International Studies Quarterly (2007) **51**, 231–241

Detecting Model Dependence in Statistical Inference: A Response

GARY KING

Harvard University

LANGCHE ZENG

University of California-San Diego

When counterfactual questions are posed too far from the data, statistical inference can become highly dependent on seemingly trivial but indefensible modeling assumptions. Choosing one or only a few specifications to publish in the presence of model dependence makes conclusions largely nonempirical and subject to criticism from others who make different choices and find contradictory results. Until now, only relatively narrow attempts to assess model dependence have been used in the literature (such as trying a few different functional form specifications) and, as a result, many published inferences are far more uncertain than standard errors and confidence intervals indicate. When researchers ignore the uncertainty due to model dependence, scholarly works tend to have the flavor of merely showing that it is possible to find results consistent with *ex ante* hypotheses, rather than demonstrating their veracity. Consequently, the opportunities for researchers to use methods such as the ones we offer to learn new facts about existing studies and avoid problems in their research are considerable.

Although the comments contain a diversity of views, we are gratified to see that all three endorse our central message. Morrow: "It is important to know when counterfactual statements drawn from statistical estimates wander far from the data used in estimates." Schrodt: "I am not contesting the general cautions made by the authors: models with different functional forms can diverge substantially, predictions made for sets of independent variables similar in value to those used to estimate a model are more likely to be accurate than predictions for more distant values, and specification error can really mess up regression models." Sambanis and Doyle (S&D): "statistical results should not be taken too far [from the data] and . . . any extrapolation depends on the model."

On some other issues, the reviewers disagree with each other or with us. Fortunately, most of the disagreements are based on well-posed mathematical questions that can be cleanly resolved with easily provable answers, without the need for philosophical interpretation and lengthy debate. We appreciate this opportunity to resolve these issues in the same volume where our work originally appears, thank the reviewers for their time, effort, and thoughtful comments, and point out to readers who may desire further information that the mathematical proofs underlying our methods appear in a technical companion article (King and Zeng 2006a),

Author's note: We thank Jim Alt, Neal Beck, Michael Doyle, Mitchell Duneier, Jim Morrow, Phil Schrodt, Nicholas Sambanis, and the editors and reviewers for helpful comments on an earlier version of this piece and the National Institutes of Aging (P01 AG17625–01) and the National Science Foundation (SES-0318275, IIS-9874747) for research support.

easy-to-use software is available to implement all the methods discussed herein (Stoll, King, and Zeng 2006), and replication data sets are available for our original article and this response (King and Zeng 2006b, 2006c).

Our article proposed three innovations. First, we introduced a method based on the convex hull concept for distinguishing extrapolations from interpolations. A wide range of statistics and methods literatures recognize that extrapolation is generally more hazardous, and produces more model-dependent inferences, than interpolations, although technical limitations in convex hull algorithms have prevented application of these ideas to data sets with more than a few variables. The algorithm we developed makes the computation feasible for the larger numbers of variables common in political science. Second, we prove mathematically for the first time what was understood only intuitively before, that the greater the distance from a counterfactual to the available data, the more model-dependent inferences can be about that counterfactual. We also recommended that the Gower distance (and geometric variability threshold) be used in combination with our convex hull test to compute measures of how much data is nearby a counterfactual of interest. The importance of these tests lies in their ability to indicate the likelihood of model dependence without having to run any alternative models, and so they are much more broadly applicable, powerful, and easy to use than standard sensitivity testing. Finally, we offer a new four-part decomposition of the potential biases affecting causal inference, which generalizes that of Heckman et al. (1998), to include and highlight problems more common in political science, and especially prevalent and misunderstood in IR and comparative research. This decomposition helps convey some of the pitfalls of counterfactual inference in causal effect estimation, and would be useful for teaching too.

We now discuss the main issues that have arisen in the comments.

Morrow on Theory

Whereas the methods introduced in our article attempt to answer the question "when can history be our guide?," James Morrow emphasizes that theory can be a guide, too. We agree with Morrow's point, which is especially crucial when theoretical information is available and empirical information is not. Our point is merely that researchers should not be hoodwinked into thinking that their conclusions are based on historical evidence when in fact they are based only on (sometimes unstated, undefended, or unjustifiable) theory. We add to Morrow's collection of fictional law enforcement officers Sherlock Holmes, who said "It is a capital mistake to theorize before you have all the evidence. It biases the judgment" (Doyle 1888:30). The methods we offer help ensure that your judgments are based on evidence.

