AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

Bias and Overconfidence in Parametric Models of Interactive Processes*

William D. Berry Florida State University
Jacqueline H.R. DeMeritt University of North Texas
Justin Esarey Rice University

We assess the ability of logit, probit and numerous other parametric models to test a hypothesis that two variables interact in influencing the probability that some event will occur [Pr(Y)] in what we believe is a very common situation: when one's theory is insufficiently strong to dictate a specific functional form for the data generating process. Using Monte Carlo analysis, we find that many models yield overconfident inferences by generating 95% confidence intervals for estimates of the strength of interaction that are far too narrow, but that some logit and probit models produce approximately accurate intervals. Yet all models we study generate point estimates for the strength of interaction with large enough average error to often distort substantive conclusions. We propose an approach to make the most effective use of logit and probit in the situation of specification uncertainty, but argue that nonparametric models may ultimately prove to be superior.

Political science journals are replete with articles presenting hypotheses about the effects of variables on the probability that some event will occur, Pr(Y). In some cases, the theory underlying these hypotheses implies that the data generating process (DGP) for Pr(Y) matches a logit or probit model, so that an unbounded latent dependent variable, Y^*, drives some observed behavior. For example, a rational choice theory of voting implies that utility from voting (Y^*) governs the decision about whether to turn out in an election. Berry, DeMeritt and Esarey (2010, 261) advise that because a logit model specifies the effects of variables on Y^*, rather than on Pr(Y), the proper specification of a logit model designed to test hypotheses about the effects of variables on Pr(Y) must "be based [instead] on an explicit theory about the effects of variables on . . . Y^*." This is particularly important when a researcher's hypotheses include the expectation that the effect of one variable on Pr(Y) varies with the value of another variable, i.e., that two variables interact in influencing Pr(Y).

Although this advice is sound, it is not helpful in what we think is the most common situation confronting a political scientist testing hypotheses about effects of variables on the probability of an event: when her theory makes no reference to an unbounded latent dependent variable. Rather, her attention is restricted to the dependent variable of her hypotheses—Pr(Y), which is bounded between 0 and 1—and its observed indicator, a binary variable. Moreover, in the typical study using binary logit or probit, the theory introduced is not sufficiently specific to imply that logit, probit, or any other functional form is a good fit to the hypothesized DGP. Instead, logit or probit is chosen from among the countless possible functional forms for a model simply because logit and probit have come to be viewed as "default" estimators for a binary dependent variable (BDV) model—making them convenient estimation choices.[1]

William D. Berry is Marian D. Irish Professor, and Syde P. Deeb Eminent Scholar in Political Science, Department of Political Science, Florida State University, Tallahassee, FL 32036-2230 (wberry@fsu.edu). Jacqueline H.R. DeMeritt is Associate Professor, Department of Political Science, University of North Texas, Denton, TX 76203-5017 (jdemeritt@unt.edu). Justin Esarey is Assistant Professor, Department of Political Science, Rice University, Houston, TX 77251-1892 (justin@justinesarey.com).

*We are grateful to Michael Miller for sharing data; and to Miller, Holger Kern, and Carlisle Rainey for their valuable comments. Replication data and documentation are available at http://dx.doi.org/10.7910/DVN/AIRCBB, and supporting information is posted on the *AJPS* website.

[1]For example, the 2005 issues of three major journals—*American Journal of Political Science* (*AJPS*), *American Political Science Review* (*APSR*), and *The Journal of Politics* (*JOP*)—contain 49 articles studying binary dependent variables with logit or probit; 30 of these articles offer no defense for their choice of logit or probit; ten other papers defend their choice, but solely by noting that the dependent variable is dichotomous.

American Journal of Political Science, Vol. 60, No. 2, April 2016, Pp. 521–539

 DOI: 10.1111/ajps.12123

This paper uses Monte Carlo analysis to assess the ability of logit, probit, and generalized linear models (GLMs) relying on other link functions—and containing various combinations of product and quadratic (i.e., squared) terms—to test hypotheses about interaction between variables in influencing the probability of an event in the face of uncertainty about the functional form of the DGP.[2] We find that for the purpose of testing a hypothesis that two variables, X and Z, interact in influencing $\Pr(Y)$, a logit or probit model containing terms for X, Z and their product, XZ, performs as well as any of numerous other GLMs, including models containing a larger number of product and quadratic terms.

Although we can find no GLM that outperforms logit or probit, the performance of logit and probit when there is uncertainty about the functional form of the DGP leaves much to be desired. When political scientists posit interaction between two variables, X and Z, in influencing the probability of an event, and they use logit or probit (including a product term, XZ, in the model) for empirical analysis, a frequently-used test for the presence of interaction is whether the coefficient for the product term is statistically significant. On a positive note, our evidence suggests that when there is no interaction between X and Z in the DGP, the coefficient for the product term, XZ, is statistically significant at the 0.05 level about 5% of the time, precisely as one would expect if the statistical test were accurate.[3] Yet, discouragingly, we find that the coefficient for the product term often fails to be statistically significant when there *is* interaction in the DGP, even when this interaction is very strong and the sample is fairly large.

Furthermore, we find that in the face of uncertainty about the exact functional form of the DGP, logit and probit tend to generate inaccurate point estimates of the quantities typically used by political scientists to measure the strength of interaction. The consequence is that logit and probit often yield point estimates indicating (i) substantively meaningful interaction when none actually exists in the DGP, or (ii) the absence of interaction when there is meaningful interaction in the DGP. Yet, we find that a logit or probit model containing a product term generates 95% confidence intervals for quantities measuring the strength of interaction that are accurate—in the sense that they contain the true quantity about 95% of the time.

With these results in mind, we offer recommendations for the appropriate usage of logit or probit when testing for interaction in influencing the probability of an event in the absence of a strong justification that a logit or probit model accurately specifies the DGP. Since we find that both logit and probit tend to produce inaccurate point estimates of the strength of interaction in this situation of specification uncertainty, but these models seem capable of generating 95% confidence intervals with accurate coverage, we advise political scientists to resist focusing narrowly on point estimates of the strength of interaction—which may be far off the mark. They should focus, instead, on the boundaries established by the accompanying 95% confidence interval.[4] This implies that when a point estimate indicates substantively strong interaction, yet its confidence interval is wide enough so that one of its boundaries implies only weak interaction, one should avoid the temptation to claim evidence of strong interaction, and instead settle for a more cautious interpretation.

Although we recommend an approach to testing interaction hypotheses relying on logit or probit, we view our advice as short-term guidance about how to make the best of a "weak hand." Our Monte Carlo results suggest that political scientists may be wise to phase out the use of logit, probit, and other parametric models to test interaction hypotheses in the presence of uncertainty about the true functional form of the DGP. Less model-dependent nonparametric estimators may be better suited for this situation.

A Common Interaction Hypothesis

Assume that a researcher has a hypothesis that two variables, X and Z, interact in influencing the probability of an event, $\Pr(Y)$, taking the following form:

> **The Binary Dependent Variable (BDV) Interaction Hypothesis**: Two variables, X and Z, interact such that (a) the effect of each of X and Z on $\Pr(Y)$ is *monotonic*—i.e., regardless of the value of X or Z, the marginal effect of X on $\Pr(Y)$ is always positive (or always negative), and the marginal effect of Z on $\Pr(Y)$ is always positive (or always negative).[5] However, (b) at each value

[2] (i) Note that logit and probit are members of a larger set of generalized linear models in the binomial family that differ only in the nonlinear link function mapping the unbounded Y^* into $\Pr(Y)$. (ii) We do not assume complete uncertainty. In particular, we assume a researcher is confident based on theory that the effect of each independent variable on $\Pr(Y)$ is monotonic.

[3] Using statistical terminology, this result implies that the statistical test on the product term has the correct size.

[4] This advice is similar to that offered by Achen (1982) in a different context.

[5] By the *marginal effect* of X on $\Pr(Y)$, we refer to the partial derivative of $\Pr(Y)$ with respect to X, i.e., $\partial \Pr(Y)/\partial X$.

> of X, the marginal effect of X on $\Pr(Y)$ consistently gets stronger (or consistently gets weaker) as Z increases.[6]

This hypothesis is typical of propositions positing interaction between variables in influencing the probability of an event in political science research.[7] A scholar testing this hypothesis is in a situation we label *specification ambiguity*. By this we mean that her hypothesis is sufficiently restrictive to eliminate some functional forms for the DGP—any not consistent with the monotonicity prediction of part *a*—but not sufficiently specific to imply any particular functional form. Yet, it is common practice to test this hypothesis by estimating a logit or probit model including X, Z, and their product, XZ, as independent variables (Nagler, 1991; Ai and Norton 2003; Rainey 2014; but see Berry, DeMeritt and Esarey 2010 and Greene 2010). The logit or probit coefficients are then used to derive one of several quantities of interest characterizing the strength of interaction between X and Z in influencing $\Pr(Y)$.

In the situation of specification ambiguity, a logit or probit model almost certainly misspecifies the underlying DGP; the only question is by how much and with what consequence? Our primary objective is to assess whether logit and probit approximate the DGP well enough to permit these models to adequately test the BDV interaction hypothesis by estimating the strength of interaction to an acceptable degree of accuracy. To do so, we use Monte Carlo methods. We simulate a diverse set of DGPs in which two variables, X and Z, determine $\Pr(Y)$, and generate numerous data sets from each. In each DGP, the effect of each of X and Z on $\Pr(Y)$ is monotonic. We restrict DGPs in this way for two reasons. First, we know—without any need for Monte Carlo analysis—that a logit or probit model in which the only terms involving X and Z are X, Z and XZ is inconsistent with some DGPs in which the effect of X or Z is non-monotonic.[8] Second, as stated in note 7, the vast majority of hypotheses in the literature positing interaction predict that the effects of the interacting variables are monotonic.

Some simulated DGPs are fully consistent with the BDV interaction hypothesis since they also involve interaction between X and Z. Others are inconsistent with the hypothesis, since the marginal effect of X on $\Pr(Y)$ is unrelated to Z, making the effects of X and Z additive. However, importantly, none of the simulated DGPs takes the form of a logit or probit model. Then for each simulated DGP, we assess the ability of logit and probit models containing X, Z and XZ to accurately estimate quantities of interest that political scientists often examine to test the BDV interaction hypothesis. Since this exercise detects substantial weaknesses of logit and probit in estimating the strength of interaction in the situation of specification ambiguity, we also investigate whether other parametric models perform this task better. We consider dozens of specifications—involving seven different link functions from the GLM binomial family, and various combinations of product and quadratic (i.e., squared) terms involving X and Z. We also explore the utility of employing sample fit criteria—like the Akaike Information Criterion (AIC)—to choose among alternative GLM specifications for a specific data set. Unfortunately, none of these alternatives outperforms logit or probit (with X, Z and XZ) for statistical inference about the strength of interaction.

