

## Testing for Interaction Using Binary Logit in the Presence of Specification Ambiguity

*Logit is routinely used to test hypotheses that variables interact in influencing the probability of an event. But logit is a specific functional form, and researchers rarely develop theories precise enough to indicate that logit, or any other functional form, correctly characterizes the hypothesized data generating process. We use Monte Carlo analysis to investigate the consequences of using logit in the absence of a theoretical justification for a logit specification. We report three main findings: (1) The statistical significance of a product term is not a good indicator of whether there is appreciable interaction in influencing event occurrence. (2) Even when there is interaction, a model without a product term sometimes outperforms a model with one in accurately estimating the strength of interaction. (3) Using the Akaike Information Criterion to determine whether to include a product term is preferable to routinely including a product term any time interaction is hypothesized.*

## Introduction

Political science journals are replete with papers hypothesizing that variables interact in influencing the probability of an event,  $\Pr(Y)$ , and testing the hypothesis using binary logit or probit. In a recent issue of this *Journal*, Berry, DeMeritt and Esarey (2010, 248) demonstrate that contrary to conventional wisdom, “a statistically significant product term is neither necessary nor sufficient for variables to interact meaningfully in influencing  $\Pr(Y)$ .” The authors advise that the decision about whether to include a product term in a logit<sup>1</sup> model designed to test a hypothesis positing that variables interact in influencing  $\Pr(Y)$  should “be based on an explicit theory about the effects of variables on the unbounded latent dependent variable... assumed by the model” (p. 261).<sup>2</sup>

Although this advice is sound, it is not helpful in the most common situation confronting a political scientist testing hypotheses about effects of variables on  $\Pr(Y)$ : when her theory makes no reference at all to an unbounded latent dependent variable. Instead, her attention is restricted to the concept,  $\Pr(Y)$ , constrained to the interval (0,1); and its empirical indicator, the binary dependent variable (BDV),  $Y$ .<sup>3</sup> Moreover, in the typical study using binary logit, the theory introduced is insufficiently precise to indicate that logit—or any other specific functional form—is a good fit to the hypothesized data generating process (DGP).<sup>4</sup> This paper investigates the

---

<sup>1</sup> Henceforth, “logit” should be read as a shorthand for “binary logit or probit,” as all points made in this paper about logit apply equally well to the highly similar probit (Aldrich and Nelson 1984, 34).

<sup>2</sup> In logit or probit, an unbounded latent dependent variable,  $Y^*$ , is specified as a function of the independent variables (i.e.,  $Y^* = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$ ). In turn, logit and probit map the unbounded  $Y^*$  into the range (0,1) using a nonlinear link function to yield  $\Pr(Y)$ .

<sup>3</sup> For example, the 2005 issues of *American Journal of Political Science (AJPS)*, *American Political Science Review (APSR)* and *Journal of Politics (JOP)* contain 49 articles studying binary dependent variables with logit or probit, and 92% make no reference to an unbounded latent variable.

<sup>4</sup> 30 of the 49 papers reporting binary logit or probit models in the 2005 issues of *AJPS*, *APSR*

appropriate use of a product term in a logit model in this typical situation of uncertainty about the functional form of the DGP. We find, using Monte Carlo analysis, that (i) the statistical significance of a product term in a logit model is not a reliable indicator of whether there is appreciable interaction in influencing  $\Pr(Y)$ , (ii) even when there is such interaction, a model without a product term sometimes outperforms a model with a product term in accurately estimating the strength of interaction, and (iii) using the Akaike Information Criterion (AIC) to determine whether to include a product term is preferable to routinely including a product term any time interaction is hypothesized.

### **Testing for Interaction in the Face of Theoretical Uncertainty about Functional Form**

When a researcher without a theoretical justification that logit is the correct functional form uses logit for empirical analysis nonetheless, the model will almost certainly misspecify the DGP. In this case, the quality of logit's estimates of the effects of variables on  $\Pr(Y)$  depends on logit's ability to approximate a wide variety of different underlying DGPs.<sup>5</sup> Thus, a key question is whether one can reliably determine the functional form for the variables in a logit model (e.g., whether to include a product term) that maximizes the approximation's accuracy and enables hypothesis testing—all without knowing the true form of the underlying DGP. In this paper we ask, and answer, this question.

In particular, we assess logit's performance in accurately estimating effects on  $\Pr(Y)$  when testing an interaction hypothesis taking the following form:

---

and *JOP* offer no defense for their choice to employ logit or probit; ten others defend their choice, but rely solely on the binary nature of the dependent variable.

<sup>5</sup> Cramer (2003, 16), in fact, argues that the logit model can be conceived of as a first-order Taylor series approximation to a large variety of DGPs.

**BDV Interaction Hypothesis:**  $X$  and  $Z$  interact such that at each value of  $Z$ , the relationship between  $X$  and  $\Pr(Y)$  is monotonic (either positive or negative), but (at each value of  $X$ ) the magnitude of the effect of  $X$  on  $\Pr(Y)$  changes (consistently getting stronger or weaker) as  $Z$  increases.<sup>6</sup>

The common practice in recent years has been to test this hypothesis with a logit model including  $X$ ,  $Z$  and their product,  $XZ$ , as independent variables. If the coefficient estimate for the product term is statistically insignificant, the hypothesis of interaction is rejected. If the product term is statistically significant, the analyst proceeds to a comparison of (i) how the estimated response of  $\Pr(Y)$  to a discrete change in  $X$ —a first difference—varies with  $Z$  (King, Tomz and Wittenberg 2000), or (ii) how the marginal effect of  $X$  on  $\Pr(Y)$  varies with  $Z$  (Brambor, Clark & Golder 2006).

The BDV interaction hypothesis is typical of the propositions positing interaction in recent political science research involving a BDV.<sup>7</sup> Researchers testing this hypothesis are in situation we label *specification ambiguity*. By this we mean that their hypothesis is restrictive enough to eliminate many functional forms for the DGP—any in which the relationship between  $X$  and  $\Pr(Y)$  is *nonmonotonic* at some values of  $Z$ , or in which the strength of the effect of  $X$  on  $\Pr(Y)$  varies *nonmonotonically* with  $Z$  (rising over part of the range of  $Z$ , but decreasing over the other)—yet their hypothesis is not sufficiently specific to imply that logit accurately reflects the DGP, or guide the choice about the functional form with which variables should be included in the model (e.g., whether a product term should be included). Thus, unless a researcher is

---

<sup>6</sup> A monotonic relationship between an independent variable and a dependent variable is one that is either positive throughout or negative throughout, rather than positive over one set of values for the independent variable and negative over another. We also assume that it is implicit in the BDV interaction hypothesis that the DGP is smooth and well-behaved (i.e., there are no abrupt changes in the marginal effects of independent variables).

