Advance Access publication August 21, 2007

Political Analysis (2008) 16:197–212
doi:10.1093/pan/mpm021

Comparing Legislators and Legislatures: The Dynamics of Legislative Gridlock Reconsidered

Fang-Yi Chiou
Institute of Political Science, Academia Sinica, 128 Academia Road, Taipei, Taiwan
e-mail: fchiou@gate.sinica.edu.tw

Lawrence S. Rothenberg
Department of Political Science, University of Rochester, Rochester, NY 14627
e-mail: lrot@mail.rochester.edu (corresponding author)

Although political methodologists are well aware of measurement issues and the problems that can be created, such concerns are not always front and center when we are doing substantive research. Here, we show how choices in measuring legislative preferences have influenced our understanding of what determines legislative outputs. Specifically, we replicate and extend Binder's highly influential analysis (Binder, Sarah A. 1999. The dynamics of legislative gridlock, 1947–96. *American Political Science Review* 93:519–33; see also Binder, Sarah A. 2003. *Stalemate: Causes and consequences of legislative gridlock*. Washington, DC: Brookings Institution) of legislative gridlock, which emphasizes how partisan, electoral, and institutional characteristics generate major legislative initiatives. Binder purports to show that examining the proportion, rather than the absolute number, of key policy proposals passed leads to the inference that these features, rather than divided government, are crucial for explaining gridlock. However, we demonstrate that this finding is undermined by flaws in preference measurement. Binder's results are a function of using W-NOMINATE scores never designed for comparing Senate to House members or for analyzing multiple Congresses jointly. When preferences are more appropriately measured with common space scores (Poole, Keith T. 1998. Recovering a basic space from a set of issue scales. *American Journal of Political Science* 42:964–93), there is no evidence that the factors that she highlights matter.

1 Introduction

Although political methodologists are well aware of measurement issues and the problems that can be created, such concerns are not always front and center when we are doing substantive analytic research. Here, we show how choices in measuring legislative preferences have influenced our understanding of what determines legislative outputs. Specifically, we demonstrate how measurement error can distort inferences by replicating and extending Binder's (1999; see also 2003) highly influential analysis of legislative gridlock. Although her analysis depends on measuring preferences across chambers and over time,

Authors' note: Thanks to Sarah Binder and Keith Poole for furnishing data used in our analysis and to Chris Achen and Kevin Clarke for advice. All errors remain our own. Online appendix is available on the *Political Analysis* Web site.

© The Author 2007. Published by Oxford University Press on behalf of the Society for Political Methodology.
All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

https://doi.org/10.1093/pan/mpm021 Published online by Cambridge University Press

the use of W-NOMINATE instead of common space scores designed for such purposes leads to unsupported conclusions.

Our discussion is especially important because, in many respects, Binder's study is an important step forward. Beginning with Mayhew (1991), a major debate in the scholarly literature on the American Congress has revolved around determining the conditions that induce the institution to produce major legislative initiatives. A variety of factors have differentiated analyses, and results have often proven contradictory.

Much of the work in this stream of research has focused on interbranch conflict, notably on the effects of divided government (e.g., Edwards, Barrett, and Peake 1997; Coleman 1999) or on differences in preferences across branches of government (e.g., Krehbiel 1996, 1998; Chiou and Rothenberg 2003; Brady and Volden 2006). Some have confirmed that interbranch conflict is important, whereas others have concluded that no support exists.

Alternatively, in what is perhaps the seminal piece in this stream of literature, Binder (1999; see also 2003) proposes a theory of legislative gridlock that focuses on partisan, electoral, and institutional sources of disagreement beyond those associated with separation of powers. Her analysis, written in specific contrast to Mayhew's (1991), has two especially noteworthy innovations. One is that she measures legislative gridlock as the proportion of potential major initiatives that do not get passed (in contrast to Mayhew's analysis, which focuses on the sheer number of legislative initiatives enacted). Stated differently, whereas Mayhew, and others who followed him, employs only a "supply-side" measure (which we will call *legislative productivity* in the following analysis), Binder incorporates "demand" features as well by measuring the proportion of important proposals approved (which we will label *legislative gridlock*). The other is that she uses first dimension W-NOMINATE scores to incorporate preferences in the form of four independent variables: percentage of moderates, ideological diversity, bicameral distance, and severity of filibuster threat. Thus, Binder implicitly assumes that W-NOMINATE is appropriate for measuring preferences within and across chambers and over time. Arguing that "the distribution of policy preferences within the parties, between the two chambers, and across Congress more broadly is central to explaining the dynamics of gridlock" (519), she produces empirical results that stand in striking contrast to Mayhew's. Binder's analysis indicates that, beyond divided government, partisan, electoral, and institutional factors are critical for explaining gridlock in American politics. In particular, she stresses that intrabranch conflict, as principally measured by bicameral differences, is key.

