Numerical Stability vs. Specification Uncertainty: A Concluding Comment on Cho and Gaines, and Burden and Kimball[1]

Gary King[2]

January 2, 2001

[1]My thanks goes to Neal Beck, Barry Burden, Lee Epstein, Mo Fiorina, and David Kimball for helpful comments; Wendy Cho and Brian Gains for data; and the National Science Foundation (SBR-9729884, SBR-9753126, and IIS-9874747), the National Institutes of Aging (P01 AG17625-01), and the World Health Organization for research support.

[2]Professor of Government, Harvard University; and Senior Advisor, Global Programme on Evidence for Health Policy, World Health Organization (Center for Basic Research in the Social Sciences, 34 Kirkland Street, Harvard University, Cambridge MA 02138; Internet Keyword: Gary King, (`http://GKing.Harvard.Edu`), `King@Harvard.Edu`, phone (617) 495-2027, eFax (520) 832-7022.

My given task in this symposium is to comment on and clarify this interesting controversy, which features Burden and Kimball's original article (1998; hereinafter BK1), Cho and Gaines' critique (2001; hereinafter CG), and Burden and Kimball's response (2001; hereinafter BK2).

BK1's goal is "to produce more accurate estimates of split-ticket voting within districts and states" than existed before. With these estimates in hand, they develop a candidate-centered perspective on the causes of split-ticket voting, one that complements Fiorina's (1996) celebrated voter-centered "balancing" theory. Demonstrating that the configuration of candidates powerfully affects split-ticket voting patterns is an important contribution to the scholarly literature. However, like many good ideas, once they are clearly stated and evidence is marshaled, BK1's is intuitive enough to be relatively uncontroversial, in this case sufficiently so that even Fiorina agrees with it.

CG's point is not that BK1's candidate-centered theory of split-ticket voting is wrong or any substantive point should be modified. CG also offer no new methodological contributions. Instead, CG's attack appears aimed mainly at the credibility of Burden and Kimball, and their data, methods, and analyses. This may be an unusual tack for a scholarly article in a substantive journal to take but since CG's work seems meticulous, it is presumably worth examining in some detail.

Of particular relevance is a set of eccological inference methods that I (King, 1997) developed and that both pairs of authors use. These methods have come to be known as "EI," the name of one of the two computer programs that implement them (the programs EI and EzI are freely available at `http://GKing.Harvard.Edu`).

I begin with a discussion of why EI was used by BK1, CG, and BK2, and why it is essential to analyses of split-ticket voting. The next section then reports on an analysis that confirms BK2's claim about the difference between what CG say they do and what they actually did. CG claimed to be studying the *numerical stability* of EI but were actually analyzing the *specification uncertainty* of BK1's analysis. They do have a lot to say about specification uncertainty in other parts of their paper, but this mistake unfortunately eliminates the primary contribution CG intended to make. The next section then demonstrates that in BK1's application the numerical stability of EI is exactly as theory dictates. Finally, I address some unrelated methodological issues.

1 Why use EI?

BK1 needed estimates of district-level quantities such as the fraction of citizens who voted for both the Republican presidential candidate and the Democratic congressional candidate. The secret ballot prevents us from knowing the answer exactly, and other methods are not up to the task. For example, owing to the necessity of confidentiality and small numbers of respondents per district, reliable survey data are usually unavailable at this local level; even if they were, they would be prohibitively expensive to collect. Indeed, even the best voting survey available, the National Election Study, yields biased estimates of congressional voting, with biases correlated with incumbency, party, and winner (see Wright, 1993, and the citations therein); voter turnout is biased even more dramatically (Burden, 2000, and the citations therein). No survey has even been shown to produce valid estimates of split-ticket voting for cases like CG's Vermont example (which is of course extremely unrepresentative of the rest of the data). By the same token, simple methods based on the difference between the congressional and presidential vote are well-known to be biased, strictly underestimating the degree of split-ticket voting (Burnham, 1970; Rusk, 1970). Finally, some had attempted to estimate split-ticket voting using Goodman's (1953) regression, the only method of ecological inference widely used in applications during the last half-century before the development of EI, but this method frequently produces impossible estimates (e.g., over 100% split-ticket voters) and ignores a great deal of information in aggregate data about individual behavior.