Although the role of theory does not apply directly to the validity of our tests, conditional as they are on the choice of explanatory variables and counterfactual questions, Morrow's general concern about theory remains an excellent one. We thus take his lead and address the issue in a broader context.

Suppose you learn, with our methods or others, that your counterfactual question is far enough from your data so that your conclusions will be model dependent. What is a political scientist to do? This question has five possible answers: (1) Live with the uncertainties that result from model dependence and convey them to readers. (2) Obtain more observations on the same explanatory variables for values closer to the counterfactual, a valuable strategy but not always feasible. (3) Develop a theory to choose among the alternative models that would otherwise lead to model-dependent conclusions. Few theories derive specific functional forms from unassailable principles, and so this is only occasionally practical (see Signorino 1999). In applying this strategy, we must avoid theories with assumptions chosen based on convenience rather than knowledge (and where "I derived it from theory" means "I made it up")—a useful simplifying device for exploration but

misleading for empirical analysis. (4) Elaborate the original theory to find new observable implications, or "novel hypotheses" in Morrow's terminology, that are closer to the counterfactual. This is a valuable strategy in most research, even when a counterfactual question has sufficient supporting evidence. (5) Substitute the counterfactual question for one closer to the available data and still of some theoretical interest.

Finally, Morrow recalls his college friend theorizing about what "the universe [and his car] would look like if the speed of light was 30 miles per hour," a "counterfactual that has no empirical basis whatsoever and would fail King and Zeng's tests spectacularly." Despite an extrapolation from 670 million miles per hour, Morrow says this counterfactual had credibility because it relied on the well-supported theory of general relativity.[1] But as Morrow also recognizes, even good theories need empirical validation. As it happens, although it took scientists considerable efforts, in 1999, Lene Hau finally did succeed in slowing the speed of light in her laboratory to just 38 miles per hour, and eventually stopped it altogether (Hau et al. 1999). We do not know whether the considerable body of empirical work that resulted from this research has yet confirmed what a certain MG-B sports car would look like in this light, but we do have a recommended speaker for Morrow's next college reunion!

Schrodt on Dinos

We appreciate Phil Schrodt's "enthusiastic" anonymous review. However, his entertaining and critical comment here is even more useful, even though, as we demonstrate, each of its six points turn out to be false. The critique offers us the opportunity to highlight and correct the honest misunderstanding underlying each of its false claims, which should prove highly valuable in helping other researchers avoid making these same mistakes. Although the exposition of our work is much improved when supplemented by these clarifications, the accuracy, scope, and originality of the claims we offered are as originally summarized in our article and proven mathematically in King and Zeng (2006a).

First, Schrodt claims that post-treatment bias is "more conventionally known as endogeneity or simultaneity" and that techniques like "instrumental variables and simultaneous equation models" address the problem. These claims are incorrect. Endogeneity (or simultaneity) bias results when the designated dependent variable actually causes the designated explanatory variable, but the analysis assumes only the reverse is true. In contrast, post-treatment bias says nothing about the dependent variable and refers only to the bias due to controlling for variables partially caused by the key causal (or "treatment") variable. Thus, an analysis may have no endogeneity bias but still high levels of post-treatment bias. For example, suppose the key explanatory variable is employment status and the dependent variable indicates whether an individual votes. As unemployed people may be busy finding jobs, unemployment may cause voter turnout to drop. Suppose also that the data include black factory workers in the 1960s American South whose employer threatened to fire anyone caught showing up at the polls. Then voting would be a cause of employment status as well, and running a single equation model with turnout as the dependent variable would result in endogeneity bias. Now suppose none of our respondents worked for the racist employer and turnout decisions do not cause employment. If we also controlled for reported intention to vote from a poll held the day before the election, endogeneity bias would not be a problem, but we would be left with posttreatment bias, as this variable is in large part a