<sup>7</sup> Indeed, in the 2005 issues of *AJPS*, *APSR* and *JOP*, there are 11 articles that posit interaction between variables in influencing the probability that an event will occur, and nine present a proposition in the form of the BDV interaction hypothesis. The remaining two claim that a variable has a positive effect on  $\Pr(Y)$  in one context and a negative effect in another, inconsistent with the monotonicity assumption.

prepared to advance a theory that yields more specific predictions, the most reasonable stance is to assume that logit misspecifies the DGP, and therefore that logit parameter estimates and predicted probabilities of event occurrence derived from these estimates may be inconsistent. To use logit with any confidence in this situation would require evidence across a wide range of DGPs that the expected estimation error due to misspecification is small, and consequently, that a logit model approximates the DGP closely enough to allow a valid test of one's hypothesis.

For this reason, we use Monte Carlo methods. We simulate a diverse set of monotonic DGPs—none taking the form of a logit model—and assess the ability of logit to determine whether interaction is present in the GDP, and if so, the strength of the interaction. Obviously, we cannot hope to comprehensively explore the entire universe of possible monotonic DGPs, or even to choose a sample of DGPs that can be viewed as representative of some “true” population of DGPs in the world. But this is not our goal. Rather, the point of our simulation is to illustrate that the accepted practice for testing the BDV interaction hypothesis described above can yield misleading conclusions under common conditions, and determine if a superior practice can be designed.

### **Simulating a Diverse Set of “True” Relationships**

We seek to assess whether logit is able to accurately test the BDV interaction hypothesis in the situation we label specification ambiguity—in which the researcher feels confident a priori that the DGP is monotonic, but is uncertain about its specific functional form. Accordingly, we simulate 230 different DGPs that share two fundamental characteristics. First, *each is monotonic*. Second, *none takes the form of a logit model*. Except for these elements of commonality, the character of the simulated DGPs varies widely, so that we have both a diverse set of additive relationships and a diverse set of interactive relationships. This ensures that our

Monte Carlo results do not hinge on the true relationship having some highly specific functional form.<sup>8</sup> Next, we briefly describe our procedure for establishing a diverse set of simulated true relationships; Appendix S-1 in the supporting information provides a more detailed description.

Each simulated DGP specifies the effects of two independent variables— $X$  and  $Z$ , both confined to the  $[0,1]$  interval—on  $\Pr(Y)$ .<sup>9</sup> We define a DGP by associating a unique value of  $\Pr(Y)$  to each possible combination of  $X$  and  $Z$  values. For illustrative purposes, some of the 230 simulated DGPs are presented in Figure 1.<sup>10</sup> Each panel depicts a DGP by showing the relationship between  $X$  and  $\Pr(Y)$  at three values of  $Z$ : its minimum (0), its maximum (1), and its midpoint (0.5).

In most of the 230 DGPs, the relationship between  $X$  and  $\Pr(Y)$  varies with the value of  $Z$ , making the DGP interactive (e.g., panels A and B of Figure 1). However, in some, the effects of  $X$  and  $Z$  on  $\Pr(Y)$  are additive—with the effect of  $X$  having the same magnitude at all values of  $Z$  (e.g., panels C and D). In some DGPs, the effects of  $X$  and  $Z$  on  $\Pr(Y)$  are linear (e.g., panels B and D), but in most, the relationship is nonlinear (e.g., panels A and C).<sup>11</sup> Among the nonlinear DGPs, some are concave up (e.g., the left side of panel A), and others are concave down (e.g.,

---

<sup>8</sup> For analysis justifying that the DGPs we simulate are sufficiently diverse to allow our results to be generalized beyond the specific DGPs we examine, see Table S-1 in the supporting information.

<sup>9</sup> The confinement of  $X$  and  $Z$  to  $[0,1]$  is an arbitrary normalization to simplify the interpretation of simulation results. Equivalently, each variable can be assumed to have any finite range,  $[a,b]$ , corresponding to the variable's minimum and maximum values in the population.

<sup>10</sup> A plot of each of the 230 DGPs and the mathematical equation for each can be found in Figure S-1 and Appendix S-2 in the supporting information. Each DGP has a unique code number that is consistent across all tables and figures in the supporting information.

<sup>11</sup> We make most simulated true relationships nonlinear because it seems likely that in real-world relationships, the marginal effects of variables on  $\Pr(Y)$  will tend to decline in magnitude as  $\Pr(Y)$  approaches the extreme values of 0 and 1. Indeed, if  $X$  and  $Z$  were unbounded, linearity in effects would *necessarily* push  $\Pr(Y)$  outside the  $[0,1]$  interval for observations with some values for  $X$  and  $Z$ . However, since in any real-world DGP the values of the independent variables are constrained to a finite range, a linear form is possible.

the right side of panel A). We also vary the degree of departure from linearity among the nonlinear DGPs by changing the rate at which the marginal effect of each independent variable on  $\Pr(Y)$  varies with the value of that variable.<sup>12</sup> In half the simulated true relationships, both  $X$  and  $Z$  are continuous; in the other half,  $X$  is continuous and  $Z$  is dichotomous (scored either 0 or 1). Each DGP that involves a dichotomous  $Z$  is identical to one of the continuous- $Z$  true relationships except that  $Z$  is constrained to be 0 or 1 rather than free to vary between these extremes.<sup>13</sup>

### **Varying the Strength of Interaction in Simulated True Relationships**

While no single value can fully describe the magnitude of the interaction in a nonlinear DGP, we can characterize one important aspect of the strength of interaction using the *min-max second difference* (to be denoted  $\Delta\Delta_{\min-\max}$ ). The min-max second difference is defined as the difference between (i) the first difference in  $\Pr(Y)$  as  $X$  increases from its minimum to its maximum *when  $Z$  is at its maximum* (the value 1) and (ii) the first difference in  $\Pr(Y)$  as  $X$  increases from its minimum to its maximum *when  $Z$  is at its minimum* (the value 0):

$$\Delta\Delta_{\min-\max} = [\Pr(Y|X=1,Z=1) - \Pr(Y|X=0,Z=1)] - [\Pr(Y|X=1,Z=0) - \Pr(Y|X=0,Z=0)].^{14} \quad (1)$$

