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

Justin Esarey† and Jane Lawrence Sumner‡

August 18, 2017

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. Our paper makes three contributions. First, we demonstrate that the Brambor, Clark, and Golder approach produces statistically significant findings when $\partial y/\partial x = 0$ at a rate that can be many times larger or smaller than the nominal false positive rate of the test. Second, we introduce the `interactionTest` software package for R to implement procedures that allow easy control of the false positive rate. Finally, we illustrate our findings by replicating an empirical analysis of the relationship between ethnic heterogeneity and the number of political parties from *Comparative Political Studies*.

*Nathan Edwards provided research assistance while writing this paper, for which we are grateful. We received helpful feedback on earlier versions of this paper from Kyle Beardsley, William D. Berry, Christopher Gandrud, Tom Pepinsky, Meg Shannon, anonymous reviewers, and participants in our panel at the 2012 Annual Meeting of the Society for Political Methodology and in a 2012 presentation in the Emory Political Science Colloquium Series.

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

‡Assistant Professor, Department of Political Science, University of Minnesota.

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. That is, although a statistically significant product term is sufficient for concluding that $\partial y/\partial x$ is different at different values of z (Kam and Franzese, 2007, p. 50), it cannot tell us whether $\partial y/\partial x$ is statistically distinguishable from zero at any *particular* value of z.

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

[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.

> interaction term simply by looking at the significance of the coefficient on the interaction term. Numerous articles ignore this point and drop interaction terms 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.[2] 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 3300 times as of August 2016. Similar advice is given in Braumoeller (2004, pp. 815-818, esp. Figure 2), which has been cited over 660 times in the same time frame.

Our paper makes three contributions to the study of interactive relationships. First, we highlight a hazard with the Brambor, Clark, and Golder 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 that $\partial y/\partial x = 0$) has a separate probability of a false positive (i.e., falsely rejecting the null hypothesis when the null is true). For example, if z is dichotomous ($z \in \{0, 1\}$), estimating a model like $y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz$ is analogous to estimating $y = \beta_0 + \beta_1 x$ twice, once for data where $z = 0$ and once for data where $z = 1$, with two chances

[2] 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."

for β_1 to be found statistically significant by chance. A similar problem is already well-recognized in the analysis of variance for nominal treatment factors (e.g., Kutner et al., 2004, Section 19.9). 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 for each individual value of z 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 be too wide or too narrow to conduct the tests that scholars wish to perform:[3] 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 scenario may occur 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.

Second, we offer researchers guidance on strategies that are effective and ineffective at controlling the false positive rate when examining interaction relationships. Our primary recommendation is for researchers to simply be aware that marginal effects plots generated using 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 (R

[3]Note that "appropriate" width of a confidence interval is relative to the test with which the interval is associated. The pointwise 95% CIs constructed by Brambor, Clark and Golder (2006) *do* include the true value of any given $(\partial y/\partial x \mid z = z_0)$ 95% of the time in repeated samples for a fixed z_0, as expected. However, these 95% CIs do *not* cover all the true values in a set of $(\partial y/\partial x \mid z \in \{z_0, z_1, ..., z_k\})$ 95% of the time; the CIs are too narrow in this case because they too frequently exclude $(\partial y/\partial x|z) = 0$ for at least one value of z when the null is true (that is, when $(\partial y/\partial x|z) = 0$ for all z). They also falsely reject the null too infrequently in a conjoint test of multiple theoretical predictions. For example, if a theory predicts that $(\partial y/\partial x|z = 0) > 0 \wedge (\partial y/\partial x|z = 1) < 0$, the null hypothesis that $(\partial y/\partial x|z = 0) \leq 0 \vee (\partial y/\partial x|z = 1) \geq 0$ will be rejected far less than 5% of the time by 95% pointwise confidence intervals when this null is true (because the CIs are too wide for this purpose).

