal effect of ethnic heterogeneity on the number of electoral parties should be positive when district magnitude is large, and statistically insignificant when district magnitude is small. To test for the presence of this relationship, the authors construct plots depicting the estimated marginal effect of ethnic heterogeneity on number of parties at different levels of district magnitude for a pooled sample of developed democracies, for 1980s cross-sectional data (using the data from Amorim Neto and Cox (2007)), and for established democracies in the 1990s. In all three samples, they find that ethnic heterogeneity has a positive and statistically significant effect on the number of parties once district magnitude becomes sufficiently large.

Figure 2 displays our replications of the marginal effects plots from Clark and Golder (2006). We show three different confidence intervals: (i) the authors' 90% confidence intervals (using a conventional t-test), (ii) a 90% CI with a nonparametrically bootstrapped critical t designed to set the false positive rate at 5% for the pattern of predictions where $ME_x^{z<2.5}$ is statistically insignificant and $ME_x^{z\geq 2.5} > 0$, which we call the "prediction-corrected" CI, and (iii) a 90% CI constructed using the FDR-controlling procedure of Benjamini and Hochberg (1995). We also calculate and show the results of a joint-F test as prescribed by Kam and Franzese (2007).

None of the joint F-tests for the statistical significance of the marginal effect of ethnic heterogeneity yield one-tailed p-values less than 0.1. Additionally, FDR-controlling 90%

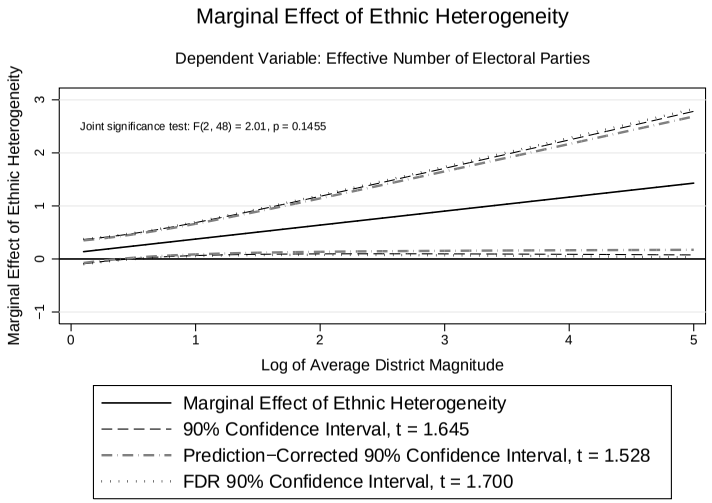
confidence intervals constructing using the procedure we describe above include zero across the entire range of district magnitude for the sample of established democracies in the 1990s. However, in the other two samples, the coverage of the 90% FDR confidence intervals confirms the authors' original results, albeit with somewhat greater uncertainty. In addition, the authors' original findings are statistically significant and consistent with their pattern of theoretical predictions when we employ the prediction-corrected 90% confidence intervals.

In summary, our analysis indicates that the authors' claims are most strongly supported by a combination of the empirical information they collect with the prior theoretical prediction of an unlikely pattern of results. Their results cannot be supported by a procedure that sets the FWER at 90%, and are only partially supported by a procedure that sets the FDR at 90%. We believe that this re-interpretation of the authors' findings is important for readers to understand in order for them to grasp the strength of the results and the assumptions upon which these results are based.

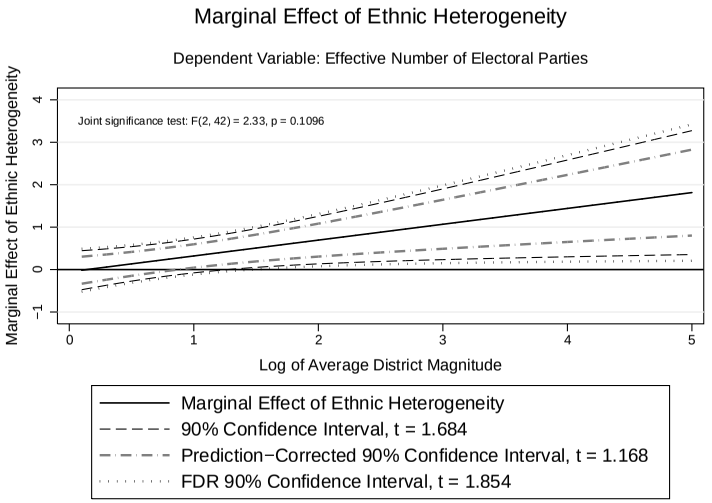
Conclusion

The main argument of this study is that, when it comes to the contextually conditional (interactive) relationships that have motivated a great deal of recent research, the Brambor, Clark and Golder (2006) procedure for testing for a relationship between x and y at different values of z does not effectively control the probability of a false positive finding. The probability of at least one relationship being statistically significant is higher than one expects because the structure of interaction models divides a data set into multiple subsets defined by z, each of which has a chance of showing evidence for a relationship between x and y under the null hypothesis. On the other hand, the possibility of simultaneously confirming multiple theoretical predictions by chance alone can be quite small because this requires a large number of individually unlikely events to occur together, making the combination of these events collectively even more unlikely. The consequence is that false positive rates may