Simulating a Diverse Set of Data Generating Processes (DGPs)

We simulate a diverse set of 115 DGPs in which two variables, X and Z, determine $\Pr(Y)$. In each DGP, each of X and Z is confined to the [0,1] interval, and the DGP associates a unique value of $\Pr(Y)$ with each possible combination of X and Z values. All simulated DGPs share two fundamental characteristics: (i) in each, the effects of X and Z are monotonic; yet (ii) none takes the functional form of a logit model, a probit model, or any other GLM. Except for these elements of commonality, the character of simulated DGPs varies widely.

Most importantly, our simulated DGPs exhibit variation in the strength of interaction between X and Z, as measured by what we call the *min-max second difference*, to be denoted $\Delta\Delta$*min-max*:

$$\Delta\Delta\textit{min-max} = [\Pr(Y|X=1,\,Z=1) - \Pr(Y|X=0,\,Z=1)] - [\Pr(Y|X=1,\,Z=0) - \Pr(Y|X=0,\,Z=0)] \quad (1)$$

[6] Part *b* of the hypothesis is the prediction that X and Z interact; using calculus, the claim is that the second derivative, $\partial^2 \Pr(Y)/\partial X \partial Z$, is always positive (or always negative). Part *a* predicts the sign of the marginal effect of X and of Z across the range of possible values for X and Z, which is critical to clarifying the nature of the expected interaction (Berry, Golder and Milton 2012).

[7] To illustrate, in the 2005 issues of *AJPS*, *APSR* and *JOP*, there are twelve articles that predict interaction between two variables in influencing the probability of an event, and ten offer a proposition in the form of the BDV interaction hypothesis, predicting that the effects of the two variables are monotonic.

[8] For example, if in the DGP, there are values of Z at which the marginal effect of X on $\Pr(Y)$ is positive for values of X below some threshold and negative above, a logit model containing only the terms X, Z and XZ could never reveal empirical evidence of the changing sign of the effect of X.

The third line in equation (1) is a *first difference* in Pr(Y); it reflects the effect of X on Pr(Y) when $Z = 0$ by measuring the response of Pr(Y) to a change across the total range of X when $Z = 0$. The second line contains a similar first difference reflecting the effect of X when $Z = 1$. The min-max second difference is the difference between the two "first differences." It measures how the response of Pr(Y) to a shift in X across its total range—which one might conceive as X's maximum effect on Pr(Y)—changes as one moves across the total range of Z.[9]

We simulate 49 additive DGPs, i.e., DGPs with no interaction between X and Z, and thus for which the min-max second difference ($\Delta\Delta$*min-max*) is zero.[10] The remaining DGPs exhibit interaction at one of five different strengths. Even the weakest strength we simulate—for which $\Delta\Delta$*min-max* $= \pm 0.1$—is intended to reflect nontrivial, and substantively meaningful, interaction. The strongest interaction we simulate is in DGPs for which $\Delta\Delta$*min-max* $= \pm 0.5$. We also simulate DGPs for which $\Delta\Delta$*min-max* $= \pm 0.2$, ± 0.3, or ± 0.4.[11] To ensure diversity among both our additive and interactive DGPs, some are linear and others are non-linear; and among the non-linear ones, we vary the degree of non-linearity.[12]

For illustrative purposes, eight DGPs are presented in Figure 1; our on-line appendix shows a graph (in Figure S-1) and an equation (in Document S-2) for each of the 115 DGPs. Each panel in Figure 1 depicts a DGP by showing the relationship between X and Pr(Y) at three values of Z: its minimum (0), maximum (1), and midpoint (0.5). Counterclockwise starting at the upper left, we exhibit linear additive DGPs (in panel A), non-linear additive DGPs (panel C), non-linear interactive DGPs (panel D), and linear interactive DGPs (panel B). As an example, consider the DGP in the left part of panel D, for which the min-max second difference is -0.2. To confirm this value for $\Delta\Delta$*min-max*, we note that when $Z = 0$ (the lowest curve), a shift across the range of X from 0 to 1 causes Pr(Y) to decrease from 0.5 to 0.3, for a change of -0.2. In contrast, when $Z = 1$ (the highest curve), Pr(Y) decreases by 0.4 (from 0.9 to 0.5) as X increases from 0 to 1. Thus, when Z moves across its range from 0 to 1, the maximum effect of X on Pr(Y) changes from -0.2 to -0.4, yielding a min-max second difference of -0.2 [$= -0.4 - (-0.2)$].[13]

Using Simulated DGPs to Construct Data Sets for Analysis

To conduct Monte Carlo analysis, we need to generate multiple data sets from each simulated DGP. We construct each observation in a data set by drawing a value for X and a value for Z from a uniform distribution over the [0,1] interval, and computing Pr(Y) for the observation using the DGP [which maps each combination of X and Z values into Pr(Y)]. We then determine whether $Y = 0$ or $Y = 1$ for the observation using a procedure that assigns Y a value of 1 with probability Pr(Y).[14] Most of our analysis of the performance of estimation models is conducted on 25 data sets of 1000 observations for each DGP (for a total of $115 \times 25 = 2875$ data sets).

The Estimation Models We Study

For each data set, we assess the performance of logit and probit models containing X, Z and their product, XZ, in testing the BDV interaction hypothesis. We also seek to determine whether other parametric models can outperform logit and probit in this task. There is certainly reason to speculate that we may be able to increase performance with an alternative model. Even if we limit attention to generalized linear models (GLMs) in the binomial family, logit and probit are just two among numerous models that are consistent with the BDV interaction hypothesis.

[9] A second difference is a quantity often reported by scholars to characterize the strength of interaction (e.g., Haspel and Knotts 2005; Basinger and Lavine 2005).

[10] Note, however, that a min-max second difference of zero does not imply the absence of interaction. Indeed, even when the min-max second difference is zero, the marginal effect of X on Pr(Y) can be different when $Z = 0$ than when $Z = 1$ at every possible value of X (for an example, see Figure S-7 in our on-line appendix). The min-max second difference is best conceived as a measure of the "global" extent of interaction.

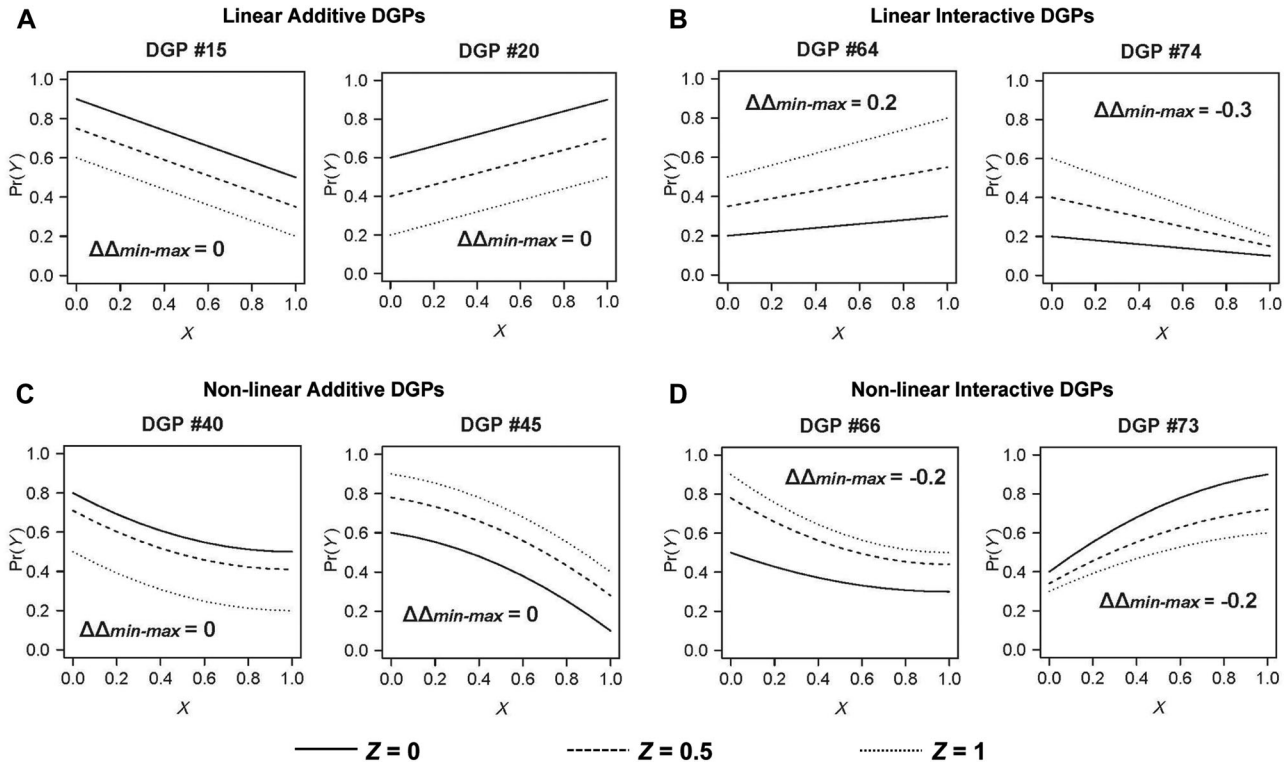
[11] The distribution of our interactive DGPs across different strengths of interaction is as follows: 12 each for which $|\Delta\Delta$*min-max*$| = 0.1, 0.2, 0.4$, or 0.5; and 18 DGPs for which $|\Delta\Delta$*min-max*$| = 0.3$.

[12] Operationally, each non-linear DGP is established by specifying a relationship between X and Pr(Y) when $Z = 0$ in the form of a quadratic function [i.e., with Pr(Y) as a function of X and X^2]. Similarly, the relationship between X and Pr(Y) when $Z = 1$ is established as a different quadratic function. Then it is assumed that as Z increases from 0 to 1, the relationship between X and Pr(Y) changes gradually—at a constant rate in some DGPs, at a variable rate in others—between these two quadratic functions.

[13] Document S-1 in our on-line appendix describes our procedure for establishing a diverse set of DGPs, and Table S-1 in this appendix lists key characteristics of each DGP, thereby documenting this diversity.