[1] In fact, a light speed of 30 miles per hour is an incredible stretch of general relativity theory in the weak gravitational field of the Earth (and even near the Sun, which is about 300,000 times heavier, Einstein 1920:88–89.) However, it is indeed allowed under the theory and empirically credible if we let light travel (and have Morrow's friend drive) through certain exotic refractive media.

consequence of the treatment variable, employment status. Intuitively, adding vote intentions would incorrectly control away most of what should be attributed to the causal effect of employment status. Our paper shows that posttreatment bias is a particularly serious problem in IR, and avoiding or treating endogeneity has no necessary effect on posttreatment bias. The problem of posttreatment bias is old, but widespread recognition of it is much more recent.[2]

Second, Schrodt claims "we already have a perfectly good measure that accomplishes what the Gower distance supposedly accomplishes and more: the variance of the error of the prediction at a specific point . . .." This claim is wrong. The issue at hand is assessing model dependence, and Schrodt would have us use a measure that is dependent on the veracity of the assumed regression model. In contrast, the Gower distance and convex hull measures make no modeling assumptions. The variance calculation, detailed in the two algebraic equations displayed prominently in his paper, is valid only "under the usual regression assumptions." These measures have indeed been "available for decades," and what they mean has been "known for decades," but they (and their statistical properties) require the assumption that the linear regression model holds exactly, so they are useless for assessing model dependence. See King and Zeng (2006a:2) for an example.

Third, Schrodt claims that the Gower distance "applies only to ratio data, the most restrictive category of measurement." This claim is wrong. Ratio-level measurement is unnecessary, as Gower distance does not require division but only normalization by range, which works even for nominal data.[3] Indeed, the advantage of the Gower distance measure is that it works for variables from any level of measurement (Gower 1971) that makes sense as explanatory variables in a statistical model, and so it does indeed apply to "almost all statistical models used and theoretical processes hypothesized in the discipline." For example, a multinomial variable like religion with J categories can be included in the model by using $J - 1$ indicators, as is taught in all our basic regression classes. But once this is done, Gower's formula applies directly. If the coding makes sense for an explanatory variable in any chosen statistical model, the distance of a counterfactual from the data used in this model can be measured by the Gower distance. No new assumptions or codings are required.

Fourth, Schrodt claims that "the right graph in [our] Figure 4 . . . involves reversing *all* of the instances of a key variable" and is therefore "nonsensical." This claim is false. The figure was based on 122 *separate and independent* counterfactual evaluations. We presented them all together, which perhaps accounts for the confusion, but reversing was done one at a time. Indeed, our methods are defined only for single-row counterfactuals. Thus, for the first civil war in the data, we held constant all the explanatory variables at their observed values, changed only whether the UN intervened, and then checked this one counterfactual against the entire (factual) data set. We then start over and evaluate a counterfactual involving the second civil war by reversing the UN decision to intervene only in this second civil war, leaving the rest of the data set unchanged, and so on. Imagining what would happen if all 122 UN decisions to intervene were simultaneously reversed would indeed be "nonsensical," but that was not done.

Fifth, Schrodt claims as his novel discovery that it is possible for a point just outside the interpolation region to be closer to a large amount of data than one

[2] Despite Schrodt's memory of taking "graduate level econometrics in the early 1970s," where he says "about half of the course dealt with this problem in various guises," no discussion of this issue appeared in standard 1970s, or even 1980s, econometrics texts or any others we have been able to find. Most of these texts, and much of comparative and IR research, do discuss endogeneity at length.

[3] For example, two observed values of a binary x can only be either the same or different, and the corresponding normalized Gower distance would be either 0 or 1, just as we would want. This is regardless of how the x values are coded (any x_1 and x_2 would do), even if $x = 0$ is not defined, and despite the fact that division operations like x_1/x_2 may make no sense.

inside the hull but in a "hole" away from most data. In fact, our article states that "it is theoretically possible (although probably empirically infrequent) for a point just outside the interpolation region . . . to be closer to a large amount of data than one inside the hull that occupies a large empty region away from most of the data," and so a discovery of Schrodt's this is not. Substantively, the exception described can happen in only a trivially small fraction of the space. This is even true in Schrodt's figures, where the potential space for counterfactuals extends miles above and to the right of the graphs at the scale he chose to draw. If we imagine re-plotting his figures by extending both vertical and horizontal axes (to their logically possible ranges, or for the "C" versions of his figures merely using the original scale of the "A" versions) and "zooming out" for a broader perspective, it is easy to see this point and how tiny the exceptions are.