EI is a method of ecological inference that begins with Goodman's widely used statistical approach but adds information known to be true. What is known as "the method of bounds" formalizes the information missed by Goodman's approach. For example, if 20,000 people vote for the Republican candidate for president, and 5,000 vote for the Democratic candidate for Congress in a district, then the number voting for *both* the Republican presidential candidate and the Democratic congressional candidate must be somewhere between 0 and 5,000. Goodman's regression ignores this obvious but critically informative fact; EI makes use of it — for the entire country and for each and every congressional district in the analysis. In some applications, this information has an enormous effect and is extremely valuable; in others, it adds only marginally. In no real case does it make inferences worse. (Scholars have conducted a great deal of research in ecologi-

cal inference since the development of EI, but all new models build on this innovation of including all available statistical and deterministic information.)

In BK1's application, about half the ecological inference problem is eliminated without any uncertainty, by the uncontroversial use of the bounds that results from moving to EI from Goodman's regression. This is the central reason for the current popularity of EI in applications in political science, in other academic fields where ecological inferences are necessary, and in public policy, judicial litigation, and business applications. EI is also widely used because it provides estimates for every district, whereas Goodman's provides only a single average estimate of split-ticket voting for the entire nation.

2 What did Cho and Gaines do?

Upon publication of their article, BK1 deposited a replication data set with the ICPSR, following the replication standard. I downloaded their data and was able to replicate their analysis without much trouble. CG's article was based on BK1's data, but, despite lengthy attempts, I was not able to replicate CG's results using only BK1's data and CG's paper.

Fortunately, EI analyses produce a special form of output — called a "replication data buffer" — that makes replication easy. A replication data buffer is a single file that includes all input data and information necessary to replicate an analysis. This file, automatically created by EI, includes the original data set, every option, recode, and statistical choice made, and the statistical results. With the program EI or EzI, a researcher can examine any of this information or, with a single command, replicate the entire analysis.

In order that it would be possible to find out what CG did, the editor "froze" the contents of CG for publication and required Cho and Gaines to provide these replication data buffers; it was then easy to ascertain what they had done and to replicate their results. As it turns out, their paper is an inaccurate report of their research.

Rerunning the same model with identical data multiple times is one way to study one particular aspect of a software program's numerical stability. CG claim to do this and, on the basis of what they thought they saw, conclude that EI "will output significantly different results for the exact same data" and so is "unstable," "highly suspect," "unhelpful," and so forth. As it turns out, and as BK2 correctly document, however, CG did not run the same program with the same data and options 10 times as CG claimed. Instead, they

ran 10 different analyses, each with different program and statistical options, variables, and data subsets. As with any statistical technique, if the investigators change the model, the input data, and the statistical options and assumptions, they will get different answers, and so what CG found, given how they actually proceeded, should come as no surprise.

Although the data buffers demonstrate that CG's evidence has nothing to do with EI's numerical stability, their evidence happens to be informative with regard to specification uncertainty — the degree to which substantive conclusions depend on perhaps minor specification decisions. CG do not say so, but most of their empirical results are indeed about specification uncertainty. It is a tribute to the robustness of BK1's conclusions that CG's tinkering with BK1's specification revealed no important substantive changes.

Since CG's article includes considerable discussion of numerical stability, but no analyses that bear on this issue, I now perform an analysis of the numerical stability of EI as applied to BK1's data and analyses.

3 The Numerical Stability of EI

Numerical stability is like plumbing. Its probably not something most substantively oriented political scientists want to spend their life theorizing about, but in the rare instances when something goes wrong nothing is more pressing.