[14] Specifically, we set $Y = 1$ if Pr(Y) $\leq u$ where $u \sim U[0,1]$, and $Y = 0$ otherwise.

FIGURE 1 Illustrative Simulated DGPs



Note: The DGP number listed in each graph refers to a reference number that is assigned to the DGP in Figure S-1 of our on-line appendix.

TABLE 1 The 42 Generalized Linear Models (GLMs) We Study

GLM Link Function		Included Terms		
1:	regression (identity)	A:	basic	X, Z
2:	logit	B:	product-term	X, Z, XZ
3:	probit	C:	X^2	X, Z, XZ, X^2, ZX^2
4:	log	D:	Z^2	X, Z, XZ, Z^2, XZ^2
5:	log-complement (log c)	E:	X^2 and Z^2	X, Z, XZ, X^2, ZX^2, Z^2, XZ^2
6:	log-log	F:	saturated	X, Z, XZ, X^2, ZX^2, Z^2, XZ^2, X^2Z^2
7:	log-log complement (cloglog)			

Note: To obtain 42 models, we use each possible combination of the seven link functions and six sets of included terms listed. The alphanumeric codes in this table are used to identify link functions and sets of included terms in later text, figures and tables.

In all, we consider the performance of the seven GLM link functions listed in the left column of Table 1.[15]

When testing the BDV interaction hypothesis, the estimation model must approximate a DGP with an unknown functional form. Including additional product terms involving X and Z would seem to increase a model's potential to approximate DGPs having a variety of functional forms. Accordingly, we assess the performance of models including one of four combinations of product terms involving quadratic components (i.e., X^2 and/or Z^2), as listed in the right column of Table 1. The specification with the most terms, labeled "saturated," contains both X^2 and Z^2, and multiple product terms involving these variables. For completeness, we also evaluate the performance of "basic" models containing only X and Z, with no product terms. Altogether, we examine models

[15]These consist of all binomial family link functions available in Stata 11.2's glm procedure except for *power* and *odds power*, each of which requires the user to specify a parameter that can assume an infinite number of values.

involving six different sets of included terms. The combination of seven link functions with six sets of terms yields 42 models for investigation, which we identify by the numerical code for the link function and the letter code for the set of included terms in Table 1.[16] For example, the logit product-term model—containing the terms *X*, *Z* and *XZ*—is denoted estimation model 2B.

It also seems reasonable that in the situation of specification ambiguity, we might be able to improve performance in testing the BDV interaction hypothesis by relying on empirical information in a sample to choose the functional form for the estimation model. Accordingly, we assess the value of two sample fit criteria: the Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC). For each simulated data set, we estimate all 42 models listed in Table 1, determine which of these candidate models has the best fit (based on the AIC or the BIC), and then estimate a quantity measuring the strength of interaction using the coefficients from the best-fitting model. We then repeat this process, except that we limit the candidate models to the 14 models in Table 1 with either no product term (i.e., specification A) or a single product term (specification B). Also, given the frequent reliance in the literature on logit and probit models with a single product term for testing the BDV interaction hypothesis, for both the logit and probit link functions separately, we use the AIC and the BIC to choose between including a product term (specification B) and excluding one (specification A). Finally, for each of logit and probit, we assess the performance of deciding whether to include a product term based on whether the coefficient for the product term, *XZ*, in the sample is statistically significant (at the 0.05 level, two-tailed test). We refer to any procedure relying on sample information to choose the best link function and set of included terms as an *adaptive* model. The ten adaptive models we examine are summarized in Table 2, which assigns each model a numerical reference code at the extreme left of the table.

Two Measures of the Strength of Interaction

Researchers testing the BDV interaction hypothesis typically examine one of two model products: a *second difference estimate* [as recommended by King, Tomz and Wittenberg (2000) and Zelner (2009); but see Hanmer and Kalkan (2013)], or a *marginal effect plot* [as recommended by Brambor, Clark and Golder (2006), and Berry, Golder and Milton (2012); see also Norton, Wang and Ai (2004)]. Each model product characterizes the strength of interaction between *X* and *Z* in influencing Pr(*Y*) by showing how the estimated effect of *X* on Pr(*Y*) changes as *Z* changes. Consequently, we examine the ability of alternative estimation models to accurately recover each of two quantities reflecting the true strength of interaction in the DGP. One is the *min-max second difference*, denoted $\Delta\Delta$*min-max* and previously defined in equation (1). The other quantity we call the *min-max marginal effect difference*, which we denote by Δ*MEmin-max*. The min-max marginal effect difference for *X* measures the response of the marginal effect of *X* on Pr(*Y*) to an increase in *Z* across its range in the sample, when *X* is fixed at its mean:

$$\begin{aligned}\Delta ME\,min\text{-}max\,(X) \;=\; & [ME\,(X)\,|X = 0.5,\, Z = 1] \\ & -[ME\,(X)\,|X = 0.5,\, Z = 0],\end{aligned} \tag{2}$$

where *ME*(*X*) denotes the marginal effect of *X* on Pr(*Y*).[17] We also compute the min-max marginal effect difference for *Z* by exchanging *X* and *Z* in equation (2).

Criteria for Evaluating the Performance of an Estimation Model

As we note above, our goal is to determine how well various estimation models recover the true value of each of two quantities measuring the strength of interaction in a DGP. We assess a model's success in estimating a quantity in two respects: (i) the accuracy of its point estimates, and (ii) the accuracy of its confidence intervals. Both kinds of accuracy are measured in a straightforward fashion. For any estimation model and any DGP, to assess the accuracy of the model's point estimates of a quantity, we determine the *mean absolute error* in the point estimate across 25 data sets generated by the DGP, i.e., the average absolute difference between the estimate of the quantity and its known true value for the DGP. A lower mean absolute error indicates a more accurate estimate, with a value of zero implying that the estimated value is equal to the true value in each of the 25 data sets. To measure the accuracy of an estimation model's confidence intervals for a quantity, we

[16] All models are estimated in Stata 11.2 using the glm, probit, or logit commands.

[17] Note that if one were to construct a marginal effect plot showing the relationship between $ME(X)|X = 0.5$ (on the vertical axis) and *Z* (on the horizontal axis) over the range of values for *Z*, Δ*MEmin-max*(*X*) would be the vertical difference between the endpoints of the plotted curve. The difference in meaning between the two measures of the strength of interaction—Δ*MEmin-max* and $\Delta\Delta$*min-max*—is illustrated in Figure S-7 in our on-line appendix.

TABLE 2 The Ten Adaptive Models We Study

Label for Adaptive Model		Choice Criterion	Candidate Link Functions (from Table 1)	Candidate Included Terms (from Table 1)	Number of Candidate GLMs from Table 1 Considered
8:	AIC	AIC	all seven	all six	42 (all)
9:	AIC simple	AIC	all seven	A, B	14
10:	logit AIC	AIC	logit	A, B	2
11:	probit AIC	AIC	probit	A, B	2
12:	BIC	BIC	all seven	all six	42 (all)
13:	BIC simple	BIC	all seven	A, B	14
14:	logit BIC	BIC	logit	A, B	2
15:	probit BIC	BIC	probit	A, B	2
16:	logit significance	stat. signif. of *XZ*	logit	A, B	2
17:	probit significance	stat. signif. of *XZ*	probit	A, B	2

calculate the percentage of the 25 data sets generated by a DGP for which the 95% confidence interval for the quantity contains its true value. We deem a model to be yielding "accurate" 95% confidence intervals if the intervals it produces contain the true value about 95% of the time.[18]

The Performance of Parametric Models in Testing the BDV Interaction Hypothesis

In this section, we use Monte Carlo analysis to examine the ability of the generalized linear models listed in Table 1 and the adaptive models listed in Table 2 to test the BDV interaction hypothesis by recovering the true min-max second difference ($\Delta\Delta$*min-max*) of a DGP in the situation of specification ambiguity. Moreover, to see whether a model's performance in estimating the min-max second difference varies with the true strength of interaction in the DGP, we consider separately (i) additive DGPs (i.e., ones where $\Delta\Delta$*min-max* $= 0$); (ii) DGPs with "weak" interaction, where $\Delta\Delta$*min-max* $= \pm 0.1, \pm 0.2$, or ± 0.3; and (iii) DGPs with "strong" interaction, where $\Delta\Delta$*min-max* $= \pm 0.4$ or ± 0.5.[19] We begin by assessing the ability of the various models to produce accurate 95% confidence intervals. Then, for each model that does so, we assess the accuracy of its point estimates.