However, although impressive in many respects, Binder's analysis is flawed by its reliance on W-NOMINATE scores as the measure of congressional preferences. Although useful for comparing preferences within the same Congress and chamber, W-NOMINATE scores are not intended for comparisons between chambers and across Congresses (Poole and Rosenthal 1997; Poole 1998).[1] Fortunately, Poole's (1998) development of "common space" scores provides a new measurement solution for the task of across-chamber and over-time comparison. When we substitute common space scores for W-NOMINATE scores in the measurement of the four key variables that Binder employs which require a preference measure, we find that none of the partisan, electoral, and institutional features that she highlights are particularly important for explaining her measure of legislative gridlock. Nor, given the use of common space scores, does a model with the same set of independent variables as Binder's do better when we employ alternative measures of

[1]Indeed, one of us has been criticized for comparing members of one chamber (the House) in consecutive Congresses by relying on W-NOMINATE scores (Herron 2004; for a rejoinder, see Rothenberg and Sanders 2004).

https://doi.org/10.1093/pan/mpm021 Published online by Cambridge University Press

the proportion of major legislative initiatives enacted. By contrast, findings for the absolute production of legislative outputs, which showed no evidence of being impacted by the partisan, electoral, and institutional forces that Binder focuses on to begin with, are relatively insensitive to preference measurement.

In the following analysis, we first briefly compare and contrast W-NOMINATE and common space scores. We then summarize and discuss Binder's original research on legislative gridlock, reestimate her analysis by using common space scores instead of W-NOMINATE, and then extend this research by substituting alternative measures of the underlying dependent variable, including both measures that focus on the proportion of major initiatives approved and those that simply look at absolute output. Finally, we show that Binder's subsequent attempt to develop another measure of bicameral difference, which empirically produces results comparable to her original research, provides no additional support due to its own measurement problems.

2 W-NOMINATE versus Common Space Scores

Before turning to the details of Binder's analysis, it is helpful to review some key differences between W-NOMINATE and common space scores. For the analysis at hand, the principal distinction that requires highlighting is that, given the way that they are generated, *W-NOMINATE scores are not directly comparable between Congresses and across chambers* (e.g., Poole and Rosenthal 1997; Poole 1998). W-NOMINATE scores use roll call votes from one chamber in a single Congress to produce a preference estimate for those serving in that chamber in that Congress. As information about voting in the other chamber in the same Congress or in other Congresses is not employed, W-NOMINATE scores have different underlying scales across chambers and Congresses. This means that comparing W-NOMINATE–based measures across chambers and Congress is inappropriate.

Common space scores offer an alternative to W-NOMINATE when we wish to compare preferences across chambers and over time (Poole 1998). Assuming that the first two W-NOMINATE dimensions are equally salient and lie within a unit circle and that the ideological preferences of legislators in both chambers remain fairly stable, common space scores use legislators sequentially serving in both the House and the Senate to provide the glue for comparable scores to be generated from W-NOMINATE.[2] Like W-NOMINATE, the first and second dimensions of common space scores are independent of one another, so the value of first dimension common space scores will not be influenced by the importance of the second dimension. In a nutshell, common space scores adjust

[2]Although a seemingly strong requirement, a number of findings suggest that the assumption that members moving across chambers are ideologically consistent is reasonable. For example, members change their behavior little over the course of their intrachamber careers (Poole 2003), implying that even changes in district composition may have a lesser effect than we might otherwise believe. Also, as most House members moving to the Senate come from small states (Snyder 1992b), implying that a politician's underlying constituency would not change much upon a move from the House to the Senate, many of the members providing the glue that allows common space scores to be estimated will be representing very similar constituencies. Admittedly, these findings that imply that members moving from the lower to the upper chamber should behave in a rather similar manner are suggestive rather than conclusive. Ideally, we would want identical items voted upon in both chambers as the glue for generating comparable scores, but such instances are too rare to exploit effectively (but for such an effort, see Bailey 2005).

We should also note that Binder (2003) questions the interchamber consistency assumption on the grounds that party loyalty scores change when a legislator switches to a different chamber. However, the evidence underlying this claim is weak, as a legislator's loyalty score may be a function of many factors, such as differences between the Senate's and House's internal rules (e.g., regarding amendments) and her ideological location relative to her party's median, rather than differences in the member's ideological behavior per se.

https://doi.org/10.1093/pan/mpm021 Published online by Cambridge University Press

otherwise incomparable W-NOMINATE scores to ensure that the same scale for ideological space is used across chambers and Congresses.

This leads to the central question that motivates our analysis: What happens when common space scores are substituted for W-NOMINATE? Does the close relationship between the way that W-NOMINATE and common space scores are estimated produce similar results in practice or do they change the way that we understand the world? Binder's (1999) analysis offers a welcome opportunity to explore this concern.

3 Binder's Analysis

For clarity sake, it is important to summarize briefly Binder's (1999) conceptual approach and the execution of her original analysis. Conceptually, Binder cautions against the single-minded focus on partisan theories that emphasize divided government. In response, she considers five features in addition to divided government and other measures of policy context. Three—percentage of moderates, ideological diversity, and time since the current legislative majority has had majority status—are designed to capture partisan and electoral influences on gridlock. Two others, severity of the filibuster threat and bicameral distance, are included to measure institutional effects. With this more complete listing of potential sources of gridlock, Binder offers several hypotheses, e.g., that the greater the bicameral distance, the more pronounced gridlock will be.