For the true relationship depicted in the left side of panel A in Figure 1,  $\Delta\Delta_{\min-\max} = (0.3 - 0.1) - (0.8 - 0.3) = 0.2 - 0.5 = -0.3$ . When there is *no* interaction between  $X$  and  $Z$  in influencing  $\Pr(Y)$ ,  $\Delta\Delta_{\min-\max}$  equals zero; and a second difference that is not zero implies some amount of

<sup>12</sup> Operationally, each nonlinear DGP was 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  $Z = 1$  was established as a different quadratic function. Then it was 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.

<sup>13</sup> Table S-2 in the supporting information lists the key characteristics of each of the 230 simulated true relationships, thereby documenting the diversity in the character of the relationships.

<sup>14</sup> Researchers frequently report a second difference to describe the extent of interaction in a logit or probit model (e.g., Haspel and Knotts 2005; Basinger and Lavine 2005).

interaction.<sup>15</sup> In general, as the interaction between  $X$  and  $Z$  gets stronger, the magnitude of  $\Delta\Delta_{\min-\max}$  will rise.

For 98 of the 230 true relationships, the min-max second difference is set at zero, making the DGP additive (e.g., panels C and D of Figure 1). In the remaining 132 true relationships, there is interaction between  $X$  and  $Z$  in influencing  $\Pr(Y)$  (e.g., those in panels A and B). In the DGPs with interaction, the min-max second difference ranges across ten values:  $\pm 0.50$ , reflecting extremely strong interaction (e.g., the right side of panel A),  $\pm 0.40$  (e.g., the right side of panel B);  $\pm 0.30$  (e.g., the left side of panel A);  $\pm 0.20$  (e.g., the left side of panel B); and  $\pm 0.10$ , reflecting weak but still nontrivial interaction.

### **Constructing Data Sets Reflecting Known “True” Relationships**

For each of the 230 DGPs we simulate, we generate 100 data sets containing 1,000 observations each through the following process. For each observation in each data set for a DGP,

- we draw a value of  $X$  and a value of  $Z$  randomly from a uniform  $[0,1]$  distribution. For DGPs in which  $Z$  is dichotomous, we recode  $Z$  values less than 0.5 to 0 and values greater than 0.5 to 1.
- Next, we determine the value of  $\Pr(Y)$  for the observation using the DGP (which maps each combination of  $X$  and  $Z$  values into a probability).
- Finally, we determine an observed value for the BDV,  $Y$ . We do so by drawing a random number from the uniform  $[0,1]$  distribution; if this number is less than the observation’s  $\Pr(Y)$  value, we assign the observation a  $Y$  value of 1; otherwise, we set  $Y$  to 0.

Given our 230 DGPs, we have a total of 23,000 data sets for use in Monte Carlo analysis.

### **The Utility of a Product-Term Coefficient for Determining Whether There is Interaction in Influencing $\Pr(Y)$**

It has been common practice when testing the BDV interaction hypothesis in recent years to

---

<sup>15</sup> Note, however, that a min-max second difference of zero does not imply that there is no interaction, since this difference reflects only the “global” extent 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 value of  $X$  between 0 and 1.

include a product term in a logit model and treat a statistically significant coefficient for this term as a necessary condition for concluding that there is empirical support for the hypothesis. Berry, DeMeritt and Esarey (2010) show that a statistically significant product-term coefficient is neither necessary nor sufficient for meaningful interaction between variables in influencing  $\Pr(Y)$ . However, because these authors offer a logical argument assuming that logit accurately specifies one's theory, their argument is largely irrelevant to the context we examine in this paper: when, in a situation of specification ambiguity, one seeks to test the BDV interaction hypothesis. In this situation, does a test of statistical significance of the product-term coefficient yield reliable information about whether the variables interact in influencing  $\Pr(Y)$ ?

We answer this question using Monte Carlo analysis on our 23,000 simulated data sets generated from known "non-logit" DGPs. Using each data set, we estimate a logit model including  $X$ ,  $Z$  and  $XZ$  as independent variables. We divide the data sets into groups based on the extent of interaction in the true model generating a data set (as measured by the absolute min-max second difference). Then, within each group, we calculate the proportion of data sets in which the coefficient for  $XZ$  is statistically significant at the .05 and .01 levels using a two-tailed test. The results are in Table 1. If the statistical significance of the coefficient for the product term constitutes a reliable indicator of the presence of interaction in influencing  $\Pr(Y)$ , data sets generated from a DGP in which there is interaction should rarely yield a statistically insignificant product-term coefficient, and data sets generated from an additive DGP should rarely produce a significant coefficient.

The good news is that when a test of significance for the product-term coefficient (in a sample of 1,000) is treated as a test of the BDV interaction hypothesis against the null hypothesis of no interaction [i.e., the hypothesis that  $X$  and  $Z$  have additive effects on  $\Pr(Y)$ ], there is a very

low probability of falsely rejecting the null hypothesis (i.e., committing a type I error). Indeed, across the 98 DGPs involving no interaction between  $X$  and  $Z$  in influencing  $\Pr(Y)$ , it is rare that an estimated coefficient for  $XZ$  is statistically significant even at the .05 level; in the 9,800 additive logit models estimated, significance is achieved just 3% of the time when  $Z$  is continuous, and only 4% of the time when  $Z$  is dichotomous. These percentages are remarkably close to those expected with a reliable test of statistical significance at the .05 level due to chance variation across samples.