[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,

Core Team, 2017) in our new `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 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.[6] In situations where marginal effects plots with pointwise confidence intervals (like those in Brambor, Clark and Golder (2006)) would be underconfident, such as when researchers are jointly testing multiple theoretical predictions, 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). The authors' original analysis, published in *Comparative Political Studies*, indicates that ethnic heterogeneity increases the number of political parties only when electoral district magnitude (in number of seats) is sufficiently large. Our re-analysis indicates that the authors' claims 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%. The strongest support for the authors' hypothesis comes from a procedure that jointly tests the authors' multiple predictions to achieve maximum power while controlling the joint false positive rate, illustrating (a) the usefulness of research designs that combine theory and empirics and (b) the sensitivity of Clark and Golder's results to pre-specification of theoretical expectations.

see Spahn and Franco (2015).

[6] As an example of this procedure, Braumoeller (2004, p. 814) recommends dropping a small and statistically uncertain interaction term in his reanalysis of Schultz (1999).

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 that $\partial y/\partial x = 0$ 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.

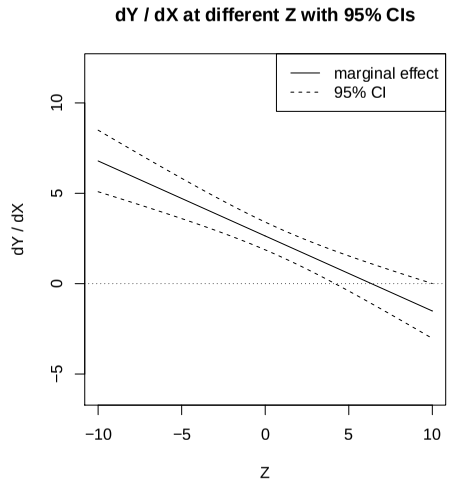
Our goal is to assess the false positive rate of this test procedure—that is, the proportion of the time that this procedure detects a statistically significant $\partial y/\partial x$ for at least one value of z when the null hypothesis that $(\partial y/\partial x|z) = 0$ is true for all z. If the false positive rate is greater than the nominal size of the test, α, then the procedure is overconfident: the confidence interval covers $(\partial y/\partial x|z) = 0$ for all z less than $(1 - \alpha)$ proportion of the time when the null is true. If the false positive rate is less than α, then the procedure is underconfident: the confidence interval could be narrower while still covering $(\partial y/\partial x|z) = 0$ for all z with probability $(1 - \alpha)$ when this null is true. 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 hypothesis that $(\partial y/\partial x|z) = 0$ for all values of z.

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 is only valid for a single t-test conducted on a single coefficient or marginal effect.[7] It is *not* valid for simultaneously

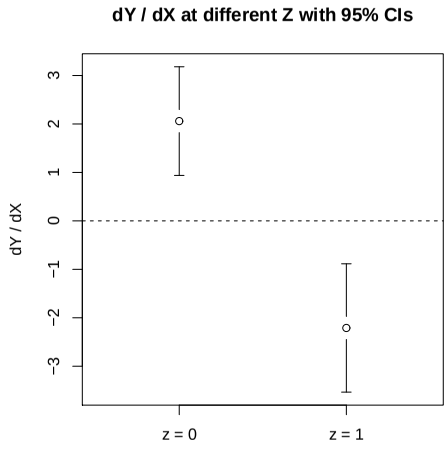
[7]Incidentally, this statement is also true for a test for the statistical significance of the product term

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, $N = 50$.

testing the statistical significance of multiple coefficients. 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 variables that have no relationship to the dependent variable into a regression, the probability that at least one of them comes up significant (in statistically independent tests) 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 when the null hypothesis of the test is true. 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 $1 - (1 - \alpha)^k$ when the null of no relationship is true and the hypothesis tests are statistically independent.

coefficient in a statistical model with interaction.

Interaction terms create a multiple comparison problem: the case of a dichotomous interaction variable

Interacting two variables in a linear regression model effectively divides a sample into subsamples, thus creating the multiple comparison problem described above. This is a well-recognized problem in the context of analysis of variance, where textbooks recommend multiple comparison adjustment when examining the marginal effect of one treatment condition whose effect is moderated by another treatment (e.g., Kutner et al., 2004, Section 19.9). 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 $(\partial\hat{y}/\partial x|z = z_0)$, 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 by doing any of the following: (i) by analytically calculating var $\left(\widehat{ME}_x^0\right)$ and var $\left(\widehat{ME}_x^1\right)$ using the variance-covariance matrix of the $\hat{\beta}$ estimates, (ii) by simulating draws of $\hat{\beta}$ out of the asymptotically normal distribution of $\hat{\beta}$ 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:

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

If the two marginal effects (and their associated statistical significance tests) in the second term are unrelated, as when the sample is split into two based on the value of z and a regression separately estimated on each subsample, then we can rewrite this as:

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

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 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 that $ME_x^{z_0} = ME_z^{x_0} = 0$ is correct for all values of z_0 and x_0. 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 at each value of z_0. Finally, we construct 1000 bootstrap samples (with replacement) for each data set, estimate $\hat{\beta}$ in each bootstrap sample, calculate $\widehat{ME}_x^{z_0}$ at $z_0 = 0$ and 1 using the $\hat{\beta}$ from each bootstrap sample, and use the 2.5th and 97.5th percentiles of the calculated marginal effects to construct a 95% confidence interval at each value of z_0.

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

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

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) when $\beta_x = \beta_z = \beta_{xz} = 0$. 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, $\text{se}\left(\widehat{ME}_x^{z_0}\right) = \sqrt{\text{var}\left(\hat{\beta}_x\right) + (z_0)^2 \text{var}\left(\hat{\beta}_{xz}\right) + 2z_0\text{cov}\left(\hat{\beta}_x, \hat{\beta}_{xz}\right)}$ and the 95% CI is $\left(\hat{\beta}_x + \hat{\beta}_{xz}z_0\right) \pm 1.96 * \text{se}\left(\widehat{ME}_x^{z_0}\right)$. 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 at each value of z_0. Bootstrapped SEs are created using 1000 bootstrap samples (with replacement) for each data set, estimating $\hat{\beta}$ in each bootstrap sample, calculating $\widehat{ME}_x^{z_0}$ using the $\hat{\beta}$ from each bootstrap sample, and using the 2.5th and 97.5th percentiles of the calculated marginal effects to construct a 95% confidence interval at each value of z_0. Theoretical false positive rates for discrete z are created using expected error rates for independent tests from the nominal α value of the test as described in equation (4).

three-category z.

Statistical interdependence between marginal effects estimates

In the section above, we assumed that marginal effects estimates (and related statistical significance tests) at different values of z are uncorrelated. But if the significance tests of $\widehat{ME}_x^0$ and $\widehat{ME}_x^1$ are related when z is dichotomous, 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. In this case, the probability of a false positive result is:

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

If $\left(\widehat{ME}_x^0 \text{ is not st. sig.} | ME_x^0 = 0\right)$ and $\left(\widehat{ME}_x^1 \text{ is not st. sig.} | ME_x^1 = 0\right)$ are perfectly correlated, then we expect the joint probability that both occur to be equal to either individual probability that one occurs $(1 - \alpha)$ and therefore $\Pr\left(\text{false positive}\right) = 1 - (1 - \alpha) = \alpha$. In that case, the individual tests have correct size. As their correlation falls, the joint probability that both occur falls below $(1 - \alpha)$ as the proportion of the time that one occurs without the other rises.[8]

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,[9] 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

[8] 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.

[9] 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) when $\beta_x = \beta_z = \beta_{xz} = 0$. 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 $\left(\hat{\beta}_x + \hat{\beta}_{xz} z_0\right) \pm 1.96 * \text{se}\left(\widehat{ME}_x^{z_0}\right)$.

theoretical expectation from Table 1) for all values of ρ_{xz}. This may be because the dichotomous nature of z creates a situation analogous to 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 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[10] (in the column labeled "uniform"), and one where x and z are drawn from a multivariate normal[11] 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.[12]

Underconfidence is possible for conjoint tests of theoretical predictions

The analysis in the prior section asks how often we expect to see $\partial\hat{y}/\partial x$ turn up statistically significant by chance when our analysis allows this marginal effect to vary with a conditioning

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

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

[12]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.