To test these hypotheses, Binder makes two important measurement choices. First, she measures legislative gridlock for the 80th to 104th Congresses as the proportion of major initiatives that pass (major initiatives being defined by a coding of the editorial pages of the *New York Times*) rather than as the sheer number of key legislative outputs à la Mayhew (1991). Second, she incorporates the first dimension of W-NOMINATE in measuring severity of filibuster threat, percentage of moderates, ideological diversity, and bicameral difference.[3] Severity of filibuster threat is measured by multiplying the number of actual filibusters by the ideological distance between the chamber median and the filibuster pivot who is furthest away from that median; percentage of moderates is measured by averaging, for the Senate and House of each Congress, the percentage of members who fall ideologically between the two party medians in the chamber; ideological diversity is measured by first taking the SD on W-NOMINATE for each member by chamber and then averaging the two chamber scores; and bicameral difference is measured by calculating the ideological distance between the House and the Senate medians for each Congress.

To see if her hypotheses are borne out, Binder estimates a grouped-logit, weighted least-squares model. She also uses ordinary least squares (OLS) to estimate a model with her independent variables and Mayhew's (1991) legislative productivity measure as the dependent variable. Her results differ dramatically by dependent variable. Except for severity of filibuster threat, her new partisan/electoral and institutional variables are statistically significant for legislative gridlock (see our Table 1).[4] When the productivity measure is substituted as the dependent variable, these same variables are essentially irrelevant. In other words, taken jointly, Binder's findings turn Mayhew's on their head and, assuming that a dependent variable incorporating the demand for new legislative initiatives is superior, indicate that the partisan/electoral and institutional features that Binder incorporates are key determinants of deadlock in the American political system.

[3]Time the majority party has been out of the majority is not measured with preferences. Instead, how many immediately previous Congresses, if any, the majority party in each chamber has been in the minority is counted, and the average is taken between the House and the Senate scores.

[4]Since we initially lacked access to Binder's complete data set, we reproduced her results in virtually identical fashion by combining variables furnished to us and the data construction documentation in Binder (1999).

https://doi.org/10.1093/pan/mpm021 Published online by Cambridge University Press

Table 1 Determinants of policy gridlock (coefficients with SEs in parentheses)

Variables	*Grouped-logit—W-NOMINATE measurement (Binder's analysis)*	*Common space measurement*	
		Grouped logit	*Prais-Winsten (robust SEs)*
Divided government	0.356* (0.148)	0.111 (0.205)	0.029 (0.051)
Percentage of moderates	−0.025* (0.013)	−0.025 (0.039)	−0.006 (0.009)
Ideological diversity	6.399* (2.819)	10.088 (25.974)	2.598 (5.862)
Time out of majority	−0.168** (0.047)	−0.044 (0.059)	−0.008 (0.011)
Bicameral distance	2.275** (0.834)	−2.069 (3.050)	−0.543 (0.441)
Severity of filibuster threat	0.035 (0.042)	0.081 (0.162)	0.009 (0.040)
Budgetary situation	−0.006 (0.010)	0.001 (0.019)	−0.001 (0.004)
Public mood (lagged)	−0.032* (0.016)	−0.006 (0.027)	0.000 (0.005)
Constant	−1.707 (1.688)	−2.305 (8.780)	−0.193 (1.866)
Adjusted R^2	0.531	0.061	−0.011

Note. Number of cases = 22; Durbin-Watson = 1.900 (value from original OLS estimation). One-tailed t test, $^*p < .1$; $^{**}p < .05$.

Of particular relevance is intrabranch conflict, as principally captured by the statistical importance of bicameral differences.

As indicated, our principal quarrel with this work involves measuring preferences with W-NOMINATE. The ideal measure that we are seeking is one that accurately captures preferences in a way that is directly comparable across chambers and over time. Despite requiring some strong assumptions of their own, common space scores, which were just becoming available when Binder conducted her analysis, offer a seemingly superior alternative to W-NOMINATE in this regard.

4 Substituting Common Space for W-NOMINATE

As we have seen, besides being an analysis of great substantive importance, Binder's (1999) study is particularly appropriate for our purposes because measuring preferences across chambers and over time is crucial for a variety of her key independent variables. Therefore, we have the opportunity not only to explore whether her results are sensitive to measurement but also to investigate how and why this happens.

As a preliminary, we first recalculate Binder's four measures that incorporate preferences by substituting first dimension common space for W-NOMINATE scores. Figure 1a–1d illustrates the differences between using W-NOMINATE and common space scores.

Overall, which measurement technique is used is clearly consequential. Not surprisingly, given the noncomparability of W-NOMINATE scores and the comparability of common space scores, there is more fluctuation and less observable path dependence for measures using W-NOMINATE rather than common space.