However, the test has a very high probability of falsely accepting the null hypothesis of no interaction (i.e., making a type II error), thereby giving the test very low power, especially when  $Z$  is continuous. Our Monte Carlo analysis shows that even among true relationships with extremely strong interaction between  $X$  and a continuous  $Z$ —a min-max second difference ( $\Delta\Delta_{\min-\max}$ ) with magnitude 0.5—the estimated coefficient for  $XZ$  is significant at the .05 level only 50% of the time. Fully half the time, we would falsely conclude that the null hypothesis of additivity should be accepted. When  $\Delta\Delta_{\min-\max}$  declines to  $\pm 0.30$ , still indicating strong interaction, the estimated coefficient for  $XZ$  is significant at the .05 level only 21% of the time with a continuous  $Z$ ; 79% of the time we would incorrectly reject the hypothesis of interaction. Although the probability of a false inference is lower in the case of a DGP in which  $Z$  is dichotomous, the risk remains substantial: the product-term coefficient is statistically significant only 52% of the time.

Thus, if one relies on whether an estimated coefficient for a product term in a logit model is statistically significant to determine whether one should reject the null hypothesis of no interaction in favor of the BDV interaction hypothesis, the power of the test is so low that even when strong interaction is present there is a huge risk of falsely concluding that it is absent. This

leads us to recommend that political scientists abandon the practice of deeming a statistically significant product-term coefficient in a logit model a necessary condition for claiming empirical support for the BDV interaction hypothesis, just as Berry, DeMeritt, and Esarey (2010) recommend in the situation of strong theory.

### **The Performance of Alternative Logit Models in Testing the BDV Interaction Hypothesis**

When using logit to test the BDV interaction hypothesis, in what functional form should variables be included? In the situation of specification ambiguity—when theory is no guide—this question essentially asks what functional form for the model maximizes the ability of logit to approximate the true—but unknown—functional form for the DGP so that estimated effects of variables most closely approximate their true values.

#### **Alternative Logit Estimation Models**

To answer this question, using each simulated data set, we estimate a variety of logit models containing different combinations of terms involving  $X$  and  $Z$  and compare the performance of the alternative models. The *no-product* model includes just  $X$  and  $Z$  as independent variables.

The *product-term* model includes  $X$ ,  $Z$ , and  $XZ$ .<sup>16</sup> We refer to the no-product and product-term

---

<sup>16</sup> We also estimate five more “complex” logit models involving one or more second-order polynomial terms in  $X$  and/or  $Z$ , gradually adding more terms to the equation: (i)  $X$ ,  $Z$ ,  $XZ$  and  $X^2Z$ ; (ii)  $X$ ,  $Z$ ,  $XZ$  and  $XZ^2$ ; (iii)  $X$ ,  $Z$ ,  $XZ$ ,  $X^2Z$  and  $XZ^2$ ; (iv)  $X$ ,  $Z$ ,  $XZ$ ,  $X^2Z$ ,  $XZ^2$  and  $X^2Z^2$ ; and (v)  $X$ ,  $Z$ ,  $XZ$ ,  $X^2$ ,  $Z^2$ ,  $X^2Z$ ,  $XZ^2$  and  $X^2Z^2$ . These models are examined because adding more terms makes the functional form less restrictive and, thus, may allow logit to better approximate a variety of DGPs. However, our Monte Carlo analysis shows that adding one or more product and/or polynomial terms to a model including  $X$ ,  $Z$  and  $XZ$  does not improve performance in estimating the nature of a monotonic relationship, regardless of whether the DGP is linear or nonlinear; and regardless of whether the DGP is additive or interactive, and if interactive, the strength of the interaction. Consequently, a researcher seeking to test the BDV interaction hypothesis using logit—and thus in the situation of specification ambiguity—can limit consideration to models with and without and a single product term; there is no reason to employ specifications including additional product and/or polynomial terms. For this reason, we do not report any detailed results about the estimation models with multiple product terms, but these

models as *unconditional* estimation models because they involve the same functional form for every sample. However, we recognize the possibility that the relative ability of logit models with and without a product term to approximate the true functional form for a DGP may vary across data sets, and explore the possibility that model selection criteria may help determine whether including a product term would be beneficial in a particular sample. We consider four specific criteria: (i) the Akaike Information Criterion (AIC), (ii) the Bayesian Information Criterion (BIC), (iii) the Receiver Operating Characteristic (ROC) score, and (iv) a test of statistical significance (2-tailed, .05 level) for the product-term coefficient when a  $XZ$  term is included.

### **Quantities of Interest Measuring the Strength of Interaction**

We then use the parameter estimates from each of the alternative logit estimation models to generate quantities of interest reflecting the extent of interaction between  $X$  and  $Z$  in influencing  $\Pr(Y)$ . (The conditional estimation models generate these quantities using the model—product-term or no-product—recommended by the relevant criterion for the particular data set being analyzed.) First, we estimate the extent of interaction in “global” terms using the min-max second difference [as defined in equation (1), and denoted  $\Delta\Delta_{\min\text{-max}}$ ].<sup>17</sup> Second, we observe quantities reflecting the “local” nature of the interaction between  $X$  and  $Z$ , i.e., the response of  $\Pr(Y)$  to infinitesimally small changes in  $X$  and  $Z$ : second derivatives ( $\partial\Pr(Y)/\partial X\partial Z$ ) at 121 points

---

results are available in Table S-3 of the supporting information.

<sup>17</sup>  $\Delta\Delta_{\min\text{-max}}$  is a second difference across the entire range of values for  $X$  and  $Z$ . We also examine second differences involving less extreme values of  $X$  and  $Z$ :  $\Delta\Delta_{.10\text{-}.90} = [\Pr(Y|X=.9, Z=.9) - \Pr(Y|X=.1, Z=.9)] - [\Pr(Y|X=.9, Z=.1) - \Pr(Y|X=.1, Z=.1)]$ , and  $\Delta\Delta_{.20\text{-}.80} = [\Pr(Y|X=.8, Z=.8) - \Pr(Y|X=.2, Z=.8)] - [\Pr(Y|X=.8, Z=.2) - \Pr(Y|X=.2, Z=.2)]$ .  $\Delta\Delta_{.10\text{-}.90}$  and  $\Delta\Delta_{.20\text{-}.80}$  are second differences across a restricted range of  $X$  and  $Z$  values excluding their extremes. Since we find that the relative performance of alternative logit estimation models is very similar across the three second difference measures, we report results only for  $\Delta\Delta_{\min\text{-max}}$ . Results for  $\Delta\Delta_{.10\text{-}.90}$  and  $\Delta\Delta_{.20\text{-}.80}$  are presented in Table S-3 of the supporting information.

spread evenly over the  $X$ - $Z$  space.<sup>18</sup>

### **Assessing Performance Using RSMEs**

For each logit estimation model (the no-product model, the product-term model, etc.), we determine the *root mean squared error* (RMSE) of estimates of each quantity of interest (e.g., the min-max second difference) across the 100 data sets generated from each DGP.<sup>19</sup> This RMSE can be conceived as a measure of the performance of a logit model in estimating the quantity of interest, as the RMSE reflects the magnitude of the typical deviation of an estimate of the quantity from its true value. If an estimation model could always recover the true DGP exactly, the RMSE value for all quantities of interest characterizing the relationship would be zero; the more a RMSE value for a quantity diverges from zero, the greater the expected error in the estimate of the quantity.