As such, we and the entire statistics community describe extrapolation as generally more difficult, and model dependent, than interpolation. Although most in the literature ignore the rare exception at issue, we introduced the Gower distance as an additional check to detect and avoid it. With both tools, we can easily identify this potential problem, even (to quote Schrodt) for the "few sane political scientists [who] would fit a quartic."

Relatedly, Schrodt claims that "the convex hull itself is dependent on outliers." In fact, like all functional form sensitivity tests, the hull test is conditional on the set of explanatory variables chosen for analysis. Just as fixing endogeneity bias will not help posttreatment bias, fixing outliers will not resolve extrapolation bias or vice versa. "We assume outliers are removed as part of the important data preprocessing procedures normally used in standard statistical modeling . . .. In the inadvisable situation where a researcher ignores the [outlier] problem and persists with checking whether the counterfactual is outside the convex hull, outliers in X would make this extrapolation-detection method overly conservative" (King and Zeng 2006a:10, fn. 10).

Finally, Schrodt questions the breadth of our claim that our procedures work "for the class of nearly all models, whether or not they are formalized, enumerated, and run, and for the class of all possible dependent variables." In fact, the advantage of the tests we proposed is that they give some of the same benefits in detecting model dependence as would come from running an infinite number of alternative specifications. Similarly, as specifying the counterfactual, and applying our tests, do not require the dependent variable and can even be performed before it is collected, inferences from our procedures apply to any possible dependent variable. If our claim sounds broad, it is only because the usual approach to sensitivity testing is so narrow and restrictive. The point in our quote above is accurate.[4]

Many of Phil Schrodt's claims—which we can imagine him summarizing as "pass the sauce, and I'll throw another dino on the barbie"—may seem somewhat more entertaining than enlightening, but we hope readers agree with us that the exchange in total has been productive and we appreciate the opportunity he has given us to help readers avoid these same misunderstandings.

Doyle and Sambanis on Sambanis and Doyle

S&D have devoted remarkable attention to the few pages in our article replicating one logit model from Doyle and Sambanis (2000). Their comment in this issue is one of *six* papers they wrote totaling 408 pages of text, and is accompanied by an additional 103 text, program, and data files, all devoted solely to studying the same few pages in our article (on the web site they cite, accessed 10/11/2006).

[4] Relatedly, Schrodt asks "Specification matters . . . This is news?" Of course not. What is news is that our methods can detect when a specification change that has no visible effect on fitting the data can be consequential for counterfactual inference, without knowing which change it is and without the need to run alternative models.

We chose Doyle and Sambanis (2000) to illustrate how to detect model dependence in practice. We addressed only this one methodological aspect of only one of their article's 10 listed hypotheses. We chose their analysis on UN peacebuilding effectiveness for further study because it was the main focus of their article on "International Peacebuilding." We take it from their response that it is at least central to their work.

Their APSR article finds an apparent effect of a conglomerate measure of UN interventions and then "unpacks" it into its five component measures. S&D now believe that the one component we studied, multidimensional peacekeeping operations, is "not the single solution to all peacebuilding challenges," but this component was the only one of the five with any supporting evidence. Of the five components unpacked, they report four as having no effect: (1) UN mediation "is insignificant and is negative," (2) observer missions have an effect that is "not large or significant," (3) UN enforcement "is not significant," (4) traditional peacekeeping "is not at all significant . . . and even has a negative sign." In contrast, multidimensional peacekeeping operations "are extremely significant and positively associated with strict peacebuilding . . . (Notice the high odds ratio of multidimensional peacekeeping operations in Model A8, Table 3)" (Doyle and Sambanis 2000:791).[5] Thus, although they write now that "we *never* offered Model A8 as the only evidence in support of our conclusions," it was the only model supportive of their hypothesis, and so any effect of the conglomerate variable is but a reflection of the effect of this one component. So we chose that model for our article and focus on it here.