However, the degree to which the measures based on W-NOMINATE and common spaces scores correspond also varies considerably. This appears to be a product of how differently the four preference-based variables are constructed. Some are built in ways that mitigate some of the noncomparability of W-NOMINATE, whereas others are not.

For instance, as seen in Fig. 1a, there is a very close relationship between the W-NOMINATE and the common space measures for the severity of filibuster threat variable (the correlation coefficient is 0.92). This is because this variable's value for any Congress, a product of the number of filibusters in a Congress by median filibuster pivot distance, is

https://doi.org/10.1093/pan/mpm021 Published online by Cambridge University Press

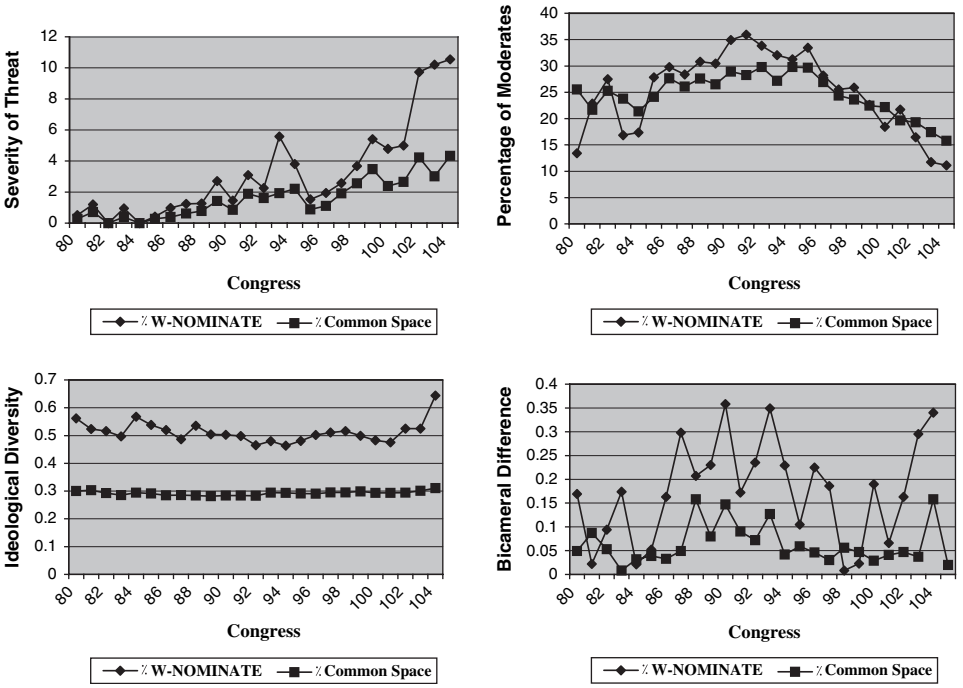


Fig. 1 (a) Severity of filibuster threat: W-NOMINATE versus common space measures. (b) Percentage of moderates: W-NOMINATE versus common space measures. (c) Ideological diversity: W-NOMINATE versus common space measures. (d) Bicameral distance: W-NOMINATE versus common space measures.

only partly determined by ideology scores. In fact, most of the variability in the filibuster threat measure is a function of substantial changes in the number of actual filibusters seen in the Senate over time. As such, W-NOMINATE's comparability issues do not affect the variable scores very much.

Although for somewhat different reasons, there is also a high correlation (0.86) between the W-NOMINATE and common space measures of the percentage of congressional moderates in a Congress (Fig. 1b). This is because this variable only requires measuring the average percentage by chamber that falls between party medians. W-NOMINATE only has to serve as an ordinal measure, lessening the amount of comparability-induced measurement error.

Noncomparability's ramifications are more pronounced when measuring ideological diversity (Fig. 1c). Recall that this variable is based on the average SD of preferences in each chamber. However, W-NOMINATE produces different scales across chambers and Congresses, leading to SDs that are not directly comparable. Even though using the average SD for the two chambers should reduce the cross-chamber noncomparability problem somewhat, considerable measurement error is nonetheless introduced. As a result, the W-NOMINATE diversity measure fluctuates more than its common space analog and the resulting correlation is only 0.56.

Comparability problems are most severe for bicameral differences (Fig. 1d). The reason for this is intuitive: There is nothing in this measure's construction, such as averaging or incorporating a measure not using preference scores, which reduces noncomparability problems. Rather, bicameral difference involves directly calculating ideological

https://doi.org/10.1093/pan/mpm021 Published online by Cambridge University Press

differences between the House and the Senate medians, which requires assuming that each chamber median comes from the same underlying ideological space. All the problems of Congress-to-Congress and chamber-to-chamber noncomparability are realized, creating a measure that fluctuates much more than its common space analog. Therefore, there is only a 0.48 correlation between the two measures.