The Accuracy of Confidence Intervals: Overconfidence is a Common Problem

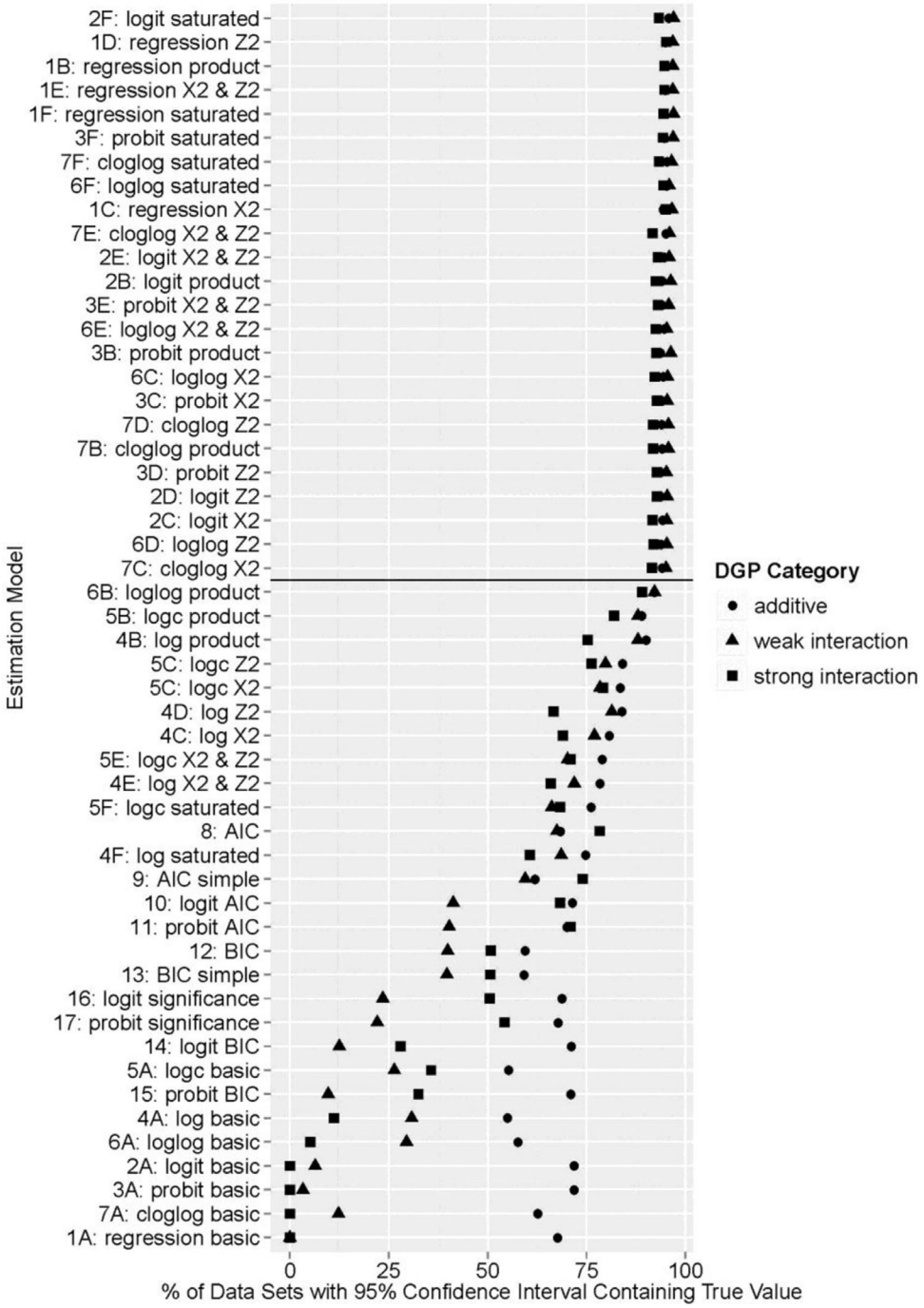
For each of the three types of DGPs—additive, weakly interactive, and strongly interactive—and each estimation model, we calculate the percentage of data sets ($n = 1000$) in which the model produces a 95% confidence interval for the min-max second difference that contains the DGP's true min-max second difference. We deem a model as yielding "accurate" 95% confidence intervals if the intervals it produces contain the true value between 90% and 100% of the time in each of the three DGP categories. The results are shown in Figure 2. Each model that produces accurate confidence intervals appears above the horizontal line superimposed on the figure.

Three kinds of estimation models tend to produce substantially overconfident inferences by generating 95% confidence intervals that are too narrow, and consequently contain the true value of $\Delta\Delta$*min-max* far less than 95% of the time. This set of poorly performing models includes all adaptive models—which are based on choosing among multiple models using a sample fit criterion, and identified in the left margin of Figure 2 by a number between 8 and 17. Importantly, this includes adaptive models that choose whether to include a product term, *XZ*, in a model based on the statistical significance

[18]It is important to recognize that finding that a model yields "accurate" 95% confidence intervals does not imply that the estimate of the standard error of the sampling distribution on which this confidence interval is based is accurate. Thus, a model yielding accurate 95% confidence intervals does not necessarily yield accurate intervals at other levels of confidence (e.g., 90% or 50%).

[19]Note that the terms "weak" and "strong" are convenient labels when collapsing the five categories of interaction strength into two groups to simplify presentation of the results. However, our sense is that in most applications, a min-max second difference of ± 0.3 would actually indicate strong interaction, and even a $\Delta\Delta$*min-max* value of ± 0.1 would represent substantively meaningful interaction.

FIGURE 2 Accuracy of the 95% Confidence Intervals for a Min-Max Second Difference ($\Delta\Delta$*min-max*) Produced by 52 Estimation Models



Note: The horizontal axis indicates the percentage of simulated data sets for which the 95% confidence interval for the min-max second difference ($\Delta\Delta$*min-max*) produced by an estimation model listed in the left margin—and identified by the alphanumeric codes established in Tables 1 and 2—contains the true value of $\Delta\Delta$*min-max*. (Note that in the left margin, "X^2" and "Z^2" are denoted as "X2" and "Z2," respectively, to keep the figure compact.) The estimation models are listed in order of the percentage of *all* simulated data sets for which the confidence interval contains the true value, but confidence interval accuracy is reported separately for each of three types of DGPs defined by the strength of interaction in the DGP: additive (i.e., $\Delta\Delta$*min-max* $= 0$); "weak" interaction, where $\Delta\Delta$*min-max* $= \pm 0.1, \pm 0.2,$ or ± 0.3; and "strong" interaction, where $\Delta\Delta$*min-max* $= \pm 0.4$ or ± 0.5; according to the legend in the right margin.

of that term.[20] All models without a product term (denoted as "basic" models in Figure 2) also yield confidence intervals that are too narrow. This is true even among DGPs that are additive, and thus involve no interaction. Finally, models relying on the log or log-complement link functions produce inaccurate 95% confidence intervals regardless of the variables included in the model.

In contrast, the vast majority of estimation models that include at least one product term produce 95% confidence intervals for the min-max second difference with accurate boundaries in all three DGP categories. Indeed, there is very little difference in performance among GLMs with a wide range of link functions and included product/quadratic terms. Altogether, 24 of the 52 models in Figure 2—the first 24 listed—produce accurate confidence intervals. This includes the logit and probit product-term models (with only *X*, *Z*, and *XZ* included).[21]

The Accuracy of Point Estimates: Small Differences Among Many Poorly Performing Models

Our success in identifying numerous models that produce accurate 95% confidence intervals for the min-max second difference ($\Delta\Delta$*min-max*) in the situation of specification ambiguity is important, because it guarantees that there are estimation models that should allow researchers doing empirical analysis to determine a range within which the true min-max second difference is very likely to lie. Indeed, given that there are many models that yield 95% confidence intervals with accurate boundaries, we see no reason to consider further models that fail to do so. The remaining question is whether any of the models that yield accurate confidence intervals also produce accurate point estimates. To answer this question, for each DGP and each estimation model among those producing accurate 95% confidence intervals, we compute the *average absolute error* in the point estimate of $\Delta\Delta$*min-max* over the 25 data sets generated from the DGP. Figure 3 portrays for each estimation model the distribution of average absolute error values across DGPs within each of the same three DGP categories used in Figure 2.

Figure 3 shows that there is little difference among the alternative estimation models in their ability to recover a DGP's true min-max second difference. In each of the 72 box plots portrayed—characterizing all combinations of 24 estimation models and the three DGP categories—the median value (across DGPs) of average absolute error is in the narrow range between 0.115 and 0.145, and each 25th and 75th percentile value is between 0.105 and 0.165.[22] Thus, there is no evidence that one can improve appreciably on the performance of a logit or probit model with a single product term by (i) relying on a GLM with an alternative link function, or (ii) adding product terms involving quadratic components.

Although the implications of error of the magnitude depicted in Figure 3 in any research project must be evaluated based on the specific nature of the political process being studied, we think that such error has the potential to produce misleading conclusions about the strength of interaction in many studies. We believe that in most research contexts, a min-max second difference of 0.1 indicates substantively meaningful interaction between two variables in influencing the probability of an event. If so, the magnitude of estimation error we detect is sufficient to often make (i) an additive DGP (i.e., one with a true min-max second difference of zero) appear interactive, and (ii) a DGP with a true min-max second difference in the range between 0.10 and 0.15—and thus characterized by substantively meaningful interaction—seem additive.

A Different Quantity for Measuring the Strength of Interaction, But the Same Conclusion

What happens if we assess the strength of interaction by estimating the min-max marginal effect difference (Δ*MEmin-max*) rather than the min-max second difference? In short, the results are largely the same. Figure 4 replicates the analysis of Figure 3 for each of 25 estimation models that produce accurate 95% confidence intervals (by the same criterion used above) for the min-max marginal effect difference[23]; these models include

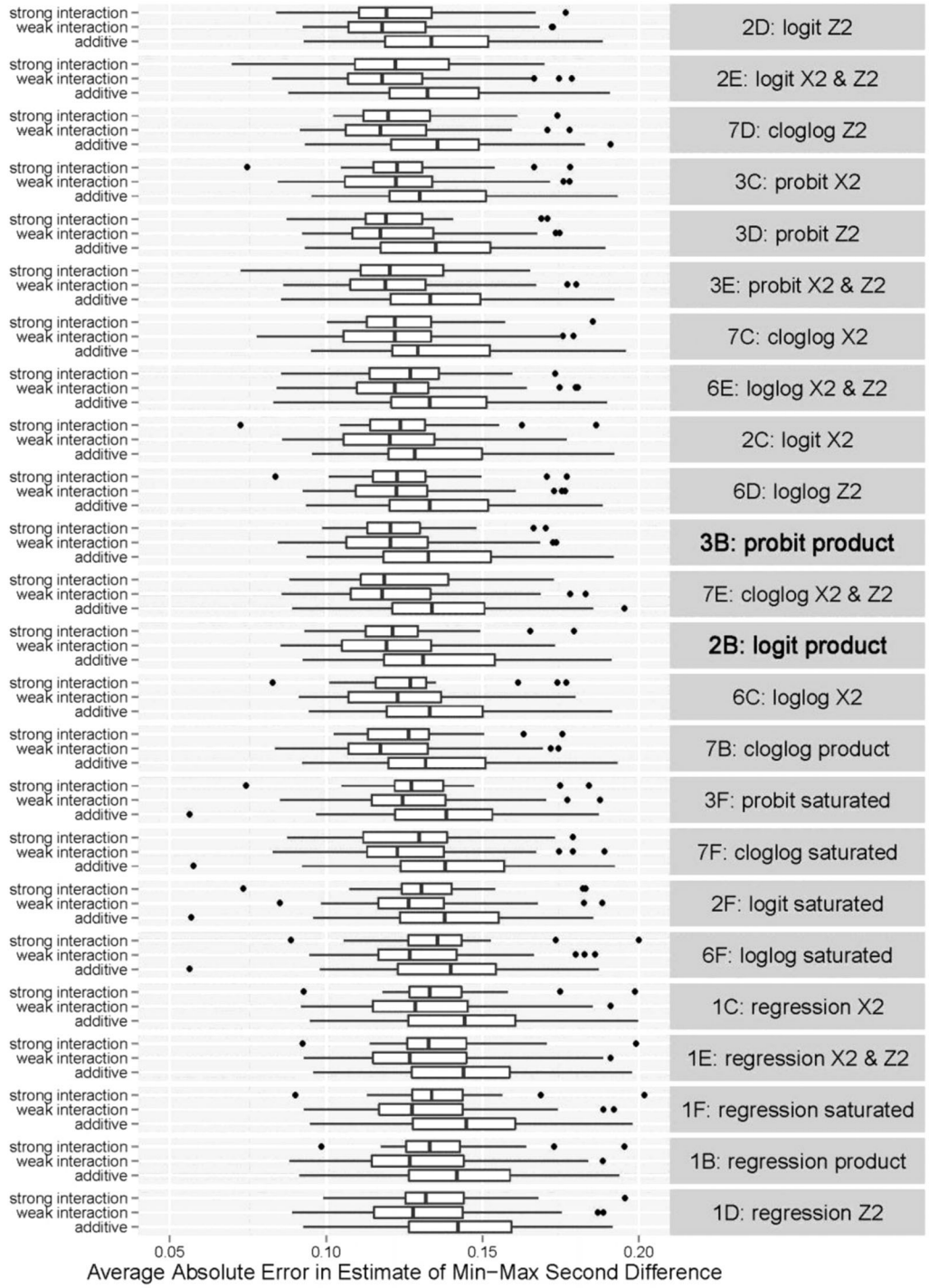
[20]Note that we have not yet discussed whether the statistical significance of the product term, *XZ*, in a model containing a single product term is an accurate test for the presence of interaction against the null hypothesis that the DGP is additive; we consider this below.