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; the null hypothesis is that either or both propositions are false, $(\partial y/\partial x|z = 0) \geq 0 \vee (\partial y/\partial x|z = 1) \leq 0$. This situation creates a different kind of multiple comparison 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 for this test. For example, in the situation noted above, 90% confidence intervals (corresponding to $\alpha = 0.05$ for a one-tailed test) will not include both $(\partial y/\partial x|z = 0) \geq 0$ and $(\partial y/\partial x|z = 1) \leq 0$ far less than 5% of the time when both are true. 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 be.[13] If our default expectation is that all marginal

[13]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 (e.g., Rainey, 2014), but are not yet common practice. 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)

effects are nonexistent ($ME_x^0 = ME_x^1 = 0$), 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) under these conditions?[14] When the statistical significance tests for $\widehat{ME}_x^0$ and $\widehat{ME}_x^1$ are statistically independent and $\alpha = 0.05$ for a one-tailed test, this probability must be:

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

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 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 2 in the next section for more details of our corrected procedure.

[14]For this theory's alternative hypothesis:

$$ME_x^0 > 0 \wedge ME_x^1 = 0$$

the appropriate null hypothesis is:

$$ME_x^0 \leq 0 \vee ME_x^1 \neq 0$$

We instead propose to assume a different default expectation when calculating the probability of a false positive:

$$ME_x^0 = ME_x^1 = 0$$

We make this assumption because this corresponds to the default expectation of "no relationship" that most political scientists bring to a study. Using the appropriate null hypothesis as our baseline would make our point even stronger, as then the probability of a false positive would then be:

$$\alpha(1 - \beta)$$

where β is the power of a test to reject the point null $ME_x^1 = 0$, typically much higher than α. But calculating β requires us to make assumptions about the probability distribution of a marginal effect's magnitude when it is not equal to zero; using the default expectation of $ME_x^z = 0$ for all values of z allows us to avoid making such restrictive and complicating assumptions. A uniform expectation of zero effects is also consistent with a proper null hypothesis for testing multiple directional hypotheses; for example, when testing the alternative hypothesis that:

$$ME_x^0 > 0 \wedge ME_x^1 < 0$$

the matching appropriate null hypothesis is:

$$ME_x^0 \leq 0 \wedge ME_x^1 \geq 0$$

and the frequentist supremum probability of a false positive is:

$$\Pr\left[\left(\widehat{ME}_x^0 \text{ is stat. sig. and } > 0 | ME_x^0 = 0\right) \wedge \left(\widehat{ME}_x^1 \text{ is stat. sig. and } < 0 | ME_x^1 = 0\right)\right]$$

as shown in the text and calculated in the tables.

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

That is, the probability of finding results that match the researcher's suite of predictions when both marginal effects are false 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 for this set of predictions when $ME_x^0 = ME_x^1 = 0$.

The situation is even better if a researcher hypothesizes that $ME_x^0 > 0$ and $ME_x^1 < 0$. In this case, when the statistical significance tests for $\widehat{ME}_x^0$ and $\widehat{ME}_x^1$ are independent and we conduct a one-tailed test where $\alpha = 0.05$ with a corresponding null hypothesis of $[ME_x^0 \leq 0 \vee ME_x^1 \geq 0]$, the largest possible probability of a false positive corresponding to the set of possible of null marginal effect values is:

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

That is, the largest 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 the Brambor, Clark and

Golder (2006) 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 a default expectation that all marginal effects were nonexistent ($\beta_x = \beta_z = \beta_{xz} = 0$) 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

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 conclude that the predictions are empirically consistent with the evidence. 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) when $\beta_x = \beta_z = \beta_{xz} = 0$. 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: $\text{se}\left(\widehat{ME}_x^{z_0}\right) = \sqrt{\text{var}\left(\hat{\beta}_x\right) + (z_0)^2 \,\text{var}\left(\hat{\beta}_{xz}\right) + 2z_0 \text{cov}\left(\hat{\beta}_x, \hat{\beta}_{xz}\right)}$.

> 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, 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. 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 statistically significant findings separately could be made much 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 (that $ME_x = 0$) on the statistical significance of $\hat{\beta}'_x$

Braumoeller (2004, p. 814), one of the foundational pieces in the political science literature concerning the analysis of interacted relationships, uses this procedure in reanalyzing work originally published by Schultz (1999). 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. We repeated the analysis of Table 1 with a binary $z \in \{0, 1\}$ under the null hypothesis (that $(\partial y/\partial x|z_0) = 0$ for all z), conditioning inference on the statistical significance of the interaction term. This procedure results in a 8.17% false positive rate when $\alpha = 0.05$ (two-tailed); the false positive rate is 9.60% when z is continuous.[15] 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.