Given the considerable differences between W-NOMINATE and common space measures, it should not surprise us that when reestimating Binder's analysis using variables with common space scores we get much different results. As Table 1 shows, we find that support for the claim that Binder's partisan/electoral and institutional features are of great importance for explaining gridlock evaporates. Additionally, joint F tests demonstrate that our nonresults are not a function of multicollinearity among these variables. Nor does use of alternative estimation techniques—OLS, White, and Prais-Winsten (results for the latter are shown in Table 1 for illustrative purposes)—yield different findings.[5]

More specifically, there are two striking differences if we compare our results with Binder's (for brevity, we will only rely on our grouped-logit findings). First, all the variables whose coefficients are statistically significant in Binder's analysis are insignificant in our analysis. Although three of the four preference-based variables retain the same sign—the lone exception being bicameral difference where the sign is actually reversed—none remain significant. Second, the model fit declines dramatically, with the adjusted R^2 decreasing from 0.531 to 0.061.

Both discrepancies are consistent with what we know about measurement error. Although perhaps counterintuitive, changes in significance and sign such as those that we observe can occur when measurement error is reduced or eliminated. Whereas measurement error, such as that which we claim is introduced by using W-NOMINATE, would asymptotically bias an independent variable in a bivariate regression toward zero, the opposite can be the case for the kind of multivariate analysis conducted here (Cole 1969; Hodges and Moore 1972; Levi 1973; Rao 1973; DeVaro and Lacker 1995; Greene 2003). Although all estimates in the equation will be biased, the direction and magnitude of these estimates are indeterminate. For example, Cole (1969, 94) summarizes that "the magnitude and direction of bias depend on the correlations between the explanatory variables and on the relative magnitude of the data errors as well as their intercorrelations." Thus, insignificant results, and as a result poor fit, may go along with correcting measurement error.[6]

Even the coefficient for the bicameral difference measure reversing its sign is consistent with established results. For instance, in his analysis of the asymptotic properties of measurement error in a multivariate regression, Achen (1985, 310) finds that "correlated errors among independent variables can lead to inappropriate signs for regression coefficients" and that the coefficient of the variable with the most serious measurement error can have a sign reversal. As using W-NOMINATE to measure our four variables of interest produces correlated measurement errors, and with bicameral difference being the variable most impacted by measurement error, it is consistent that the bicameral difference coefficient is the only coefficient reversing its sign.

Table 2 illustrates this somewhat differently. We take Binder's model and variable measurement except that we "correct" the measurement of one preference-based variable

[5]The only exception is bicameral distance. For this measure, we find, contrary to Binder's prediction, a statistically significant *negative* coefficient when we drop some variables and use Prais-Winsten estimation (we will explore reasons for this result shortly). Our results are also not sensitive when, as does Binder (1999), we substitute alternative measures of public mood and party control and are also qualitatively unchanged when we substitute five alternative gridlock measures based on issue salience that Binder (2003) has subsequently developed.

[6]For an empirical example, see Hashimoto and Kochin's (1980) analysis of economic discrimination.

https://doi.org/10.1093/pan/mpm021 Published online by Cambridge University Press

Table 2 Determinants of policy gridlock—effects of substituting common space for W-NOMINATE measures (grouped-logit coefficients with SEs in parentheses)

	Common space measure substituted			
Variables	*Percentage of moderates*	*Severity of filibuster threat*	*Ideological diversity*	*Bicameral distance*
Divided government	0.313* (0.153)	0.351* (0.154)	0.326* (0.176)	0.096 (0.160)
Percentage of moderates	−0.039 (0.031)	−0.029* (0.011)	−0.034* (0.015)	0.008 (0.022)
Ideological diversity	6.156* (3.203)	6.628* (2.869)	−0.148 (22.544)	5.686 (3.559)
Time out of majority	−0.158** (0.050)	−0.177** (0.051)	−0.097* (0.046)	−0.089 (0.054)
Bicameral distance	2.223* (0.923)	2.512** (0.771)	1.648 (0.940)	−3.753 (3.200)
Severity of filibuster threat	0.039 (0.047)	0.087 (0.110)	0.028 (0.057)	0.105* (0.050)
Budgetary situation	−0.006 (0.017)	−0.003 (0.012)	−0.001 (0.012)	0.002 (0.012)
Public mood (lagged)	−0.038* (0.017)	−0.033* (0.017)	−0.023 (0.022)	−0.022 (0.020)
Constant	−0.870 (2.193)	−1.725 (1.692)	1.442 (7.222)	−2.222 (2.407)
Adjusted R^2	0.470	0.526	0.342	0.327

Note. Number of cases = 22. One-tailed t test, *$p < .1$; **$p < .05$.

by substituting a W-NOMINATE measure with its common space analog. We then do the same thing for another W-NOMINATE measure, and so on. For example, the first column of results replicates Binder's analysis except that a common space measure of percentage of moderates is substituted for the W-NOMINATE measure. The second column measures percentage of moderates, ideological diversity, and bicameral distance with W-NOMINATE but measures severity of filibuster threat with common space. The next two columns are analogous, except that ideological diversity (column three) and bicameral distance (column four) are the only variables measured with common space. As the first two columns of results show, Binder's empirical results remain similar if the only variable measured with common space scores is either percentage of moderates or severity of filibuster threats. As the third column of results demonstrate, change is somewhat greater when the replaced variable is ideological diversity and, as shown in the final column, the results essentially collapse when we use the common space measure of bicameral difference. In other words, even if we measure moderates, diversity, and filibuster threat with W-NOMINATE, the results dramatically change if we correct the measurement of bicameral difference. Consistent with the spirit of Achen's analysis, the variable with the most measurement error, bicameral difference, is the key for producing Binder's results.[7]