[21]To ensure that our results are not an artifact of our choice of sample size for simulated data sets ($n = 1000$), we use the glm function in R to redo our simulation analysis for the probit product-term model (3B in Table 1) with much smaller data sets: $n = 100$. We find that 95% confidence intervals contain the true value of $\Delta\Delta$*min-max* 94.2% of the time for additive DGPs, 93.8% of the time for DGPs with weak interaction, and 92.5% of the time for DGPs with strong interaction.

[22]The only evidence of systematic variation in performance of the estimation models is across DGP categories, with each model performing somewhat worse for data sets generated by strongly interactive or additive DGPs than for data sets generated by weakly interactive DGPs.

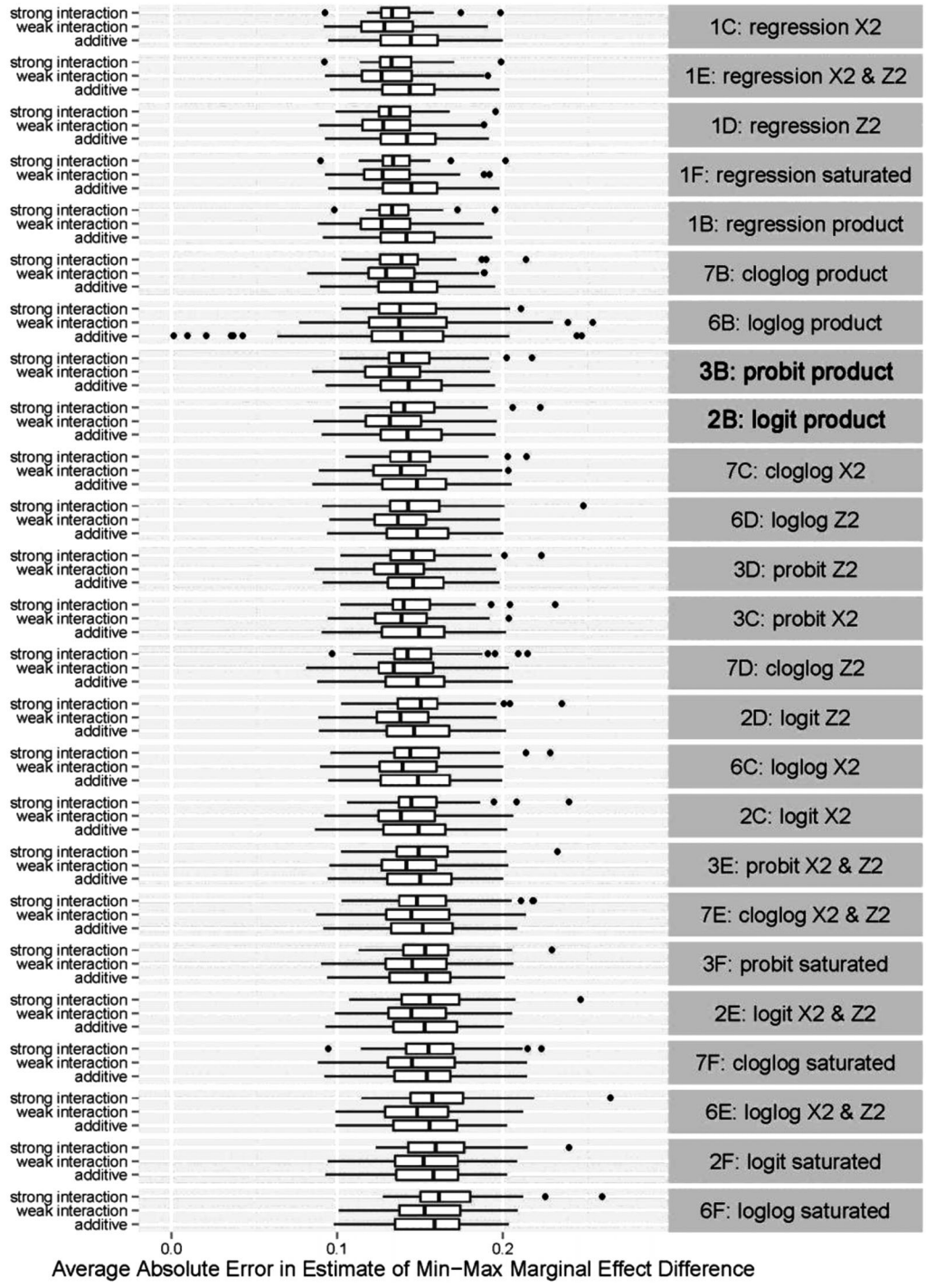
[23]The 25 models include the 24 that produce accurate confidence intervals for the min-max second difference—providing evidence

FIGURE 3 Accuracy of the Point Estimates of the Min-Max Second Difference ($\Delta\Delta$*min-max*) by 24 Estimation Models that Produce Accurate 95% Confidence Intervals



Note: Each box plot shows the distribution across all DGPs in a specified category of the average absolute error in an estimate of a DGP's min-max second difference produced by the estimation model listed in the right margin (and identified by the alphanumeric codes established in Table 1). (Note that in the right margin, "X^2" and "Z^2" are denoted as "X2" and "Z2," respectively.) The estimation models are listed in order of the overall accuracy of their point estimates across *all* data sets analyzed. However, accuracy is reported separately for each of the three DGP categories defined by the strength of interaction in the DGP: additive (i.e., $\Delta\Delta$*min-max* $= 0$); "weak" interaction, where $\Delta\Delta$*min-max* $= \pm 0.1, \pm 0.2$, or ± 0.3; and "strong" interaction, where $\Delta\Delta$*min-max* $= \pm 0.4$ or ± 0.5. DGP categories are listed in the left margin.

FIGURE 4 Accuracy of the Point Estimates of the Min-Max Marginal Effect Difference [$\Delta MEmin\text{-}max(X)$ or $\Delta MEmin\text{-}max(Z)$] by 25 Estimation Models that Produce Accurate 95% Confidence Intervals



Note: Each box plot shows the distribution across all DGPs in a specified category of the average absolute error in an estimate of a DGP's min-max marginal effect difference (for X or for Z) produced by the estimation model listed in the right margin (and identified by the alphanumeric codes established in Table 1). (Note that in the right margin, "X^2" and "Z^2" are denoted as "X2" and "Z2," respectively.) The estimation models are listed in order of the overall accuracy of their point estimates across *all* data sets analyzed. However, accuracy is reported separately for each of the three DGP categories defined by the strength of interaction in the DGP: additive (i.e., $\Delta\Delta min\text{-}max = 0$); "weak" interaction, where $\Delta\Delta min\text{-}max = \pm 0.1, \pm 0.2$, or ± 0.3; and "strong" interaction, where $\Delta\Delta min\text{-}max = \pm 0.4$ or ± 0.5. DGP categories are listed in the left margin.

the logit and probit models with a single product term.[24] Just as before, there is little variation among models that yield accurate confidence intervals in the accuracy of their point estimates. Overall, point estimates of the min-max marginal effect difference have slightly more error than point estimates of the min-max second difference; in each of the 75 box plots in Figure 4, the median value (across DGPs) of average absolute error is in the narrow range between 0.125 and 0.165, and each 25th and 75th percentile value is between 0.110 and 0.180.

To summarize the key results thus far, our Monte Carlo analysis shows that the logit and probit product-term models are capable of producing 95% confidence intervals for the strength of interaction with accurate boundaries, whether measured by the min-max second difference ($\Delta\Delta$*min-max*) or the min-max marginal effect difference (Δ*MEmin-max*).[25] Although logit and probit generate accurate confidence intervals, they tend to yield point estimates of the strength of interaction with enough error to create a substantial risk that a researcher will draw a misleading substantive conclusion—claiming substantively meaningful interaction when there is no interaction in the DGP, or concluding that interaction is absent when it is actually present.[26] Nevertheless, there is no evidence that one can reduce the error in point estimates by abandoning a logit or probit model with a single product term in favor of a GLM with a different link function or that includes additional product terms with quadratic components.

The Ability of Probit and Logit to Produce Accurate Statistical Tests for the Presence of Interaction

It is common practice by political scientists to treat a statistically significant coefficient for a product term, *XZ*, in a logit or probit model as a necessary condition for concluding that there is interaction between *X* and *Z* (Nagler 1991; Rainey 2014). Berry, DeMeritt and Esarey (2010) show that this practice is misguided when they demonstrate that a statistically significant product term in a logit model is neither necessary nor sufficient for substantively meaningful interaction between two variables in influencing Pr(*Y*). However, their argument assumes that logit accurately specifies the theory being tested, and thus their reasoning is irrelevant to the context we examine: the situation of specification ambiguity. In this situation, does a test of statistical significance for the product-term coefficient in a logit or probit model yield reliable information about whether *X* and *Z* interact in influencing Pr(*Y*)? Similarly, in the situation of specification ambiguity, does a test of the statistical significance of the min-max second difference or the min-max marginal effect difference produce reliable information about whether X and *Z* interact?