Statistical Claims

With our methods, we found that the counterfactuals from Model A8 were far from the data. This implied that there exists plausible alternatives to Model A8 that are not ruled out by existing theory, fit in-sample data approximately the same, but give very different predictions for the same counterfactuals and which imply different substantive results. Our methods do not say which alternative specifications these are, but they were easy to find, which is a clear confirmation of the value of our approach. (The specific alternative model we chose added to A8 an interaction between UN intervention and war duration, a model consistent with their self-described "interactive" peacebuilding theory.)

We showed that the modified and the original models could not be distinguished on the basis of model fit. S&D try to argue their model fits better by only reporting evidence maximally in favor of it. For example, they ran a procedure in Stata that calculated *fourteen* measures of fit, from which they report only the *single* measure that gave the strongest support to their preferred hypothesis and suppressed the rest, the majority of which do not support their argument. (The authors of the Stata command they use emphasize that "there is no convincing evidence that selecting a model that maximizes the value of a given measure of fit results in a model that is optimal in any [relevant] sense" (Long and Freese 2006:104). In fact, the bulk of the evidence shows that the fit of neither model is unambiguously better. Similarly, they ignore the .001 p-value on the standard t-test on the interaction term in the modified model, and instead "drop the clustering" to run a likelihood-ratio test to claim that the interaction is "non-significant." They also ignored the fact that five of

[5] The tests they ran compared each component to all others, including no UN action, but even if they had correctly compared each component to no UN action (only), the conclusion would be the same: of all the components, only multidimensional peacekeeping meets their threshold for importance or statistical significance. (The only other measure of UN interventions they found significant in the 18 models they ran is an ordinal variable, constructed from the components. Unfortunately, the ordinal ranking is logically inconsistent given the findings about the effects of its components. Using it as a continuous measure, as they do, confounds the problem further.)

the seven cases of UN intervention fit better under the modified model. Basing conclusions on all evidence, whether or not it supports your case, is not "flip-flopping" (a term they take from political campaigns applied to politicians expected to take the same position no matter what); it is the only appropriate scientific procedure. In this case, the evidence indicates that there does not exist sufficient evidence to choose between the two models.

Second, S&D argue that "the two models do not produce diametrically opposed policy implications." Consider the crucial issue of determining for what type of wars UN interventions would have maximum effect on postwar peacebuilding. Under the original model, the UN has the largest effect in shorter duration wars, a pattern statistically significant for all wars. Under the modified model, the UN is maximum effective for wars lasting 10 years, and for short wars (where the original model says the UN is most effective), the effects are not statistically significant. It is difficult to imagine how these divergent policy implications (all appearing in S&D's Figure 1, although without the confidence interval for the original model) could reasonably be described as unimportant.[6]

Third, S&D argue that "the logit specification inherently estimates interaction effects." Nagler (1991) showed this claim false long ago. Indeed, in S&D's Figure 1, the slight variability due to the logit form in the estimated effects from the original model has nothing to do with the evidence in the data; if the two variables had a stronger, weaker, or no interaction, that line would remain unchanged. The proper way to estimate an interaction effect in logit models, as shown by Nagler, is to include the interaction, an example of which is the modified model we ran, which appears in the same graph.

Fourth, in their Figure 2, S&D apply our procedure to a different explanatory variable that we did not examine and find less model dependence than for multidimensional UN peacekeeping (as shown in our Figure 4). The only alternative model they examined was the interaction with control variables, and when they found less model dependence than we did, they concluded that our tests are invalid. This is fallacious reasoning. Testing only one alternative model does not evaluate the conclusion of our test, which predicts that *some* alternative model, not necessarily one with specific interaction terms, will give nonrobust results. In addition, their Figure 2 does offer examples of considerable model dependence, the most dramatic of which is in the graph in the middle of their figure: where the original model (on the vertical axis) gives a prediction of peacebuilding success near .5, the modified model gives predictions that range all the way from about .05 to .95.