Admittedly, we must exercise caution in suggesting how measurement error influences Binder's findings. The theoretical work on the direction and magnitude of biases caused by measurement error looks at asymptotic properties of linear models, and we are dealing with a small sample using nonlinear estimation. Although all of Binder's and our findings are qualitatively unchanged if we use a linear model, we cannot rule out that some of the changes in direction and magnitude of coefficients may be a product of the sensitivity of

[7]We also conduct similar analyses replacing sets of two and three of the variables measured with W-NOMINATE with their common space analogs (results available from authors). The inference drawn remains the same—the variables with more measurement error are principally responsible for the significant results for W-NOMINATE. One difference is that F tests for the joint significance of the four variables in question, which are significant in the models presented in Table 2, tend to be insignificant.

https://doi.org/10.1093/pan/mpm021 Published online by Cambridge University Press

any small sample to changes in variable measurement or to the inclusion or exclusion of certain variables.[8]

The second striking discrepancy between our analysis and Binder's, the dramatic decline in adjusted R^2, arises naturally from the fact that variables are no longer significant. Our adjusted R^2 is essentially equivalent to the model fit one gets from rerunning Binder's analysis without her four preference-based measures. Closer investigation (results not shown) similar to that presented in Table 2 further reveals that changing ideological diversity and bicameral difference measurement from W-NOMINATE to common space has the greatest dampening effect on the model. This is intuitive as these are the two variables that suffer most from measurement error.

To conclude, our analysis suggests that Binder's findings reflect the effects of measurement error. As we have seen, such measurement error impacts all four of the independent variables that Binder measures with W-NOMINATE to one extent or another. The variables measured with the most error, ideological diversity and, especially, bicameral difference, have the biggest effect on findings of statistical significance and the level of model fit. Although we might be surprised that measurement error induces false positives, this is consistent with what econometric theory tells us and may also be a function of small sample size.

5 Alternative Measures of Gridlock and Productivity

Although our results highlight the importance of measurement concerns for Binder's analysis, we wish to go beyond this and examine whether our findings using common space scores hold for alternative measures of legislative gridlock and productivity (recall that Binder finds dramatically different results depending upon whether one examines the proportion or number of key initiatives passed). To do this, we first turn to several competing measures of legislative gridlock; subsequently, we investigate different measures of legislative productivity.

With respect to legislative gridlock, we substitute two alternative measures for the dependent variable that scholars have employed:

(1) Coleman's (1999) measure, which basically divides Mayhew's (1991) enumeration of significant bills by the sum of significant bills passed and failed (regarding failed bills, see Edwards, Barrett, and Peake 1997);[9] and

(2) Chiou and Rothenberg's (2003) measure, which is equivalent to the measure above except that it uses only Mayhew's Sweep One measure of significant enactments (i.e., it uses contemporary journalistic selections but not retrospective evaluations by policy experts).

It is important to note that these measures are empirically quite different from Binder's measure (correlations with Binder's measure are 0.498 and 0.487 for the Coleman and Chiou-Rothenberg measures, respectively).[10] This is not to suggest that the Coleman and Chiou-Rothenberg measures are necessarily better than Binder's. Both the Coleman and the Chiou-Rothenberg measures may be critiqued as potentially endogenous because

[8]For example, if we take Binder's analysis, including her W-NOMINATE measurement but dropping the variables measuring percentage of moderates and ideological diversity, all of her primary findings evaporate.

[9]To make Coleman's measure comparable to Binder's, we use the number of significant bills that pass divided by the number of significant bills that pass and fail.

[10]However, although we include both for the sake of completeness, there is little difference between the latter two alternatives (the correlation is 0.947).

https://doi.org/10.1093/pan/mpm021 Published online by Cambridge University Press

what gets on the agenda, which forms the denominator for each measure, may be a function of legislative characteristics. However, Binder's reliance on *New York Times* editorials may also be endogenous to legislative characteristics; additionally, relying on the *Times* has other potential problems, such as changing editorial policies, the weighting of all editorials equally, fluctuating levels of competition for editorial page space, and idiosyncrasies that distinguish the *Times* from other potential news sources. Regardless, if we find comparable results for these very different measures, then we should have more confidence that our emphasis on the measurement of preferences for understanding legislative gridlock is justified.

As Table 3 shows, the inference that, once we use common space scores, Binder's sets of partisan/electoral and institutional variables do not explain gridlock level is also drawn when the Coleman and Chiou-Rothenberg measures are adopted. Except for budgetary situation, none of the explanatory variables in these equations' estimates are consistently significant.[11] Interestingly, these results are qualitatively invariant whether W-NOMINATE or common space scores are used with grouped-logit estimation, but adopting W-NOMINATE does change results when Prais-Winsten estimation is employed.[12] Regardless, despite the higher adjusted R^2 found when alternative measures of legislative gridlock are adopted, meaningfully explaining the gridlock level remains problematic.