Using digital computers to perform numerical calculations always involves issues of numerical accuracy, precision, and stability. Numbers are represented in computers in binary notation, but the binary representation of the simple decimal number 0.1, for example, requires an infinite number of binary digits, which is obviously more than any real computer can handle. In practice, computers truncate or round when they run into problems like this. Usually this yields a difference so small that we do not notice it. For occasional applications, however, numerical problems can be substantial, especially for those statistical procedures that are iterative, involve many random numbers, or are otherwise computationally intensive or algorithmically complicated.

One key numerical issue in EI, along with many of the statistical techniques that have been developed in the past decade, is that it uses *statistical simulation* to make computations. The idea is simple but very powerful (as explained in Tanner, 1996, and King, Tomz, and Wittenberg, 2000): Often computing a quantity of interest, such as the

mean μ of a probability density $P(y)$, is difficult (i.e., it requires integrals such as $\mu = \int yP(y)dy$ that may range from annoying to impossible), but drawing random numbers from $P(y)$ is easy. In that situation, we can employ the theory underlying survey sampling to approximate μ: take m random draws from $P(y)$ and compute the average. If m were infinite, we could compute μ exactly. In practice, we choose m to be large enough to give as many digits of accuracy as needed for our application.

Most quantities of interest computed by EI use statistical simulation, and so some important questions are whether the approximation error is correct and whether uncertainty in the entire EI algorithm, including parts that do not use random numbers, is as theory dictates. If sources of approximation uncertainty exist for which the statistical or numerical theory cannot account, or the random number generator does not produce sufficiently random numbers, or rounding, truncation, or other numerical accuracies have a big enough effect, then we would need to take specially developed steps.

To study these issues, I reran each of BK1's analyses 1,000 times for each of several values of m (10, 30, 50, 75, 100, 150, ... , 500). For each value of m and each quantity of interest in BK1, I computed the mean and standard deviation across the 1,000 simulations and compared it to the value theory dictates it should be.[1]

Figure 1 depicts these results on numerical stability and simulation variance for two of BK1's empirical analyses. The vertical axis of each graph is the standard deviation across the 1,000 simulations. (The means were in all cases the same as those reported in BK2 and so are not reported here.) The horizontal axis is the number of simulations, m. The theoretical relationship that is supposed to exist between the standard deviation and m appears as a solid line: the actual results of the 1,000 simulations appears as open circles. Clearly the theoretical line is very closely approximated by the empirically based circles, thus confirming this aspect of the numerical stability of EI in the context of these two of BK1's applications.

I ran the same analysis for many other quantities reported in BK1, with the resulting analogous graphs revealing identical relationships and the same high level of numerical

[1]Let $\tilde{\beta}_j$ be simulation j of a quantity of interest. When EI summarizes the posterior distribution with the mean, it approximates the theoretically correct quantity, that is impossible to compute directly, with the average of m simulations, $\hat{\beta} = \sum_{j=1}^{m} \tilde{\beta}_j/m$. Since each simulation is supposed to be an independent random draw from the same density, the variance of the reported mean should be $V(\hat{\beta}) \propto 1/m$. The vertical axis of Figure 1 is the standard deviation of the average, the square root of $V(\hat{\beta})$, and the solid line plots this relationship between this standard deviation and m, which is $\sqrt{V(\hat{\beta})} \propto 1/\sqrt{m}$.