Fifth, they complain that we chose an extreme counterfactual where the UN intervenes simultaneously in all the civil wars. As explained under the fourth point in "Schrodt on Dinos," we do not do this, but rather evaluate single counterfactuals, each corresponding exactly to the observed data, except for the single value of the UN variable in the one civil war of interest. The counterfactuals we evaluated involved changing the treatment value and leaving the control variables at their real observed values. S&D created other counterfactuals unnecessarily far from the data by changing both the treatment *and* some of the control variables to hypothetical values. They set some control variables to their observed values, some to their means, and others to their 25th and 75th percentiles. Unfortunately, no civil war has control variable values equal to the chosen hypothetical values. Indeed none of their counterfactual controls is even within the convex hull of the observed controls. In these practices, they missed an important point Morrow highlighted:

[6] They also argue (in footnote 12) that the two models do not differ much as the correlation between the model predictions "is large and positive at 68%." To understand how irrelevant this correlation is, consider one model predicting probabilities of .5 for the first half of the data, .51 for the second half; another model predicting .01 for the first half and .99 for the last half. These drastically different predictions will have a correlation of exactly 1.

"I particularly like the subtlety of the observation that even when all the values of the individual variables occur in the data, it may not be the case that all combinations of those values do."

Finally, their comment puts forward random selection as a magic cure-all procedure for eliminating risks to inference and counterfactuals that result from switching random treatment assignments as evidence of the absence of extrapolation. Both claims are wrong. The benefit of randomization is realized in any given sample only as the sample size grows large. For a simple example, consider 10 subjects, differing greatly from one another on the explanatory variables, so that no two are comparable. With these data, no random assignment would lead to comparability for reliable inference. Thus, randomization per se does not provide any guarantee for counterfactuals to be close to the data, and should not be used as a standard for counterfactual evaluation. Generally, the smaller the sample size, the more the variables, and the more sparse the data, the more difficult inferences will be due to extrapolation, with randomization or not. But sufficient experimental control enables one to generate even small data sets that are well suited to making inferences about given counterfactuals. All randomized data are not created equal; all data sets of the same size are not equally informative. Our tests can be used to sort out the good from the bad.

To shore up their fallacious assertions, S&D claimed to reanalyze a "good dataset with experimental data" that has "more than 1,000 observations" from Hiscox (2004). Unfortunately, the procedures described in their paper were not the ones they used. Instead of using Hiscox's full experimental data (a sample of 942 observations), as they claim, they used a nonrandom subsample of 562 observations, and then entirely made up and added three variables of their own not in Hiscox's data, without discussing it in the paper.[7] When we replicated what S&D *claimed* to have done, using Hiscox's original data, we found 98.7 percent of the counterfactuals to be inside the hull. Even a random sample of only 122 points from these data has 78.7 percent of the counterfactuals inside the hull. Indeed, even a random sample of just 22 points from these data has 50 percent of the counterfactuals inside. Thus, not only do our tests indicate when sufficient information exists to evaluate counterfactuals in observational data, they can also be used to evaluate experimental data. Hiscox's original experimental data set was good; the version S&D constructed was not.

Mathematical Claims

S&D make the general claim that our "diagnostic tools perform badly in small datasets," and that in data sets like theirs "with [122 observations and 11 variables], no query point will be inside the convex hull" (S&D supplement., p. 126). This is a strong claim—one that would be a stunning development given the extensive literature on convex hulls—but constructing many counterexamples is easy, so the claim is false. Each counterexample takes the form of a data set the size of their data (or smaller) with 122 observations, one treatment variable, and 10 control variables but, unlike their data, has as much as 100 percent of counterfactuals that fall within the convex hull.[8] This proves that the convex hull check results reflect properties of

[7] Hiscox also ran an experiment with a different treatment. Data from the two experiments total more than 1,000 observations but cannot be meaningfully combined for estimating treatment effects.