Table 2 shows the average RMSE values for estimates of the min-max second difference and second derivatives derived from various logit estimation models.<sup>20</sup> Average RMSEs were originally calculated separately for three types of additive models—those in which neither  $X$  nor  $Z$  has an effect on  $\Pr(Y)$ , those in which only  $Z$  has an effect, and those in which both  $X$  and  $Z$  have effects—and five types of interactive models of varying strength, as reflected by absolute min-max second differences of 0.1, 0.2, 0.3, 0.4 and 0.5. However, because the performance of the estimation models for the three types of additive true relationships were very similar, we pool all additive DGPs into a single category in the table. The table shows results for models in

---

<sup>18</sup> For true DGPs in which  $Z$  is dichotomous, this second derivative is meaningless since  $Z$  is discrete but the derivative indicates the response of  $\Pr(Y)$  to an infinitesimally small change in  $Z$ . We measure instead, at 11 different values of  $X$  (0, 0.1, 0.2, ..., 0.9, 1), a close analog to the second derivative: the change in  $\partial\Pr(Y)/\partial X$  when  $Z$  shifts from one value (0) to the other (1).

<sup>19</sup> Appendix S-3 in the supporting information provides a detailed description of the RMSE measure of the performance of an estimation model.

<sup>20</sup> All Monte Carlo analyses for logit were replicated using probit with very similar results. Empirical results for probit estimations are presented in Table S-3 of the supporting information.

which  $Z$  is continuous and those in which  $Z$  is dichotomous separately—the former on the left, the latter on the right.<sup>21</sup>

### **Which Unconditional Logit Estimation Model is Best?**

We first examine the conditions under which each unconditional model performs best. Table 2 shows that the relative performance of the no-product and product-term models in estimating quantities of interest measuring the strength of interaction depends on the magnitude of interaction in the true relationship. If the BDV interaction hypothesis is wrong and the DGP is actually additive, a no-product model dramatically outperforms a product-term model, especially for DGPs in which both  $X$  and  $Z$  are continuous. For example, across all additive DGPs in which both  $X$  and  $Z$  are continuous, the average RMSE for a min-max second difference estimate based on a no-product model is 0.027—a value 0.135 lower than the average RMSE generated from a product-term model (0.162). In contrast, when the true DGP is strongly interactive (i.e.,  $|\Delta\Delta_{\min-\max}| = 0.3, 0.4$  or  $0.5$ ), a product-term model outperforms a no-product model by a substantial amount. For example, with both  $X$  and  $Z$  continuous, for DGPs for which true  $|\Delta\Delta_{\min-\max}| = 0.4$ , the average RMSE for a product-term estimate of the min-max second difference is 0.150 lower than the average RMSE for a no-product estimate (0.156 vs. 0.306).

However, the situation is quite different when the true DGP is characterized by less extreme—but hardly trivial—amounts of interaction:  $|\Delta\Delta_{\min-\max}| = 0.1$  or  $0.2$ . For such true relationships, the difference in performance between no-product and product-term estimates of the min-max second difference and second derivatives is much less pronounced, and which is superior depends on whether  $Z$  is continuous or dichotomous. For example, average RMSEs are

---

<sup>21</sup> We also compute average RMSE values for *linear* DGPs and *nonlinear* DGPs separately. Since the relative performance of estimation models is largely insensitive to whether the true relationships are linear or not, we present only analyses pooling linear and nonlinear DGPs. Disaggregated results are presented in Table S-3 of the supporting information.

lower for a product-term estimate when  $Z$  is dichotomous and  $|\Delta\Delta_{\min\text{-max}}| = 0.2$ , but greater for a product-term estimate when  $Z$  is continuous and  $|\Delta\Delta_{\min\text{-max}}| = 0.1$ . This latter result illustrates a conclusion from the Monte Carlo analysis that may seem counter-intuitive to those used to analyzing interaction in a regression model with an observed (unbounded) dependent variable: *in a situation of specification ambiguity, even when the independent variables interact in influencing  $\Pr(Y)$ , a logit model without a product term will sometimes yield a more accurate estimate of the strength of the interaction than a model including a product term.*

In summation, when the DGP is additive, one can expect to estimate the true min-max second difference and second derivatives (all of which are, in fact, zero) with less error using a logit model without a product term than using one with a product term. As the extent of interaction in the true DGP—as measured by the absolute value of the min-max second difference—increases, the performance of the no-product estimation model tends to decline and the performance of the product-term model tends to rise. When true interaction reaches moderate strength (i.e.,  $|\Delta\Delta_{\min\text{-max}}|$  in the range between 0.1 and 0.2), neither estimation model seems systematically superior. When the magnitude of interaction surpasses a tipping point in this moderate range, the expected error in estimating these quantities of interest is lower when estimates are derived from a logit model with a product term than when they are derived from a model without one.

Knowing which unconditional estimation model—no-product or product-term—is superior for DGPs with known extents of interaction is of little practical significance, however, because we can never know the true nature of the DGP (additive, moderately interactive, or strongly interactive) in an applied setting. Consequently, we next investigate whether one can outperform the two unconditional estimation models by deciding whether to include a product

term based on empirical model selection criteria.

### **Does a Conditional Model—Based on Model Selection Criteria—Outperform the Unconditional No-Product and Product-Term Estimation Models?**

Recall that we investigate four *conditional* logit estimation models, each informed by a different model selection criterion:

- which specification (including a product term, or excluding it) has the lowest Akaike Information Criterion (AIC) value, yielding the *AIC-informed* estimation model;
- which specification has the lowest Bayesian Information Criterion (BIC) value, yielding the *BIC-informed* model;
- which specification has the higher Receiver Operating Characteristic (ROC) score, yielding the *ROC-informed* estimation model; and
- a test of statistical significance (at the .05 level, 2-tailed) for the product-term coefficient when a *XZ* term is included, yielding the *significance-informed* model.