Figure 2 already provides a partial answer to the second question. It shows that for additive DGPs (i.e., ones with no interaction), the logit and probit product-term models produce 95% confidence intervals for the min-max second difference that contain the true min-max second difference—i.e., zero—about 95% of the time. This result establishes that when the DGP is additive, a two-tailed test for statistical significance (at the 0.05 level) of the min-max second difference against the null hypothesis of no interaction yields a statistically significant quantity about 5% of the time, just as one would

of the robustness of our results across alternative measures of the strength of interaction—plus the log-log product-term model (6B from Table 1).

[24] As above (using the glm function in R), we conduct supplemental simulations evaluating the performance of the probit product-term model in estimating 95% confidence intervals for Δ*MEmin-max* in data sets with fewer cases: $n = 100$. 95% confidence intervals contain the true value of Δ*MEmin-max* 94.2% of the time for additive DGPs, 94.7% of the time for DGPs with weak interaction, and 94.1% of the time for DGPs with strong interaction.

[25] It is possible that of the numerous estimation models yielding accurate 95% confidence intervals, there is substantial variation in the average width of these intervals. If this were the case, a model producing *narrower* accurate confidence intervals should be considered superior. Indeed, for each quantity measuring the strength of interaction, there is Monte Carlo evidence of nontrivial variation in the average width of the intervals produced by models yielding accurate 95% confidence intervals. However, for each quantity, the logit and probit product-term models are among the set of estimation models yielding intervals with the smallest average width (see Figures S-10 and S-11 in our on-line appendix). [Figure S-11 shows that the log-log product-term model (6B from Table 1) yields somewhat narrower confidence intervals for the min-max marginal effect difference than the logit and probit product-term models, but Figure 4 shows that this model's point estimates have average absolute error levels with much greater variability across additive DGPs than do the logit and probit product-term models—making logit and probit preferable estimators.]

[26] (i) In this sentence, and others below, we make a claim about the magnitude of error in a point estimate. Such claims *always* refer to the *average* magnitude of error over repeated estimations, and should never be interpreted as asserting that the described amount of error will be present in any estimate from a single sample. (ii) Monte Carlo analysis (see Figures S-8 and S-9 in our on-line appendix) shows that there is no consistent pattern about the direction of error in point estimates of the two quantities measuring the strength of interaction—whether this error is in the direction of (a) overestimating the strength of interaction, or (b) underestimating it. (Note that bias toward *under*estimation, when larger in magnitude than the true value of the quantity, yields an expected estimated value with a sign opposite that of the true value.) In applied practice, the direction of error will depend on unknowable aspects of the DGP being studied.

expect with a valid test for statistical significance.[27] We now provide more extensive results regarding the performance of tests for statistical significance of the min-max second difference, and evaluate the performance of tests for statistical significance of the min-max marginal effect difference and the product-term coefficient too.

We estimate a probit product-term model in each of 100 data sets ($n = 1000$) generated from each of our 115 simulated DGPs, and use the coefficients to estimate the min-max second difference in each data set.[28] We then calculate for each DGP (i) the percentage of data sets in which the min-max second difference is statistically significant at the 0.05 level, and (ii) the percentage of data sets in which the product term in the model, *XZ*, is statistically significant. Figure 5 depicts the distribution of each of these two percentages across the 49 DGPs that are additive. Panel A shows that for additive DGPs, the median percentage of data sets in which the min-max second difference is statistically significant is 6, which is nearly equal to the percentage we would expect if the statistical test were perfectly accurate (i.e., 5). Moreover, there is not a single additive DGP among those we simulate for which this percentage exceeds 12. Similarly, panel B shows that the probit product-term model only rarely yields a statistically significant product term when there is no interaction in the DGP.[29]

Figure 6 shifts attention to DGPs in which there is interaction. In it, we plot the percentage of data sets in which the product term (panel A) or the min-max second difference (panel B) is statistically significant against the true strength of interaction in the DGP. It is evident that neither statistical test does a good job of detecting interaction in our simulated data sets ($n = 1000$). Indeed, even when the DGP is very strongly interactive—with a min-max second difference of ± 0.5—the estimated min-max second difference and the product term each fails to be statistically significant more than 35% of the time. The presence of weaker interaction in a DGP ($|\Delta\Delta \textit{min-max}| \leq 0.40$) is detected by a statistical test of one of the two quantities less than half the time.[30]

Similar Findings When One of the Interacting Variables is Dichotomous

Many tests of the BDV interaction hypothesis in the literature involve a binary independent variable. To ensure that our Monte Carlo results are not contingent on the independent variables being continuous, we repeat our analysis making one of the variables, *Z*, binary. Detailed results are presented in Document S-3 of our on-line appendix, but broadly speaking, the results are similar to those reported above for the continuous case. To summarize the two most important findings:

- The logit and probit product-term models are among a set of generalized linear models (GLMs) that produce approximately accurate 95% confidence intervals for both quantities measuring the strength of interaction: the min-max second difference, and the min-max marginal effect difference for *X*.[31]
- All GLMs that produce 95% confidence intervals with accurate boundaries for the two quantities measuring the strength of interaction have roughly equal, but disappointing, levels of point estimate accuracy.

The Accuracy of Logit and Probit Estimates of a Central Marginal Effect

Our simulation results reveal that in the situation of specification ambiguity, logit and probit have systematic difficulty yielding accurate point estimates of quantities measuring the strength of interaction. However, estimating the strength of interaction in the absence of knowledge about the exact functional form of the underlying DGP is a very difficult task,[32] and many scholars use logit or

[27] Thus, there is a very low probability of falsely rejecting the null hypothesis of no interaction (i.e., committing a type I error).

[28] Models are estimated using the glm function in R.

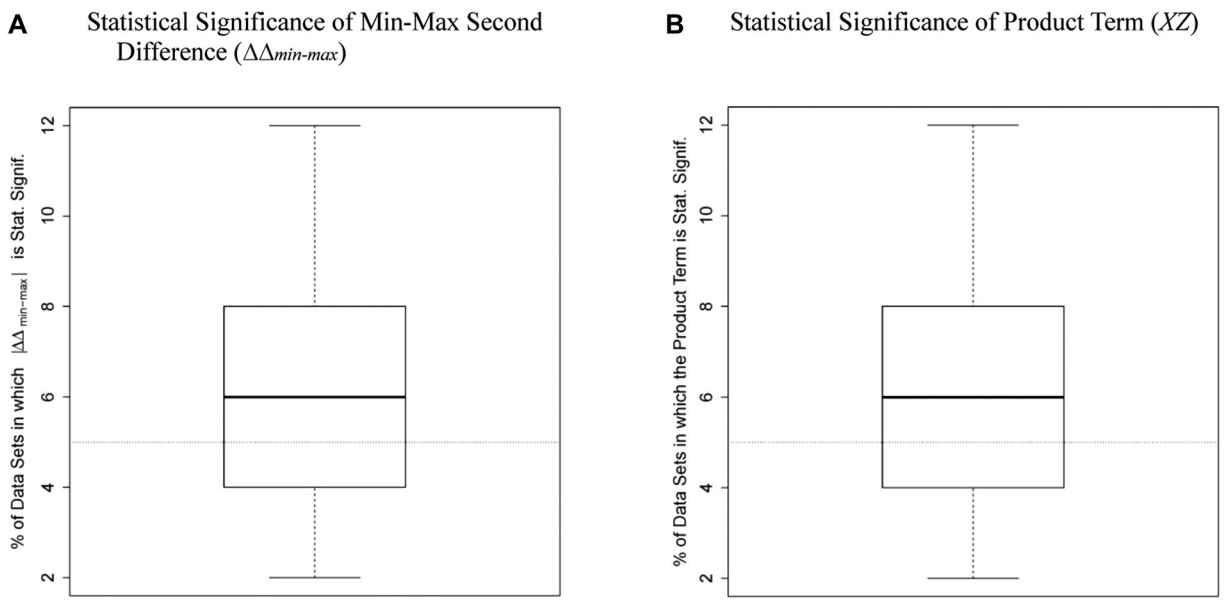
[29] A figure comparable to Figure 5, but for a sample size of either 100 or 500, would convey similar conclusions.

[30] (i) The performance of these statistical tests declines dramatically for data sets of 100 cases. For example, even when the true value of $|\Delta\Delta \textit{min-max}|$ is 0.5, an estimate of the min-max second difference (or of the product term) is statistically significant less than 13% of the time.(ii) When the analyses reported in Figures 5(A) and 6(B) are repeated for our other measure of the strength of interaction, the min-max marginal effect difference, the results—which are portrayed in Figure S-12 of our on-line appendix—are very similar. The same is true when the analyses in each panel of Figures 5, 6 and S-12 are replicated using logit rather than probit; detailed results are, respectively, in Figures S-13, S-14 and S-15 of our on-line appendix. Consequently, the conclusions reported in this section apply to both logit and probit, and to both measures of the strength of interaction.

[31] This claim is restricted to *X* because one cannot conceive of the marginal effect of *Z* when *Z* is binary.The marginal effect of a variable, *V*, is the response of the dependent variable to an instantaneous change in *V*, and a binary variable is not subject to an instantaneous change. Put differently, $\partial \Pr(Y)/\partial V$ is undefined when *V* is binary.

[32] The difficulty can be attributed to the "curse of dimensionality" (King and Zeng 2001).

FIGURE 5 The Ability of the Probit Product-Term Model to Fail to Reject the Null Hypothesis of Additivity When the DGP is Additive



Note: Each box plot shows the distribution across our 49 additive DGPs—for which $\Delta\Delta min\text{-}max = 0$—of the percentage of data sets drawn from a DGP in which the estimate of the min-max second difference (panel A) or the product term (panel B) is statistically significant at the 0.05 level (two-tailed test). The dotted horizontal line marks the conventional 5% threshold for statistical significance. The fact that the two box plots are identical is not entirely surprising since each plotted value is a percentage of 100 data sets, and hence constrained to be an integer.