As for legislative productivity, the absolute number of initiatives passed, we include not only Mayhew's measure, which Binder also analyzes, but also three alternative measures of Howell et al. (2000) that classify enactments with different criteria of legislative significance.[13] Additionally, although our results are essentially invariant to estimation technique, in Table 3 we present Poisson regression results.[14]

The findings in Table 4 have two interesting features. First, Binder's partisan/electoral and institutional variables actually seem to do better explaining legislative productivity than

[11]We might be surprised that there are high adjusted R^2s but few significant coefficients in Table 3. This is a function of multicollinearity among control variables and one or two of the institutional variables. For the analyses of the Coleman measure, joint F tests on budgetary situation, public mood, and either bicameral distance or severity of filibuster threat show multicollinearity. But we also find both that the signs of any impact of bicameral distance or filibuster severity on the dependent variable are in the wrong direction and that it is these four variables which contribute to the high adjusted R^2s. Similarly, for the analyses of the Chiou-Rothenberg measure, a joint F test on the combination of budgetary situation, public mood, time out of majority, and severity of filibuster threat shows strong multicollinearity that is responsible for high adjusted R^2s. However, we can again discover no evidence that significant effects in the expected direction are being obscured (indeed, the direction of the filibuster threat variable is in the wrong direction). Overall, despite the model fit, these joint F tests provide no reason for inferring that the key factors that Binder (1999) highlights impact legislative gridlock.

[12]For the Coleman measure, percentage of moderates, time out of majority, and policy mood are significant in the right direction with one-tailed tests, although diversity, bicameral distance, and filibuster work in the opposite direction of that posited (only diversity is significant); for the Chiou and Rothenberg measure, divided government, percentage of moderates, time out of majority, and policy mood are significant, whereas diversity and filibuster work in the opposite direction. These findings are consistent with our discussion about how measurement error, especially in a small sample, can affect statistical results.

[13]Howell et al. (2000) distinguish between Groups A, B, and C: Group A essentially contains those laws in Mayhew's (1991) Sweep One; Group B contains all other public laws mentioned in either the *New York Times* or the *Washington Post* and which received at least six pages of coverage in the *CQ Almanac*; and Group C includes all other public laws mentioned in the *CQ Summary*.

[14]Results for OLS (which Binder uses), generalized negative binomial, and Prais-Winsten estimation are roughly comparable (and, not surprisingly, using OLS makes our results even more similar to Binder's). Since the data are time series, we also use a log-linear model using iterative weighted least squares, developed by Katsouyanni et al. (1996) and Schwartz et al. (1996), which estimates a Poisson regression first to obtain parameter starting values and then incorporates its residuals into subsequent iteration of the model. The construction of the variance-covariance matrix is flexible enough to allow for both autocorrelation and overdispersion (*Stata*'s arpois procedure implements this model). Although we find highly significant first-lag autocorrelation for Mayhew's data and second-lag autocorrelation for the data of Howell et al., results are not qualitatively changed.

https://doi.org/10.1093/pan/mpm021 Published online by Cambridge University Press

Table 3 Reestimation of determinants of policy gridlock—alternative measures (coefficients with SEs in parentheses)

	Coleman (1999)		*Chiou and Rothenberg (2003)*	
	Means of estimation		*Means of estimation*	
Variables	*Grouped logit*	*Prais-Winsten (robust SEs)*	*Grouped logit*	*Prais-Winsten (robust SEs)*
Divided government	0.167 (0.232)	−0.023 (0.034)	0.304 (0.228)	0.047 (0.031)
Percentage of moderates	−0.009 (0.041)	0.003 (0.005)	0.025 (0.041)	0.009 (0.006)
Ideological diversity	9.741 (29.378)	5.607 (4.799)	16.279 (29.529)	4.796 (5.143)
Time out of majority	−0.083 (0.209)	−0.049 (0.036)	−0.078 (0.199)	−0.027 (0.039)
Bicameral distance	−1.473 (3.088)	−1.450 (0.742)	−1.241 (3.067)	−0.804 (0.957)
Severity of filibuster threat	−0.066 (0.160)	0.024 (0.033)	−0.046 (0.156)	−0.006 (0.037)
Budgetary situation	−0.050*** (0.019)	−0.005 (0.005)	−0.046** (0.018)	−0.008* (0.005)
Public mood (lagged)	0.011 (0.031)	0.003 (0.004)	0.004 (0.030)	0.003 (0.005)
Constant	−3.055 (10.146)	−1.179 (1.480)	−5.252 (10.190)	−1.169 (1.659)
Adjusted R^2	0.526	0.892	0.513	0.743
Durbin-Watson[a]	2.697		2.320	

Note. Number of cases = 20. One-tailed t test, *$p < .1$; **$p < .05$; ***$p < .01$.
[a]Value from original OLS estimation.

legislative gridlock, i.e., along with divided government and budgetary situation, the percentage of moderates and ideological diversity are statistically significant using some measures of productivity. Furthermore, findings for Mayhew's measure qualitatively correspond to those reported by Binder (except for divided government); similarly, many of the findings for the three measures of Howell et al. are roughly comparable to the results found by Binder.[15]

Thus, when the sum total of our empirical results for gridlock and productivity is considered, there is not much support for the claim that considering a broader universe of potential enactments substantially changes inferences concerning legislative dynamics. Rather, what differences seemed to exist are largely a function of measurement error produced by using W-NOMINATE scores.