Each conditional model generates estimates of quantities of interest measuring the magnitude of interaction by employing the model—with or without a product term—recommended by the relevant selection criterion for the particular sample being analyzed.

Our Monte Carlo analysis shows that the AIC-informed estimation model outperforms the other three conditional models. Therefore, to save space, we do not report performance results for the BIC-informed, ROC-informed or significance-informed estimation models. (Results for these three models are presented in Table S-3 in the supporting information, and a comparison of their performance relative to that of the AIC-informed model is offered in Table S-4.) Table 2 permits a comparison of the performance of the AIC-informed model to that of both the no-product and product-term models. In particular, the relative performance of the three models is observed in 24 contexts formed by the intersection of six magnitudes for the true min-max second difference (0, 0.1, 0.2, 0.3, 0.4 and 0.5), two types of *Z* variables (continuous and dichotomous), and the two quantities of interest reflecting the extent of interaction. Based on

mean RMSEs, the performance of the AIC-informed model is nearly always better than whichever unconditional model (no-product or product-term) is worst in a context, but nearly uniformly worse than whichever unconditional model is best in the context. Of the 24 cases, this result holds in all but three, and in each of the three exceptions, the RMSE value for the AIC-informed model is always within 0.01 of making the claim true. The AIC-informed estimation model might, therefore, be conceived as a “risk-averse” choice—avoiding the worst possible performance by sacrificing the opportunity for the best. Consequently, a researcher who is risk averse might favor the AIC-informed model over the no-product and product-term models on this ground alone.

However, there is another reason to prefer the AIC-informed model over the two unconditional models. When shifting from an unconditional estimation model to the AIC-informed model, the *decrease* in expected RMSEs by moving away from the worst unconditional model tends to be larger than the *increase* in expected RMSEs by moving away from the best model. Let us label the amount by which the average decrease in RMSE by using the AIC-informed model rather than the worst unconditional model exceeds the average increase in RMSE by using the AIC-informed model instead of the best unconditional model the *net benefit from using the AIC-informed model*; when this value is negative, it will be called a *net cost*. For example, in the continuous  $Z$ /additive context, the average RMSE for the min-max second difference is 0.027 for the no-product estimate, 0.126 for the AIC-informed estimate, and 0.162 for the product-term estimate. One reduces the expected RMSE by 0.036 ( $= 0.162 - 0.126$ ) by using the AIC-informed estimation model rather than the inferior product-term model. In turn, one increases the expected RMSE by 0.099 ( $= 0.126 - 0.027$ ) by employing the AIC-informed model instead of the superior no-product model, thereby suffering a net cost from using the AIC-

informed model of 0.063 (= 0.099 – 0.036). Indeed, it is in this context in which the net cost from using the AIC-informed model is at its largest. But across the 24 contexts, there are nine in which there is a net *benefit* from using the AIC-informed model greater than 0.063, and in six of these contexts the net benefit exceeds 0.10.<sup>22</sup> Moving to a lower threshold, across the 24 contexts, there are five in which the net cost from using the AIC-informed model exceeds 0.03, but ten contexts in which there is a net benefit larger than 0.03. Summarizing, our Monte Carlo analysis shows that when testing the BDV interaction hypothesis by estimating how the effect of  $X$  on  $\Pr(Y)$  varies with  $Z$ , the expected benefit of shifting from an unconditional model to the AIC-informed model exceeds the cost.

### Conclusion and Recommendations

Political scientists are often interested in testing what we label the *BDV interaction hypothesis*: two variables,  $X$  and  $Z$ , interact such that at each value of  $Z$ , the relationship between  $X$  and  $\Pr(Y)$  is monotonic (either positive or negative), but (at each value of  $X$ ) the magnitude of the effect of  $X$  on  $\Pr(Y)$  changes (consistently getting stronger or weaker) as  $Z$  increases. When a researcher tests this hypothesis without a strong theory that implies a specific functional form for the data generating process (DGP), we describe her as being in the situation of *specification ambiguity*: her theory is precise enough to imply that the DGP is monotonic but is insufficiently specific to imply that logit is the correct functional form for the DGP. Logit is certainly consistent with this hypothesis, but so are many other functional forms.

Despite this specification ambiguity, the usual practice is to test the BDV interaction hypothesis using a logit model including a product term ( $XZ$ ). If the coefficient estimate for the

---

<sup>22</sup> The nine contexts are: for the min-max second difference, continuous  $Z/|\Delta\Delta_{\min-\max}| = 0.4$  or **0.5**, and dichotomous  $Z/|\Delta\Delta_{\min-\max}| = \mathbf{0.3}$ , **0.4** or **0.5**; and for second derivatives, continuous  $Z/|\Delta\Delta_{\min-\max}| = 0.5$ , and dichotomous  $Z/|\Delta\Delta_{\min-\max}| = 0.3$ , **0.4** or **0.5**. (The six contexts in which net benefit exceeds 0.1 have their  $|\Delta\Delta_{\min-\max}|$  value in bold text in this note.)

product term is statistically insignificant, the hypothesis of interaction is rejected. If it is significant, the logit coefficients are used to estimate the magnitude of interaction, i.e., the extent to which the effect of  $X$  on  $\Pr(Y)$  varies with  $Z$ .

A test of this procedure using Monte Carlo analysis reveals that it can yield misleading results about whether interaction is present. The problem is that a test for statistical significance of the coefficient for the product term in a logit model is a highly misleading indicator of whether the BDV interaction hypothesis is true, due to the test's low power. With a sample size as large as 1,000, even when the DGP is very strongly interactive, the coefficient for  $XZ$  will fail to be significant at the .05 level a large percentage of the time. For this reason, researchers should abandon the practice of treating a significant product-term coefficient as a necessary condition for the BDV interaction hypothesis to hold. They should also recognize that a logit model with a product term is not always the best specification for testing the BDV interaction hypothesis. Although a model with a product term can be expected to outperform a model without one in estimating the effects of  $X$  and  $Z$  on  $\Pr(Y)$  when the true relationship is very strongly interactive, as the magnitude of interaction in the DGP declines, the expected performance of the no-product model rises, surpassing that of the product-term model well before the strength of interaction declines to trivial levels. Since the character of the true DGP is unknown to the analyst, an unconditional reliance on either specification for testing the BDV interaction hypothesis is unwarranted, and the analyst can outperform both by employing model selection criteria to choose the best specification. Among the criteria we examined, the Akaike Information Criterion (AIC) proved superior.