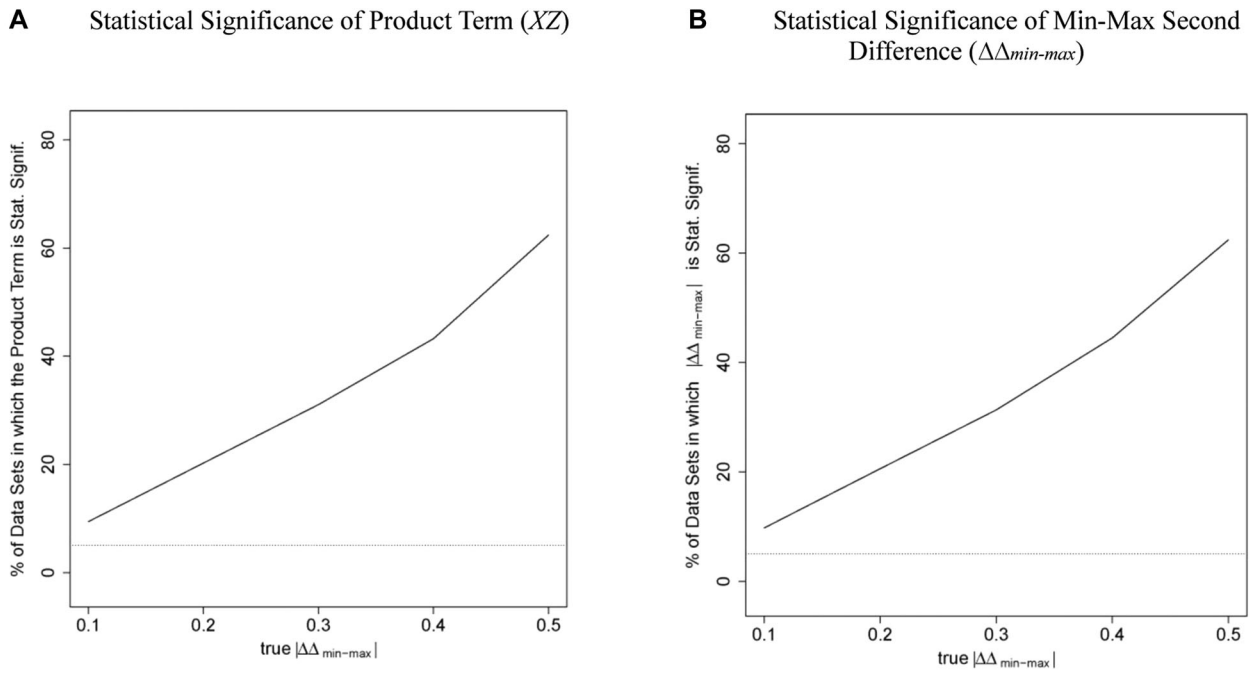
probit for what we would expect to be a far simpler task: estimating a *central marginal effect* (e.g., Holian 2009; Gilbert and Oladi 2012; Dreher and Gassebner 2012). By the central marginal effect of X on $\Pr(Y)$, we refer to the marginal effect of X when each independent variable is held constant at a central value (mean, median or mode). This quantity is often estimated to test an unconditional hypothesis taking the form, "X has a positive (or negative) effect on $\Pr(Y)$." It is important to know whether logit's poor performance in estimating the strength of interaction in the situation of specification ambiguity extends even to this easier task. We therefore ask, "When a researcher's theory is sufficiently strong to justify an assumption that the effects of independent variables are monotonic, but insufficiently strong to assume that any particular parametric model accurately specifies the DGP, does a logit or probit model yield an accurate estimate of the central marginal effect of a variable on the probability of an event?"

To answer this question, we apply the same methodology we use above to assess the performance of GLMs in estimating quantities measuring the strength of interaction. However, this time we assess the performance of each of four models—logit with and without a product term (models 2B and 2A in Table 1), and probit with and without a product term (3B and 3A)—in estimating the central marginal effect of each of X and Z for each of our 115 simulated DGPs. As earlier, we consider the accuracy of both point estimates and confidence intervals, and assess performance separately for DGPs with different strengths of interaction.[33]

The combination of four estimation models (2A, 3A, 2B, or 3B) and six categories of DGPs ($|\Delta\Delta min\text{-}max| = 0, 0.1, 0.2, 0.3, 0.4$, or 0.5) yields 24 assessment contexts. In 23 of the 24 contexts, 95% confidence intervals for the estimated central marginal effect of X or Z contain the true central marginal effect between 90% and 100% of the time. In the remaining context, the percentage exceeds 89. Thus, there is evidence that both logit and probit produce approximately accurate 95% confidence intervals in the situation of specification ambiguity. Furthermore, our analysis shows that the quality of point estimates of the central marginal effect is very stable across DGPs with different strengths of interaction; in each of the 24 contexts, the average absolute error in an estimate of the central marginal effect falls within the narrow range between 0.044 and 0.054. We believe that

[33] Results are summarized below; complete results can be found in Table S-4 of our on-line appendix.

FIGURE 6 The Ability of the Probit Product-Term Model to Reject the Null Hypothesis of Additivity When the DGP is Interactive



Note: Each point plotted is the average percentage of time that the product term (panel A) or an estimate of the min-max second difference (panel B) is statistically significant at the 0.05 level (two-tailed test) in data sets drawn from a DGP for which the magnitude of the true min-max second difference, $|\Delta\Delta min\text{-}max|$, is the value indicated on the horizontal axis. The dotted horizontal line marks the conventional 5% threshold for statistical significance.

in many applications by political scientists, estimation error of such magnitude is unlikely to distort conclusions about the substantive significance of a central marginal effect.

The Accuracy of Logit and Probit Estimates of Non-Central Marginal Effects

The relatively strong performance of logit and probit in estimating a central marginal effect in the situation of specification ambiguity does not ensure a similar level of performance in estimating a marginal effect farther from the "center" of the data. Indeed, King and Zeng (2006) find that model dependence rises with the distance from the center, suggesting that logit/probit performance may also decline with distance.

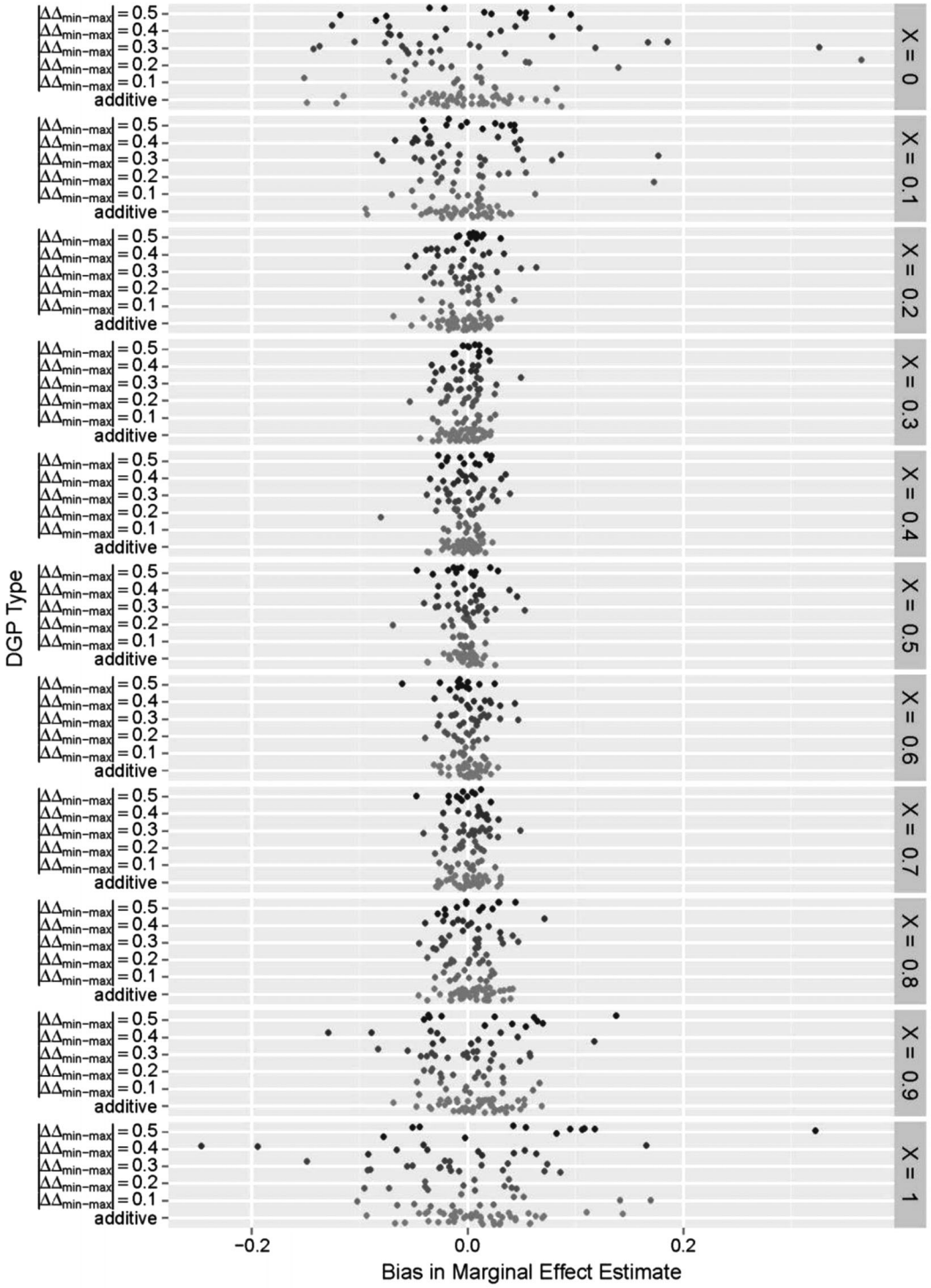
To assess whether this speculation is correct, for each of our simulated DGPs, we use logit to estimate the marginal effect of X on $\Pr(Y)$ at various combinations of X and Z values. Estimated values are generated from a "saturated" logit model containing multiple product terms involving quadratic components (model 2F in Table 1), using an extremely large data set ($n = 100000$) drawn from the DGP.[34] For each DGP, we compute the *bias* in the estimate of the marginal effect of X—i.e., the estimated effect minus the true effect—at each of eleven values of X spread across the range from 0 to 1, when Z is held at its central value of 0.5.[35] In Figure 7, each of these bias values is denoted by a point, and points are plotted separately for DGPs with different strengths of interaction (as measured by the min-max second difference).

If our speculation is correct, the magnitude of bias should be greater when X or Z takes on an extreme value of 0 or 1 than when each of X and Z is near its central value of 0.5. Consistent with our expectation, Figure 7

[34]We use a very large data set for this illustration to eliminate the possibility that weak performance in estimation is due to sampling error. We rely on a saturated model to give logit the best possible shot to closely approximate the DGP. Recall that in our Monte Carlo analysis, a saturated logit model performs about as well as a logit or probit model with a single product term (see Figures 2, 3 and 4).

[35]Note that the *sign* of a bias value does not reveal whether the strength of the marginal effect is over- or underestimated. This is because for some DGPs (e.g., #64 in Figure 1), positive bias would imply overestimation of strength, while for others (e.g., #73 in Figure 1), negative bias would imply overestimation of strength.