[15] These numbers are calculated using simulation-based standard errors, as described in Table 1.

Suggestion 2: use tests designed to minimize false discoveries and maximize power

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 control the probability of a false positive (when overconfidence is a potential problem) or false disconfirmation of a theory that makes multiple predictions (when underconfidence is the relevant threat). We describe two separate test procedures, depending on whether the researcher believes overconfidence or underconfidence to be the likely problem. In this section, we will discuss each procedure in turn. In brief, for overconfidence we recommend adapting the Benjamini and Hochberg (1995) procedure to control the false discovery rate. For underconfidence, we suggest finding a critical t-statistic that produces a specified joint false positive rate using a nonparametric bootstrapping technique. Both of these procedures can be implemented using our R library, `interactionTest`.

Overconfidence corrections for estimated marginal effects

When a multiple comparison problem creates the danger of excess false discoveries, 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 (i.e., $ME_x^z = 0$) 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 statis-

tically 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 positives 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 for interacted relationships 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, their procedure suggests that the researcher should rank order each of the p-values, p_k for $k \in \{1...m\}$; p_1 is the smallest p value and p_m is the largest, with k the rank index. Then, find the largest rank, $k = k^\star$, that satisfies $p_k < \alpha\frac{k}{m}$. The researcher then rejects the null hypothesis for all $\widehat{ME}_x^{z_j}$ from $j = 1...k^\star$ 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).[16] To visually depict which marginal effects are statistically significant, a researcher can use the critical t-statistic $t^\star$ corresponding to $\alpha\frac{k^\star}{m}$ when constructing a 95% CI using $\hat{\beta} \pm t^\star * \text{se}\left(\widehat{ME}_x^{z_0}\right)$ at all values of z_0. Note that this procedure also imposes a weak limit on the FWER: when all null hypotheses are true, i.e. $(\partial y/\partial x \,|\, z = z_0) = 0$ for all values of z_0, the FDR is equivalent to the FWER (Benjamini and Hochberg, 1995, p. 291).

Put another way, this procedure orders the p-values for all relevant values of z, and determines how many rejections of the null hypothesis can be made such that all p-values for the rejected hypotheses are less than the value of α multiplied by $\frac{k^\star}{m}$. The $\frac{1}{m}$ multiplier is a Bonferroni-type adjustment for multiple comparisons; this multiplier 'deflates' α to account for the joint probability of at least one false positive when m-many tests are conducted (Benjamini and Hochberg, 1995, p. 293). The innovation of Benjamini and Hochberg (1995)

[16] 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`.

is to shift the statistic of interest to the proportion of rejected null hypotheses for which the null is true (instead of the probability of at least one rejection). This allows us to throw out highly statistically insignificant results from consideration in declining order of p-value, starting with $k = m$, until we find $k^\star$. As each subsequent p-value is discarded and k gets smaller, the size target ($\alpha\frac{k}{m}$) to which all remaining p-values are compared also gets smaller. The process stops when all p-values are less than the deflated size target, $\alpha\frac{k^\star}{m}$, which is then used to find a critical t-statistic. This critical t-statistic, $t^\star$, can then be used to construct marginal effects plots with confidence intervals visually similar to those of Brambor, Clark and Golder (2006); if a researcher uses these CIs to test multiple hypotheses, at most α proportion of the rejected null hypotheses will be false; the FDR is controlled at α. If $k^\star = 1$, the Bonferroni and Benjamini-Hochberg deflation factors are identical. The procedure to find an appropriate FDR-controlling $t^\star$ for marginal effects calculated from an interaction model is included as a part of the new `interactionTest` R library that we developed for this paper.

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).[17]

[17]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 $\beta_x = \beta_{xz} = 0$. 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 sta-

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) when $\beta_x = \beta_z = \beta_{xz} = 0$ 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) when $\beta_x = \beta_z = \beta_{xz} = 0$ *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} to the t-statistic with a critical value of $\alpha \frac{k}{m}$ for the appropriate value of k; for continuous values of z, m is equal to the number of points z_0 at which we assess $\partial y/\partial x|z_0$ (we use 11 points in our simulations).