[8] For a simple counterexample, consider 61 observations set to zeros for all variables except the first three of the UN (treatment) variable, which are ones, and 61 more observations for which control variables are ones and the UN variable is one for the first four observations and zero for the rest. Thus, the UN variable has seven ones and 115 zeros, just as in S&D's data. The 122 counterfactuals that result from switching values of the UN variable (one at a time) are all inside the convex hull. Indeed, even if $n = 10$, all counterfactuals would still fall in the hull. We can also use randomly generated data to produce as many counterexamples as desired, all with 100 percent of the counterfactuals within the hull.

data, and are not solely determined by data set size, as S&D claim. The fact that our tests are able to separate "good" small data sets (the $n = 122$ random subsample of Hiscox's experimental data being a real example) from "bad" ones proves the utility of our tests for small data as well. Thus, S&D's claim that our program "does not produce meaningful results" for small data sets like theirs is wrong.[9]

Second, S&D consider our Gower distance test and claim to "derive a mechanical upper bound on the average number of data points that are 'close' to the counterfactuals" as a function of "the relative number of binary treatments." Their supplement claims more specifically to "prove" that for any data set the same size as their's with a binary treatment, a counterfactual (created by switching the binary treatment values) cannot be close to more than 11 percent of the observations (p. 133). They use this claimed mathematical result to argue that our geometric variability threshold for closeness is unreasonable, and that the average of 1.3 percent of observations we found close to their counterfactuals is not that small and that our methods are not reliable. To prove that this second mathematical claim is false requires only one counterexample. The simple example data set given in footnote 8 is sufficient, as every counterfactual in those data has 50 percent of the data close by. Indeed, it is easy to construct any number of counterexamples.[10]

That these last two claims are false should not be a surprise. The convex hull concept is not new; our innovation is an algorithm enabling scholars to apply this venerable concept to data with as many variables as political scientists often have. Similarly, although they have not been used for the purpose we propose, the Gower distance was originally defined over three decades ago and the geometric variability over a decade ago, and they are known not to have the properties S&D attribute to them.

Remaining Issues

We consider two remaining issues. First, S&D claim that "the easiest way to satisfy King and Zeng's convex hull test is to drop variables from the model, which will bring more counterfactuals into the hull." All our tests, and all counterfactual specifications, are conditional on chosen covariates. If the variables are wrong, the counterfactuals and our tests can be wrong. Such is also the case with all estimates from regression-like models. Dropping control variables, and inducing omitted variable bias, is not advisable. The only valid way to drop relevant control variables would be through procedures that make their values constant in the design phase, such as in blocked experimental designs.

Second, S&D sometimes argue that their results should be taken as valid despite insufficient evidence or evidence to the contrary. For example, they write "Even if a model fails a specification test, this does not invalidate an empirical conclusion and it certainly does not invalidate a model . . .." Or, "Before we can reject an empirical result, we need to design a more flexible approach . . .." Similarly, they also argue that because an alternative model "cannot be estimated and parameter estimates are incredibly distorted" the model must be wrong, but although inestimability

[9] A related claim, that "every statistical method applied to multidimensional space will rely to some degree on extrapolation," is also wrong, as some counterfactuals do not extrapolate. They also claim "correlation among the explanatory variables alone can take the 'counterfactual' outside the hull," which is as it should be, as holding constant one variable while changing a highly correlated one is unrealistic.

[10] Distance distributions are determined by the data generating processes and vary across data sets. Some data sets may see factual data points having little other data nearby. But in any given data, "good" counterfactuals should have an amount of data nearby that are not too far from that for the factuals. In our simple example given above, the 50 percent of data near the counterfactuals is precisely the same as that for the factuals. In contrast, counterfactuals far from the data will have much less data nearby than the factuals. Counterfactuals from Doyle and Sambanis (2000) that we studied have only 1.3 percent of the data nearby; their observed data points have on average more than 12 times as much data nearby.

makes things difficult for the investigator, it must be recognized as a failing of the data. S&D explain, "We stand by our conclusions" because "we collected the data, quantified difficult concepts, learned about the cases, visited the countries, talked to the policy makers, learned about the structure and politics of the United Nations, and then applied currently available quantitative methods to analyze our question." Every point in this list is an excellent reason to stand by a set of conclusions, except the last. Using the best methods available in 2000 was a good reason to conclude in 2000 that they did everything possible. But the best methods for assessing model dependence in 2007 suggest starkly different conclusions about the value of these quantitative data. We take no position on whether this new information should outweigh their qualitative evidence, but the burden of proof must remain with the investigator.