Based on our findings, we advise researchers testing the BDV interaction hypothesis to undertake the following steps:

- **Estimate a logit model both with and without a product term. The latter should include  $X$  and  $Z$  (and any other independent variables); the former should include  $XZ$  as well. Determine which model yields the lower AIC value.**
- **Use the parameter estimates for the model with the lower AIC value—either that involving a product term, or that excluding one—to estimate how the effect of  $X$  on  $\text{Pr}(Y)$  varies with  $Z$ .**

Excellent tools are available for this purpose. King, Tomz and Wittenberg's (2000) CLARIFY add-on to Stata can produce second difference point estimates (e.g., the min-max second difference, or a second difference over a more restricted range of  $X$  and  $Z$  values), along with confidence intervals that permit a test of statistical significance. Brambor, Clark and Golder's (2006) Stata code can be adapted to generate a point estimate of the marginal (or instantaneous) effect of  $X$  on  $\text{Pr}(Y)$  at any value of  $Z$  (along with a confidence interval) so that  $X$ 's marginal effect can be compared across different values of  $Z$ .<sup>23</sup>

Finally, we recommend sensitivity analysis:

- **Use the parameter estimates for the model with the *higher* AIC value to estimate how the effect of  $X$  on  $\text{Pr}(Y)$  varies with  $Z$ . If the substantive conclusions about the BDV interaction hypothesis from this analysis closely match those derived from the model with the lower AIC value, model selection becomes a moot issue, lending greater credibility to the results.**

Our Monte Carlo analysis suggests that general agreement between conclusions based on the models with and without the product term may not be exceedingly rare. Across our 23,000 simulated data sets, (i) confidence intervals for the min-max second difference derived from the product-term and no-product models lead to the same conclusion about whether this second difference is statistically significant at the .05 level 52% of the time, and (ii) the two point estimates of the second difference have magnitudes within 0.10 of each other 32% of the time, and within 0.05 17% of the time.

---

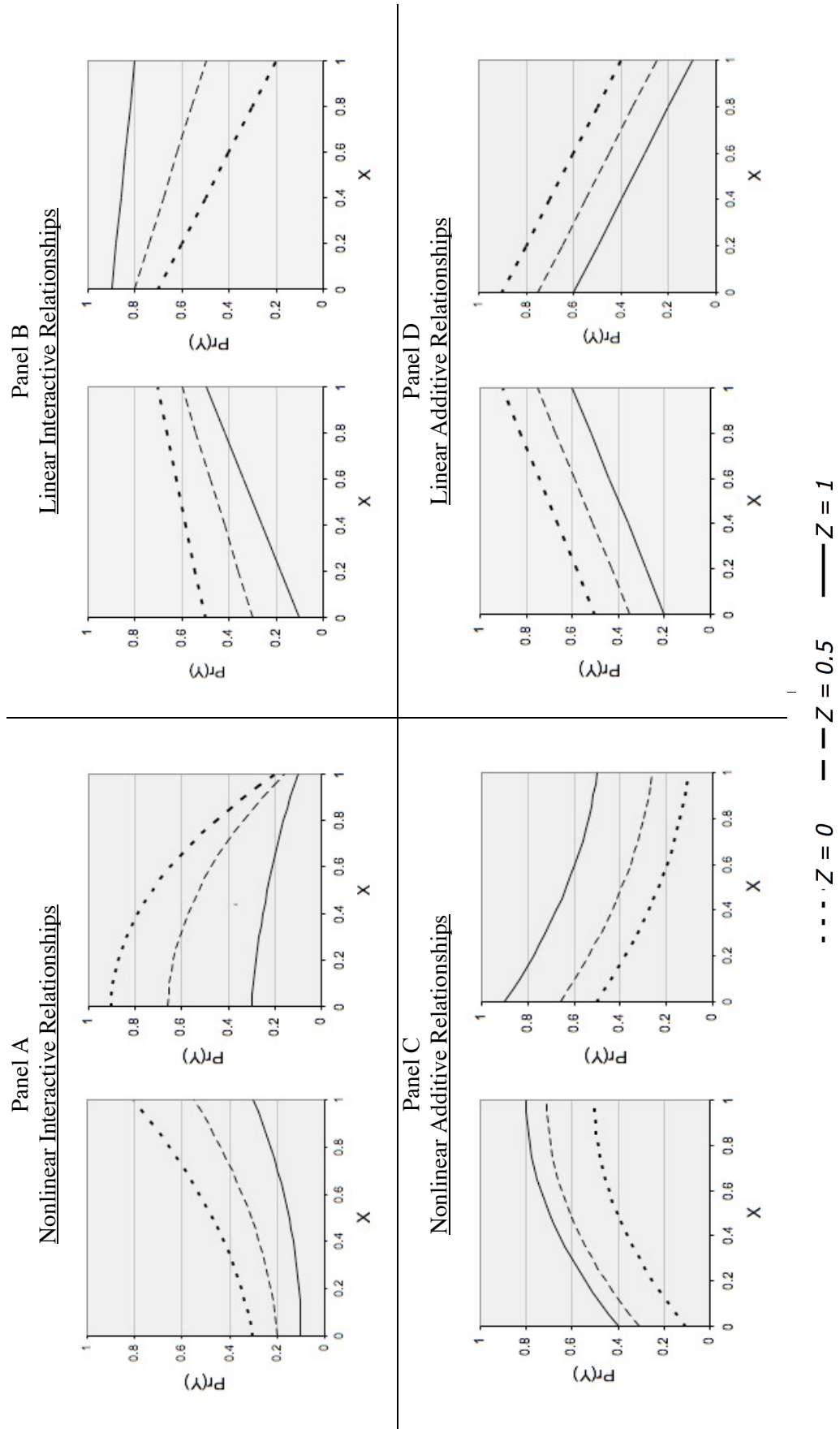
<sup>23</sup> At <http://www....edu>, we provide detailed examples of how Stata can be used to implement the analytical approach we recommend.