We used both of these procedures on the simulated data from Table 2; in each case, we set the target false positive rate (FDR or FWER) of the procedure to 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, $\widehat{ME}_x^{z_0} = 0$ for all z_0), both the procedures should yield roughly equivalent results (because the FDR in 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 $\lessapprox 5\%$.

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 intuition behind our approach is simple: using simulation, determine a critical $t^\star$ statistic that will produce joint confirmation of all a theory's marginal effect predictions α proportion of the time *when in fact all the marginal effects are zero.* If we use this $t^\star$ to construct confidence intervals for marginal effects plots (using the ordinary formula for confidence intervals and the analytically calculated standard errors from the original model) in the style of Brambor, Clark and Golder (2006), we will be able to simply look at these plots to determine whether the theory's marginal effects predictions are supported by evidence with the reassurance that this procedure will yield a false positive empirical confirmation of the predictions at most $100 * \alpha$ percent of the time when all marginal effects are zero.

The specific step-by-step details of our procedure are described in an appendix; however, we provide R code to implement this procedure for generalized linear models as a part of our `interactionTest` R library. This R library leverages the `boot` package (Canty and Ripley,

tistical significance of marginal effects (Berry, DeMeritt and Esarey, 2010). In this case, the bootstrapping procedure described in the next subsection can be adapted to limit the FWER to 5%.

2016) to perform ordinary bootstrap resampling of the target data set. The bootstrapping process can be computationally intensive and lengthy; to speed up performance, the `boot` package can interface with the `snow` library (Tierney et al., 2015) to use parallel processing with multiple CPU cores for faster computation. Our library documentation provides an example of using parallel processing through `snow`.

We tested the effectiveness of the nonparametric bootstrapping procedure in 1,000 simulated data sets with $N = 1000$ observations when all marginal effects are zero 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 of the corresponding hypothesis test.[18] 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 nearly *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: specify theories with multiple predictions in advance and use bootstrapped critical t statistics to maximize empirical power

Correcting for the overconfidence of conventional pointwise 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

[18] The specific null of each test varies according to the specifics of the prediction being tested; see the notes in Table 5 for details.

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	1.08	1.12	1.05	1.22	0.57
e.g.: ME_x^z st. insig. $\mid z < 0.5$, $ME_x^z < 0 \mid z \geq 0.5$	rejection rate	0.04	0.04	0.05	0.04	0.05
opposite-sign directional predictions	median critical t	1.35	1.33	1.24	0.77	0.15
$ME_x^z > 0 \mid z < 0.5$, $ME_x^z < 0 \mid z \geq 0.5$	rejection rate	0.04	0.05	0.06	0.05	0.05
opposite-sign directional predictions for both ME_x^z and ME_z^x	median critical t	1.24	1.23	1.14	0.66	0.10
e,g,: $ME_x^z > 0 \mid z < 0.5$, $ME_x^z < 0 \mid z \geq 0.5$, $ME_z^x > 0 \mid x < 0.5$, $ME_z^x < 0 \mid x \geq 0.5$	rejection rate	0.04	0.05	0.05	0.04	0.05
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	1.28	1.20	0.72	0.13
	rejection rate	0.04	0.05	0.05	0.05	0.04

*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 of the test rejected. The null hypothesis for each test is: (1) $ME_x^{z<0.5} \neq 0 \vee ME_x^{z\geq 0.5} \geq 0$; (2) $ME_x^{z<0.5} \leq 0 \vee ME_x^{z\geq 0.5} \geq 0$; (3) $ME_x^{z<0.5} \leq 0 \vee ME_x^{z\geq 0.5} \geq 0 \vee ME_z^{x<0.5} \leq 0 \vee ME_z^{x\geq 0.5} \geq 0$; (4) $ME_x^{z<0.5} \leq 0 \vee ME_x^{z\geq 0.5} \geq 0 \vee ME_z^x \geq 0$. The critical t row indicates the median nonparametrically bootstrapped t-statistic found to yield a 5% statistical significance rate for the predictions assessed; we use 10,000 bootstrap replicates for each simulated data set. 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; thus $\beta_x = \beta_z = \beta_{xz} = 0$. 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. Rejection rates calculated for $\beta_x = \beta_z = \beta_{xz} = 0$ correspond to sup Pr(false positive|null) for all but the first null hypothesis $\left(ME_x^{z<0.5} \neq 0 \vee ME_x^{z\geq 0.5} \geq 0\right)$; this choice is discussed further in footnote 14.

tradeoff is fundamental to all hypothesis tests and not specific to the analysis of interaction: lowering the size of the test, as we do by setting the FDR or FWER to 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.