Concluding Remarks

Valid inference about counterfactuals is essential for the key goals of social science, including prediction, answering "what if" questions, and estimating causal effects. When counterfactuals are posed too far from available data and lead to high degrees of model dependence, standard uncertainty measures such as standard errors and confidence intervals can often be massively underestimated. With model dependence and analysts choosing a particular estimator based on any informal rule they desire, formal statistical properties such as unbiasedness and consistency cannot even be uniquely defined, much less proved. Whereas previous methods to assess model dependence usually required specifying and estimating many alternative models, and are thus subject to the criticism that some relevant alternative specifications were missed, the checks we offer are not dependent on such choices and so are valid for the set of all alternative models.

We thank our commenters and editors for the valuable opportunity to clarify the results in our article, and to discuss many important related issues that have arisen. With the methods we propose and accompanying easy-to-use software, and given the impressive replication requirements of ISQ and other leading journals (King 1995; Gleditsch et al. 2003), researchers should now be able to more readily identify model dependence to improve their own work, to reanalyze data from existing articles and reevaluate statistical results and conclusions, and to work together to build new and more reliable social science scholarship we all wish to advance.

References

Doyle, Sir Arthur Conan. (1888) *A Study in Scarlet*, Adamant Media Corporation.

Doyle, Michael W., and Nicholas Sambanis. (2000) International Peacebuilding. *American Political Science Review* 94(4):779–801.

Einstein, Albert. (1920) *Relativity: The Special and General Theory*. New York: Henry Holt.

Gleditsch, Nils Petter, Patrick James, James Lee Ray, and Bruce Russett. (2003) Editors' Joint Statement: Minimum Replication Standards for International Relations Journals. *International Studies Perspectives* 4:105.

Gower, J. C. (1971) A General Coefficient of Similarity and Some of Its Properties. *Biometrics* 27:857–872.

Hau, Lene V., Z. Dutton, C. H. Behroozi, and S. E. Harris. (1999) Light Speed Reduction to 17 Metres Per Second in an Ultracold Atomic Gas. *Nature* 397(6720):594–598.

Heckman, James J., Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. (1998) Characterizing Selection Bias Using Experimental Data. *Econometrika* 66(5):1017–1098.

Hiscox, Michael J. (2004) Through a Glass and Darkly: Attitudes Toward International Trade and the Curious Effects of Issue Framing. Available at http://www.experimentcentral.org/data/data.php?pid=136.

King, Gary. (1995) Replication, Replication. *PS: Political Science and Politics* 28(3):443–499. Available at http://gking.harvard.edu/files/abs/replication-abs.shtml.

KING, GARY, AND LANGCHE ZENG. (2006a) The Dangers of Extreme Counterfactuals. *Political Analysis* 14(2):131–159. Available at http://gking.harvard.edu/files/abs/counterft-abs.shtml.

KING, GARY, AND LANGCHE ZENG. (2006b). Replication Data Set for 'Detecting Model Dependence in Statistical Inference: A Response.' Available at http://id.thedata.org/hdl%3A1902.1%2FFGSRBXXIYT hdl:1902.1/FGSRBXXIYT UNF:3:K4/CgnMYDMV6izc5RVOZTA== Murray Research Archive [distributor(DDI)].

KING, GARY, AND LANGCHE ZENG. (2006c) Replication Data Set for 'When Can History Be Our Guide? The Pitfalls of Counterfactual Inference.' Available at http://id.thedata.org/hdl%3A1902.1%2FDXRXCFAWPK hdl:1902.1/DXRXCFAWPK UNF:3:DaYlT6QSX9r0D50ye+tXpA== Murray Research Archive [distributor(DDI)].

LONG, J. SCOTT, AND JEREMY FREESE. (2006) *Regression Models for Categorical Dependent Variables Using Stata*. College Station: Stata Press.

NAGLER, JONATHAN. (1991) The Effect of Registration Laws and Education on U.S. Voter Turnout. *American Political Science Review* 85(4):1393–1405.

SIGNORINO, CURTIS. (1999) Strategic Interaction and the Statistical Analysis of International Conflict. *American Political Science Review* 93(2):279–298.

STOLL, HEATHER, GARY KING, AND LANGCHE ZENG. (2006) WhatIf: Software for Evaluating Counterfactuals. *Journal of Statistical Software* 15(4). Available at http://gking.harvard.edu/whatif/.