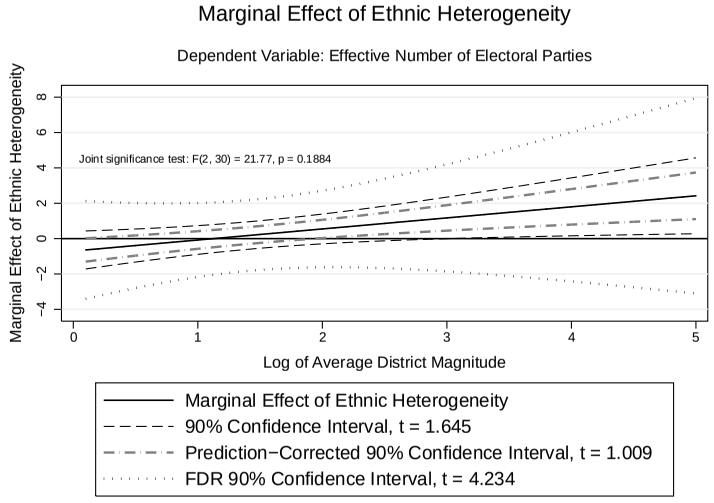
Figure 2: Marginal effect of ethnic heterogeneity on effective number of electoral parties (Figure 1 of Clark and Golder (2006)), with original and prediction- and discovery-corrected confidence intervals



(a) Pooled Analysis, Established Democracies



(b) 1980s, Amorim Neto and Cox



(c) 1990s, Established Democracies

be considerably higher *or* lower than researchers believe when they conduct their tests. A further consequence is that researchers using this procedure are implictly applying inconsistent standards to assess whether evidence tends to support or undermine a theory when that theory makes multiple empirical predictions.

Fortunately, we believe that specifying a consistent false positive rate for the discovery of interaction relationships is a comparatively simple matter of following a few rules of thumb:

1. do not condition inference about marginal effects on the statistical significance of the product term alone;

2. if a relationship is close to statistical significance under conventional tests, use procedures that control the overall false discovery rate and/or familywise error rate, such as the sequential test procedure of Benjamini and Hochberg (1995) or the joint F-test recommended by Kam and Franzese (2007); and

3. if possible, generate multiple hypotheses about contextual relationships before consulting the sample data and test them as a group using a nonparametric bootstrapping procedure to generate the appropriate critical t value, because it maximizes the power of the study.

None of these recommendations constitutes a fundamental revision to the way we conceptualize or depict conditional relationships. Rather, they allow us to ensure that evidence we collect is compared to a counterfactual world under the null hypothesis in a controlled fashion and consistent with the hypothesis tests that we perform in other situations. All of our recommendations can be implemented in standard statistical packages; we hope that researchers will keep them in mind when embarking on future work involving the assessment of conditional marginal effects.

References

Abdi, Herve. 2007. The Bonferonni and Sidak Corrections for Multiple Comparisons. In *Encyclopedia of Measurement and Statistics*, ed. Neil Salkind. Thousand Oaks, CA: Sage pp. 103–106.

Ai, Chunron and Edward C. Norton. 2003. "Interaction tterm in logit and probit models." *Economics Letters* pp. 123–129.

Amorim Neto, Octavio and Gary W. Cox. 2007. "Electoral Institutions, Cleavage Structures, and the Number of Parties." *American Journal of Political Science* 41:149–174.

Benjamini, Y. and Y. Hochberg. 1995. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 289–300.

Berry, W.D., J.H.R. DeMeritt and J. Esarey. 2010. "Testing for interaction in binary logit and probit models: is a product term essential?" *American Journal of Political Science* 54(1):248–266.

Berry, William, Matthew Golder and Daniel Milton. 2012. "Improving Tests of Theories Positing Interaction." *Journal of Politics* 74(3):653–671.

Brambor, Thomas, WR Clark and Matthew Golder. 2006. "Understanding interaction models: Improving empirical analyses." *Political Analysis* pp. 1–20.

Braumoeller, Bear F. 2004. "Hypothesis Testing and Multiplicative Interaction Terms." *International Organization* 58(4):807–820.

Clark, WR and Matthew Golder. 2006. "Rehabilitating Duverger's theory." *Comparative Political Studies* 39(6):679–708.

Hochberg, Y. 1988. "A sharper Bonferroni procedure for multiple tests of significance." *Biometrika* 75(4):800–802.

Holm, S. 1979. "A simple sequentially rejective multiple test procedure." *Scandinavian Journal of Statistics* 6:65–70.

Kam, Cindy D. and Robert J. Franzese. 2007. *Modeling and interpreting interactive hypotheses in regression analysis*. University of Michigan Press.

King, Gary, Michael Tomz and Jason Wittenberg. 2000. "Making the most of statistical analyses: Improving interpretation and presentation." *American Journal of Political Science* 44(2):347–361.

Lehmann, E.L. 1957*a*. "A theory of some multiple decision problems, I." *The Annals of Mathematical Statistics* pp. 1–25.

Lehmann, E.L. 1957*b*. "A theory of some multiple decision problems. II." *The Annals of Mathematical Statistics* 28(3):547–572.

Rainey, Carlisle. 2014. "Arguing for a Negligble Effect." *American Journal of Political Science* 58(4):1083–1091.

Rom, D.M. 1990. "A sequentially rejective test procedure based on a modified Bonferroni inequality." *Biometrika* 77(3):663–665.

Shaffer, JP. 1995. "Multiple hypothesis testing." *Annual Review of Psychology* 46:561–584.

Sidak, Z. 1967. "Rectangular confidence regions for the means of multivariate normal distributions." *Journal of the American Statistical Association* 62:626–633.

Spahn, Bradley and Annie Franco. 2015. "A False Discovery Framework for Mitigating Publication Bias." Online. URL: http://polmeth.wustl.edu/mediaDetail.php?docId=1617.