This suggestion is subject to two important caveats. First, researchers must use bootstrapped-derived critical t statistics (as in Table 3) in order to reap the benefit of a more powerful test; simply testing each prediction separately using pointwise confidence intervals (as suggested by Berry, Golder and Milton (2012)) would result in diminished power as a result of using overly conservative tests (as shown in the previous section of this paper). Second, 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 when all marginal effects are zero, 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 for a marginal effect in the linear model with interaction terms, Clark and Golder (2006) went on to apply this advice in a *Comparative Political Studies* paper examining 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 (a 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).

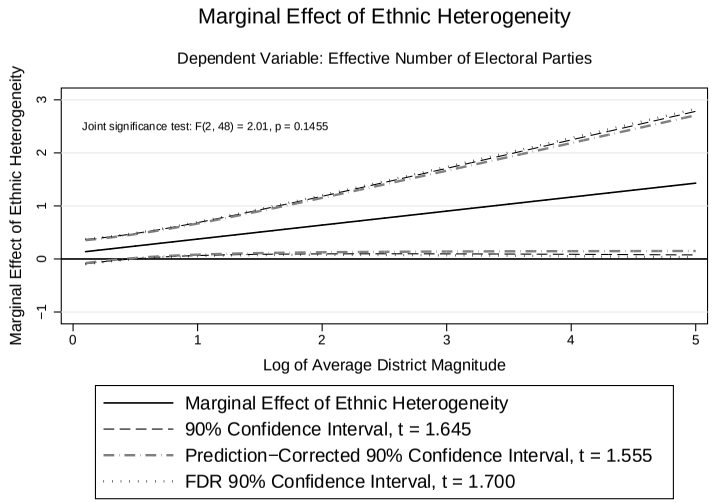
We interpret their hypothesis to mean that the marginal effect of ethnic heterogeneity on the number of electoral parties should be positive when district magnitude is large, and statistically insignificant when district magnitude is small. To test for the presence of this relationship, the authors construct 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 replication 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 $ME_x^{z\geq 2.5} > 0$ (where x is ethnic heterogeneity and z is log average district maginitude), 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).

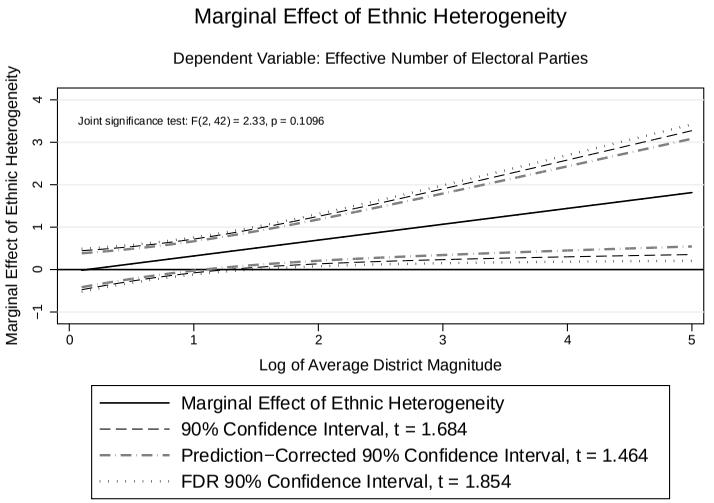
None of the joint F-tests for the statistical significance of the marginal effect of ethnic heterogeneity yield one-tailed p-values less than 0.1. Additionally, FDR-controlling 90% confidence intervals include zero across the entire range of district magnitude for the sample of established democracies 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.