FIGURE 7 The Declining Performance of Logit in Estimating the Marginal Effect of *X* When Moving Away From the "Center" of the Data



Note: The graph shows the bias in an estimate of the marginal effect of X on Pr(Y) (i.e., the estimated effect minus the true effect in the DGP) derived from a saturated logit model (Model 2F in Table 1) fitted on an asymptotically large data set ($n = 100{,}000$). Bias values are denoted on the horizontal axis. Each data point represents the bias in the estimated marginal effect of X for a DGP with the strength of interaction—as measured by the absolute min-max second difference ($|\Delta\Delta min\text{-}max|$)—specified in the left margin, at the value of X indicated in the right margin, when Z equals its central value of 0.5. The plotted points are "jittered" to avoid overlapping, and get darker as the strength of interaction in the DGP increases.

indicates that regardless of the strength of interaction in the DGP, the magnitude of bias in an estimated marginal effect is relatively low when both X and Z are near the center of the data. Indeed, with Z fixed at 0.5, at any value of X within 0.2 of 0.5 (i.e., when X is 0.3, 0.4, 0.5, 0.6 or 0.7), the magnitude of bias in the estimated marginal effect of X exceeds 0.05 for at most two of the 115 DGPs. But when X is 0.1, 0.2, 0.8 or 0.9 (i.e., close to the "edge" of the data), the marginal effect of X is estimated with bias of substantially greater magnitude for many DGPs.

Recommendations and Final Thoughts

Political scientists testing the hypothesis that two variables, X and Z, interact in influencing the probability of an event, $\Pr(Y)$, are typically in what we label the "situation of specification ambiguity." By this we mean that they are willing to assume, based on a priori theory, that the effects of X and Z on $\Pr(Y)$ are monotonic—i.e., the marginal effect of each of X and Z always has the same sign—but their theory is not strong enough to justify an assumption that logit, probit, or any other model, matches the functional form of the data generating process (DGP). Yet, researchers in this situation nearly always estimate a logit or probit model including X, Z and the product term, XZ; and use the resulting coefficients to estimate the strength of interaction between X and Z in influencing $\Pr(Y)$.

Our Monte Carlo results validate this practice in some respects, but they also raise concerns about some types of inferences researchers routinely draw from their logit or probit analysis. First, our simulations indicate that in the situation of specification ambiguity, a logit or probit model including X, Z and XZ performs at least as well in estimating the strength of interaction as any model in a large set of generalized linear models (GLMs) involving (i) a variety of link functions, and (ii) one or more product terms with quadratic components (e.g., XZ^2). Second, a logit or probit model without a product term tends to yield overly optimistic confidence intervals for the strength of interaction. Third, using a fit statistic—the AIC, the BIC, or whether a product term is statistically significant—to determine which among various GLM specifications to use to estimate the strength of interaction in a particular data set is problematic; this strategy, too, yields excessively optimistic confidence intervals. These three findings prompt the following advice:

> **Recommendation 1** (about **Model Specification**): When estimating a quantity measuring the strength of interaction between X and Z in influencing $\Pr(Y)$—the min-max second difference ($\Delta\Delta$*min-max*), or the min-max marginal effect difference (Δ*MEmin-max*)—to test the BDV interaction hypothesis (see p. 4 for a specific statement of the hypothesis), researchers should employ a probit or logit model with a single product term: XZ.[36]

Our Monte Carlo analysis suggests that some common statistical tests yield some trustworthy results even in the situation of specification ambiguity. Specifically, when no interaction is present in a DGP (i.e., the DGP is additive), a test for the statistical significance (at the 0.05 level, two tailed) of either the min-max second difference or the min-max marginal effect difference derived from a probit or logit model with a single product term yields the conclusion that there is statistically significant interaction only about 5% of the time—the same frequency of a "false positive" signal that would be expected with any sound test of statistical significance. The same is true of a test for the statistical significance of the coefficient for the product term. Yet, these tests are not very powerful; they often yield a "false negative" result by failing to detect statistically significant interaction when interaction—even very strong interaction—is present in the DGP. Thus:

> **Recommendation 2** (about the **Accuracy of Statistical Tests**): Even in the situation of specification ambiguity, a researcher can reasonably use a finding derived from a logit or probit model with a product term that an estimate of the min-max second difference ($\Delta\Delta$*min-max*) or the min-max marginal effect difference (Δ*MEmin-max*) is statistically significant at the 0.05 level to justify an inference that some interaction between X and Z is very likely present. The same is true of a finding that the product term is statistically significant. However, finding that any of these three quantities is not statistically significant should not be treated as evidence of the absence of interaction.

We find, however, that in the situation of specification ambiguity, logit or probit point estimates of quantities measuring the strength of interaction—either $\Delta\Delta$*min-max* or Δ*MEmin-max*—are characterized by error of a magnitude that can be sufficient in many

[36] This contrasts with the situation in which one's interactive theory gives one a high degree of confidence that a logit or probit model accurately specifies the DGP, where the decision about whether to include a product term should be driven by that theory (Berry, DeMeritt and Esarey 2010).

contexts to distort a conclusion about the substantive significance of interaction. Yet our analysis suggests that 95% confidence intervals for either quantity are accurate, in the sense that the true value of the quantity lies inside the estimated interval about 95% of the time. This prompts the following advice:

> **Recommendation 3** (about **Deemphasizing Point Estimates**): In the situation of specification ambiguity, an assessment of the substantive significance of an estimate of the min-max second difference ($\Delta\Delta$*min-max*) or the min-max marginal effect difference (Δ*MEmin-max*) derived from a logit or probit model with a product term should be based on the boundaries of the confidence interval for the quantity rather than on its point estimate. In particular, researchers should claim that interaction is substantively significant only if both boundaries of the interval have a magnitude strong enough to justify the claim.

Our Monte Carlo analysis suggests that specification ambiguity imposes less pernicious consequences when logit or probit is used to estimate the *central marginal effect* of a variable on Pr(Y)—i.e., the marginal effect of the variable when each independent variable is held at its mean—than when one of these models is used to estimate the strength of interaction. Indeed, our results suggest that as long as one's theory is strong enough to justify an assumption that the effect of a variable on Pr(Y) is monotonic, (i) a 95% confidence interval for the central marginal effect of the variable is likely to have accurate boundaries, and (ii) the expected error in the point estimate of the central marginal effect should be small enough to avoid a deceptive conclusion about the substantive importance of the effect in many research contexts. This leads us to a final suggestion:

> **Recommendation 4** (about the **Accuracy of Central Marginal Effect Estimates**): In the situation of specification ambiguity, a researcher using logit or probit has a stronger justification for giving credence to a point estimate of the central marginal effect of a variable than for giving credence to a point estimate of a quantity measuring the strength of interaction. This is not, however, a license to focus exclusively on the point estimate of a central marginal effect and ignore its confidence interval—an action that cannot be justified even in absence of any specification uncertainty (Achen 1982).

We believe that following our recommendations would likely lead political scientists to claim compelling evidence of interaction much less frequently than they do now.[37] For those, like us, who believe that the effects of variables tend to be contextual—and consequently, that our theories should be conditional—this would be a discouraging outcome. Much interaction likely present in the world would be undetected. Stronger theory permitting more confident assumptions about functional form would solve the problem, but we are skeptical that political scientists are capable of developing the kinds of precise theories that would be required without resorting to largely arbitrary assumptions. Thus, we encourage the research community to search for estimation models that perform better than logit and probit in the presence of specification uncertainty.

Researchers have shown that a variety of semiparametric and nonparametric models outperform logit and probit in some settings (e.g., Kennedy 2008; Härdle 2004). These models include generalized additive models (GAMs), and several machine learning algorithms—such as neural network models and random forests. Some of these models may prove to be superior to the generalized linear models (GLMs) we examine in this paper for testing the BDV interaction hypothesis, a possibility that should be investigated using Monte Carlo methods.

References

Ai, Chunrong and Edward C. Norton. 2003. "Interaction terms in logit and probit models." *Economics Letters* 80:123–129.

Basinger, Scott J. and Howard Lavine. 2005. "Ambivalence, Information, and Electorat Choice." *American Political Science Review* 99:169–84.

Berry, William D., Jacqueline H.R. DeMeritt and Justin Esarey. 2010. "Testing for Interaction in Binary Logit and Probit Models: Is a Product Term Essential?" *American Journal of Political Science* 54:248–266.

Berry, William D., Matt Golder and Daniel Milton. 2012. "Improving Tests of Theories Positing Interaction." *The Journal of Politics* 74:653-671.

Dreher, Axel and Martin Gassebner. 2012. "Do IMF and World Bank Programs Induce Government Crises? An Empirical Analysis." *International Organization* 66:329-358.

Gilbert, John and Reza Oladi. 2012. "Net Campaign Contributions, Agricultural Interests, and Votes on Liberalizing Trade with China." *Public Choice* 150:745-769.

[37] In Document S-4 in our on-line appendix, we illustrate the potential implications of following our recommendations by revisiting analysis recently presented by Miller (2012).

Greene, William. 2010. "Testing Hypotheses about Interaction Terms in Nonlinear Models." *Economics Letters* 107: 291–296.

Härdle, Wolfgang, Marlene Müller, Stefan Sperlich and Axel Werwatz. 2004. *Nonparametric and Semiparametric Models.* Berlin: Springer-Verlag.

Haspel, Moshe and H. Gibbs Knotts. 2005. "Location, Location, Location: Precinct Placement and the Costs of Voting." *The Journal of Politics* 67:560–573.

Holian, Matthew J. 2009. "Outsourcing in US Cities, Ambulances and Elderly Voters." *Public Choice* 141:421-445.

Kennedy, Peter. 2008. *A Guide to Econometrics*, 6th ed. Malden, MA: Blackwell.

King, Gary and Langche Zeng. 2001. "Improving Forecasts of State Failure." *World Politics* 53:623-658.

———. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14:131–159.

Miller, Michael K. 2012. "Economic Development, Violent Leader Removal, and Democratization." *American Journal of Political Science* 56:1002–1020.

Nagler, Jonathan. 1991. "The Effect of Registration Laws and Education on US Voter Turnout." *American Political Science Review* 85:1393–1405.

Rainey, Carlisle. 2014. "Compression and Conditional Effects." Working Paper. URL: http://www.carlislerainey.com/research/compression-and-conditional-effects/.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

On-Line Appendix