We conclude with a final caution. Our advice to readers to use the Akaike Information Criterion to guide the choice about the functional form of a logit model when testing the BDV interaction hypothesis rests on the assumption that the theory being tested is insufficient to favor one monotonic functional form over any other. When there are strong reasons to prefer a specific logit specification over other functional forms—e.g., firm theoretical guidance shaped by past scientific findings—these reasons should take precedence over the AIC value in a specific sample.

## References

- Aldrich, John H. and Forrest D. Nelson. 1984. *Linear Probability, Logit, and Probit Models*.  
Newbury Park, CA: Sage.
- Basinger, Scott J. and Howard Lavine. 2005. "Ambivalence, Information and Electoral Choice."  
*American Political Science Review* 99 (May):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 (January):248-66.
- Brambor, Thomas, William Clark and Matt Golder. 2006. "Understanding Interaction Models:  
Improving Empirical Analyses." *Political Analysis* 14 (Winter):63-82.
- Cramer, J. S. 2003. *Logit Models from Economics and Other Fields*. Cambridge: Cambridge  
University.
- Haspel, Moshe and H. Gibbs Knotts. 2005. "Location, Location, Location: Precinct Placement  
and the Costs of Voting." *Journal of Politics* 67 (February):560-73.
- King, Gary, Michael Tomz and Jason Wittenberg. 2000. "Making the Most of Statistical  
Analyses: Improving Interpretation and Presentation." *American Journal of Political Science*  
44 (April):341-55.

**Figure 3. C'Uco rrg of vj"True Relationships Involving X, Z'and Pr(Y)**



Note: The equations for these DGPs are in Appendix S-2 in our unpublished supplement (pp. 6-17). Panel A shows DGP #4 (left) and DGP #24 (right). Panels B, C and D show DGPs with codes 80, 32, 43, 45, 53 and 56.

**Table 1. The Ability of Information about the Statistical Significance of the Product-Term Coefficient to Yield a Correct Inference About Whether a Relationship Among  $X$ ,  $Z$  and  $\Pr(Y)$  is Interactive**

Type of DGP	continuous-Z DGPs		dichotomous-Z DGPs	
	proportion of time product-term coefficient is significant at .01 level (2-tail test)	proportion of time product-term coefficient is significant at .05 level (2-tail test)	proportion of time product-term coefficient is significant at .01 level (2-tail test)	proportion of time product-term coefficient is significant at .05 level (2-tail test)
additive (i.e., $\Delta\Delta_{\min-\max} = 0$ )	0.00	0.03	0.01	0.04
interactive: $ \Delta\Delta_{\min-\max}  = 0.10$	0.02	0.06	0.06	0.14
interactive: $ \Delta\Delta_{\min-\max}  = 0.20$	0.04	0.13	0.20	0.35
interactive: $ \Delta\Delta_{\min-\max}  = 0.30$	0.09	0.21	0.37	0.52
interactive: $ \Delta\Delta_{\min-\max}  = 0.40$	0.15	0.32	0.60	0.76
interactive: $ \Delta\Delta_{\min-\max}  = 0.50$	0.28	0.50	0.79	0.88

**Table 2. The Performance of Alternative Models in Estimating Effects on Pr(Y)**

Type of DGP	Estimation Model	Average Root Mean Square Error (RMSE)		Type of DGP	Estimation Model	Average Root Mean Square Error (RMSE)	
		Difference	2nd Derivatives Spread Over X-Z Space			Difference	2nd Derivatives Spread Over X-Z Space
continuous Z: additive (i.e., $\Delta\Delta_{\min-\max} = 0$ )	no product	0.027	0.070	dichotomous Z: additive (i.e., $\Delta\Delta_{\min-\max} = 0$ )	no product	0.031	0.062
	product term	0.162	0.194		product term	0.095	0.124
	AIC-informed	0.126	0.160		AIC-informed	0.079	0.109
	signif-informed	0.073	0.111		signif-informed	0.056	0.085
continuous Z: interactive $ \Delta\Delta_{\min-\max}  = 0.10$	no product	0.091	0.152	dichotomous Z: interactive $ \Delta\Delta_{\min-\max}  = 0.10$	no product	0.094	0.135
	product term	0.162	0.226		product term	0.095	0.144
	AIC-informed	0.149	0.213		AIC-informed	0.098	0.147
	signif-informed	0.122	0.187		signif-informed	0.099	0.146
continuous Z: interactive $ \Delta\Delta_{\min-\max}  = 0.20$	no product	0.176	0.204	dichotomous Z: interactive $ \Delta\Delta_{\min-\max}  = 0.20$	no product	0.166	0.195
	product term	0.159	0.221		product term	0.094	0.143
	AIC-informed	0.174	0.229		AIC-informed	0.109	0.155
	signif-informed	0.186	0.229		signif-informed	0.130	0.171
continuous Z: interactive $ \Delta\Delta_{\min-\max}  = 0.30$	no product	0.243	0.279	dichotomous Z: interactive $ \Delta\Delta_{\min-\max}  = 0.30$	no product	0.233	0.263
	product term	0.157	0.253		product term	0.093	0.166
	AIC-informed	0.188	0.272		AIC-informed	0.112	0.178
	signif-informed	0.227	0.290		signif-informed	0.147	0.202
continuous Z: interactive $ \Delta\Delta_{\min-\max}  = 0.40$	no product	0.306	0.357	dichotomous Z: interactive $ \Delta\Delta_{\min-\max}  = 0.40$	no product	0.295	0.329
	product term	0.156	0.291		product term	0.090	0.177
	AIC-informed	0.200	0.318		AIC-informed	0.101	0.185
	signif-informed	0.259	0.353		signif-informed	0.137	0.210
continuous Z: interactive $ \Delta\Delta_{\min-\max}  = 0.50$	no product	0.375	0.429	dichotomous Z: interactive $ \Delta\Delta_{\min-\max}  = 0.50$	no product	0.350	0.393
	product term	0.143	0.319		product term	0.093	0.193
	AIC-informed	0.183	0.342		AIC-informed	0.099	0.197
	signif-informed	0.265	0.389		signif-informed	0.123	0.213