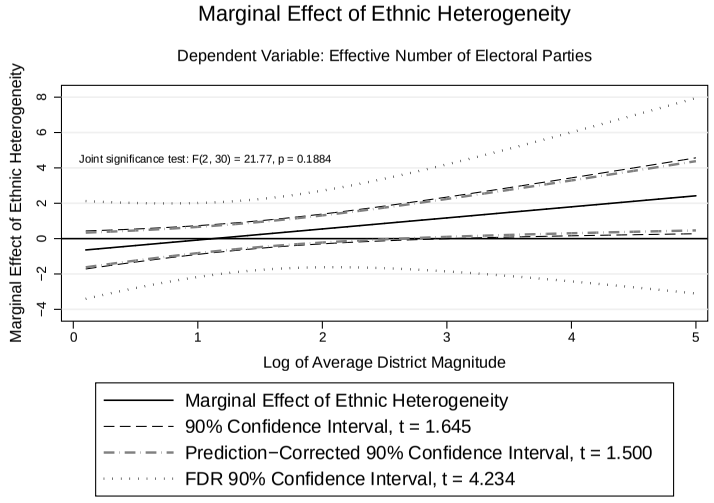
Figure 2: Marginal effect of ethnic heterogeneity on effective number of electoral parties (Figure 1 of Clark and Golder (2006)), with original and prediction- and discovery-corrected confidence intervals



(a) Pooled Analysis, Established Democracies



(b) 1980s, Amorim Neto and Cox



(c) 1990s, Established Democracies

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 when none really exists. 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 be considerably higher *or* lower than researchers believe when they conduct their tests. A further consequence is that researchers using the Brambor, Clark, and Golder 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 interactive 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;

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.

Our new `interactionTest` software package for R makes it easy for applied researchers to control the false positive rate when they create marginal effects plots in the mode of Brambor, Clark and Golder (2006), even in the complex case where multiple theoretical predictions present a threat of underconfident statistical hypothesis tests.

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 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, Chunrong and Edward C. Norton. 2003. "Interaction terms in logit and probit models." *Economics Letters* 80(1):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(1):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)* 57(1):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, William R. Clark and Matthew Golder. 2006. "Understanding interaction models: Improving empirical analyses." *Political Analysis* 14(1):63–82.

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

Canty, Angelo and Brian Ripley. 2016. "boot: Bootstrap R (S-Plus) Functions." R Package. version 1.3-18.

Clark, William R. 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(2):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.

Kutner, Michael, Christopher Nachtsheim, John Neter and William Li. 2004. *Applied linear statistical models*. Fourth ed. New York: McGraw Hill.

Lehmann, E.L. 1957*a*. "A theory of some multiple decision problems, I." *The Annals of Mathematical Statistics* 28(1):1–25.

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

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. URL: `https://www.R-project.org/`.

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.

Schultz, Kenneth A. 1999. "Do democratic institutions constrain or inform? Contrasting two institutional perspectives on democracy and war." *International Organization* 53(2):233–266.

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(318):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.

Tierney, Luke, A. J. Rossini, Na Li and H. Sevcikova. 2015. *snow: Simple Network of Workstations*. R package version 0.4-1. URL: `https://CRAN.R-project.org/package=snow`.

Appendix: Bootstrapping Procedure for Controlling False Positive Rates in Conjoint Theoretical Prediction Tests

The specific procedure that our `interactionTest` library uses to calculate a critical t statistic using nonparametric bootstrapping is as follows:

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 $\widetilde{ME}_x^{z_0}$, $\widetilde{ME}_z^{x_0}$, and their standard errors ($se(\widetilde{ME}_x^{z_0})$ and $se(\widetilde{ME}_z^{x_0})$) using the model for multiple values of x_0 and z_0 in the range of x and z respectively; the standard errors can be analytically derived and calculated using each model estimate. (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 when the marginal effect equals zero.)

5. Repeat steps 2-4 many times; we use 10,000 bootstrap replicates.

6. Using the bootstrapped values of t_x and t_z, find a critical t statistic $t^\star$ such that all theoretical predictions are confirmed α proportion of the time when all marginal effects equal zero. 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 $\widehat{ME}_x^{z_0} \pm t^\star * \text{se}\left(\widehat{ME}_x^{z_0}\right)$.

The confidence intervals for marginal effects plots constructed using this $t^\star$ will yield a false positive empirical confirmation of all the tested predictions at most $100 * \alpha$ percent of the time when all marginal effects are zero.