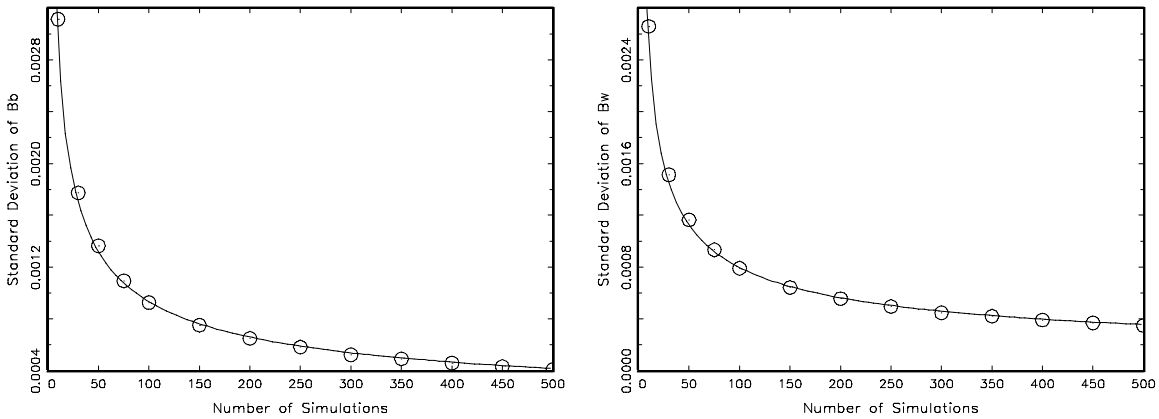

Figure 1: *Numerical Stability in EI: Each graph evaluates a different quantity of interest in BK1's work. The vertical axis is the standard deviation over 1,000 replications of the same analysis; and the number of simulations is the horizontal axis. The solid line is the theoretical relationship, and the circles are the results of the analysis.*

stability. Analyses of other quantities, such as the log-likelihood value, standard errors, and different features of the posterior distribution, lend themselves to the same conclusion.

4 Other Methodological Issues

CG also include much material intended to be helpful pedagogy but turns out to be incorrect. I discuss a few of these and related points here. Most of the rest are discussed in BK1.

First, CG say that "EI claimed to have computed a Hessian matrix that was not positive definite." and then state emphatically that "A Hessian matrix *must* be positive definite." (p.22). Although we might prefer Hessian matrices to be positive definite, nothing requires them to be such. Indeed, anyone who has ever seen a computer complain of "multicollinearity" in least squares regression will know one of the causes, and have seen the consequences, of a Hessian that is not positive definite. When CG ran EI and found Hessians that were not positive definite, they should have reexamined, or respecified, their model or otherwise explored the consequences. What they did do was inappropriate.

Second, CG explain that they "have a prior that aggregation bias exists in the data set" (p.12). They do not explain where they get these beliefs, but whenever one has such priors they need to be included in the analysis. EI includes facilities to specify such priors. Unfortunately, as revealed in their replication data buffers, CG chose EI options

that ignored this information, thus not following what they said they did, the EI manual, or one of the most basic basic tenets of Bayesian analysis.

Third, CG write that "Producing out-of-bounds estimates is thus a very useful feature of the linear probability model" (p.12). In fact, out-of-bounds estimates are known to be incorrect and so, although these estimates from the linear model may be a useful diagnostic tool, the model is clearly not useful for providing accurate estimates.

Fourth, In Figure 5, CG present two tomography plots from two runs of the second stage of a 2×3 analysis. They point out that the two plots are almost the same, but "close inspection reveals many discrepancies between these supposedly identical tomography plots," pointing especially to a few lines in one that are "curiously absent" from the other. What CG missed is that a tomography plot for their 2×3 analysis would normally require a four-dimensional display, not the two-dimensional plot they present. As the EI manual explains, this two-dimensional projection is offered as an easily interpretable alternative to (somehow!) interpreting multiple hyper-planes intersecting a four-dimensional hypercube. The advantages of this approach are obvious, but they are only realized if the methodology is understood. The only sacrifice from this more easily interpretable approach offered by EI is the trivial differences due to simulation across multiple graphs, as illustrated by their Figure 5.

Finally, CG worry about how to choose covariates, always an important concern. Unlike CG, BK1 base their decisions on substantive knowledge and graphical diagnostics, which is normally the best first approach. CG are more interested in formal statistical tests. When criticizing BK1, they find that the diagnostics reveal sufficient information about correct covariate choice (e.g., "Unfortunately, this covariate did not perform the necessary function of removing aggregation bias, and their analyses suffered accordingly." p.16), but when criticizing the methodology they argue the opposite and conclude that "a formal method is needed," citing Cho's previous calls for the same thing. I agree that new formal specification tests might be useful additions to substantive knowledge and existing graphical diagnostics, but CG seem to have missed what would presumably be the place to start, the only efficient formal specification tests offered in the literature (see King, 1997, and King, Rosen, and Tanner, 1999).

5 Concluding Remark

The numerical instability that CG reported that they they saw in BK1's EI results was in fact induced by CG's changing the model specification for each of their runs contrary to the claims in their paper. Since a proper evaluation that repeats the same analysis multiple times reveals no meaningful numerical instability of this type in BK1's application of EI, CG's analysis provides no reason to reject the central substantive findings reported in BK1.[2]

Burden and Kimball's article is not perfect (who's is?), but the improvements that might be made, and the important research questions they raised, will no doubt stimulate scholars of American politics to try to build on their work for some time. However, nothing in Cho and Gaines' paper should cause readers to alter any of Burden and Kimball's methodological choices or substantive conclusions.

References

Altman, Micah and Michael McDonald. 1999. "The Robustness of Statistical Abstractions: A Look Under the Hood," presented at the annual meetings of the Society For Political Methodology, College Station, Texas.

Burden, Barry C. 2000. "Voter Turnout and the National Election Studies," *Political Analysis* 8 (Autumn): 389–98.

Burden, Barry C. and David C. Kimball. 2001. "***," *American Political Science Review*, forthcoming.

Burden, Barry C. and David C. Kimball. 1998. "A New Approach to the Study of Ticket Splitting," *American Political Science Review*, 92, 3 (September): 533–544.

Burnham, Walter Dean. 1970. *Critical Elections and the Mainsprings of American Politics*, New York: Norton.

[2]Numerical stability is a much broader concept than the specific aspect of it discussed here. More detailed analyses of different aspects of EI's numerical properties have been conducted by me and others, the most comprehensive of which appear in the ongoing more general work of Micah Altman, Jeff Gill, and Mike McDonald (see Altman and McDonald, 1999). These studies have identified some numerical issues that have been used to improve EI over the years. As detailed in EI's documentation, most stem from the procedure that computes the volume above the unit square for the bivariate normal density. If anyone discovers any other ways to improve EI, I would be grateful to hear about them or to include improvements or alternative methods suggested in the program.

Cho, Wendy K. Tam and Brian J. Gaines. 2001. "Reassessing the Study of Split-Ticket Voting," *American Political Science Review*, forthcoming.

Fiorina, Morris P. 1996. *Divided Government*. 2nd ed. Needham, MA: Allyn & Bacon.

Goodman, Leo. 1953a. "Ecological Regressions and the Behavior of Individuals," *American Sociological Review*, 18: 663–666.

King, Gary. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*, Princeton: Princeton University Press.

King, Gary; Ori Rosen; and Martin Tanner. 1999. "Binomial-Beta Hierarchical Models for Ecological Inference," *Sociological Methods and Research*, 28, 1 (August): 61–90.

King, Gary; Michael Tomz; and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation," *American Journal of Political Science* 44, 2: 341–355.

Rusk, Jerrold G. 1970. "The Effect of the Australian Ballot Reform on Split Ticket Voting: 1876–1908," *American Political Science Review*, 64 (December): 1220–38.

Tanner, M. A. 1996. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, third edition, New York: Springer-Verlag.

Wright, Gerald C. 1993. "Errors in Measuring Vote Choice in the National Election Studies, 1952-88." *American Journal of Political Science* 39 (February): 291–316.