6 Conference Votes as an Alternative Measure

Having demonstrated that Binder's (1999) initial results are not robust, we now briefly turn to an alternative attempt to measure what Binder sees as the most critical source of intrabranch conflict, i.e., bicameral difference.[16] In an Appendix to the monograph incorporating her initial work, Binder (2003) responds to an initial version of our research with common space scores by developing a new measure of bicameral difference not relying upon W-NOMINATE.

[15]Note that, relative to the findings for the proportion of key proposal enacted, the results with common space scores for the number of major pieces of legislation passed are qualitatively more comparable to those with W-NOMINATE. For example, bicameral difference is inconsequential however it is measured. Thus, measurement error's impact on the results for the absolute numbers of enactments is not nearly as pronounced as for the proportional measure. As we have already elaborated, this is quite consistent with what can happen with measurement error in a small sample. The effects of measurement error are still being felt but within the context of the greater uncertainties and variances that come with small samples.
[16]See the online appendix for a more formal treatment than that presented here.

https://doi.org/10.1093/pan/mpm021 Published online by Cambridge University Press

Table 4 Reestimation of production of landmark laws (poisson regression coefficients with SEs in parentheses)

	Mayhew (1991)	*Howell et al. (2000)*		
		Group A	*Group B*	*Group C*
Divided government	−0.257* (0.159)	−0.449*** (0.177)	−0.185 (0.162)	−0.058 (0.065)
Percentage of moderates	0.081*** (0.032)	0.063** (0.041)	0.101*** (0.033)	0.005 (0.015)
Ideological diversity	8.707 (19.706)	14.610 (24.474)	−8.849 (21.209)	−20.486*** (9.160)
Time out of majority	−0.017 (0.041)	0.028 (0.200)	0.088 (0.174)	−0.260 (0.078)
Bicameral distance	0.810 (2.325)	0.999 (2.644)	−1.019 (2.265)	2.984 (0.963)
Severity of filibuster threat	0.370 (0.118)	0.331 (0.138)	0.312 (0.120)	0.208 (0.048)
Budgetary situation	0.046*** (0.014)	0.038*** (0.017)	0.006 (0.014)	0.028*** (0.006)
Public mood (lagged)	−0.005 (0.021)	−0.003 (0.027)	0.030* (0.023)	−0.029 (0.010)
Constant	−1.808 (6.769)	−3.367 (8.694)	0.395 (7.368)	11.667*** (3.258)
Number of cases	22	21	21	21
Log-likelihood	−51.947	−46.386	−58.061	−106.504
Durbin-Watson[a]	2.797	2.122	2.063	2.105

Note. One-tailed t test, *$p < .1$; **$p < .05$; ***$p < .01$.
[a]Value from original OLS estimation.

Specifically, she uses the absolute difference between the House's and the Senate's percentage average approval on common votes, calculating the average approval percentage for conference report bills on which either voice or recorded votes take place in both chambers. As many conference votes are done by voice, she assumes that voice votes are equivalent to unanimous votes of approval or disapproval.

Employing her new measure, Binder's results are consistent with her analysis using W-NOMINATE. Furthermore, upon finding that her conference vote measure correlates at 0.88 with the common space bicameral difference measure for the 12 Congresses before 1971 but at only −0.43 for the 15 post-1970 Congresses, she concludes that "common space coordinates are ill-suited for detecting bicameral differences after the second dimension disappears after the 1960s" (Binder 2003, 146–7). She further argues that the equal weighting of the two ideological dimensions used to calculate common space scores produces measurement error that leads to this correlative pattern and the differences in empirical findings documented above.[17] Ultimately, she infers that her theory of

[17]Binder (2003) misinterprets the assumption that common space coordinates are equally salient as implying that the varying importance of the second dimension affects the comparability of first dimension scores over time or across chambers. As with W-NOMINATE, this is not the case. To reiterate, the two estimated dimensions of common space scores are equivalent to eigenvectors that are extracted independently. Indeed, first dimension common space and W-NOMINATE scores correlate at *0.95 or higher* for all Houses and Senates. Not surprisingly, then, the historical correlative pattern discussed above with respect to common space scores is also found for W-NOMINATE, so any explanation for temporal changes goes beyond anything unique to common space scores. Moreover, as we will illustrate, it is possible that Binder's finding about the historical pattern of correlations is caused by differences in the proportion of conference votes that are voice votes before and after 1970.

https://doi.org/10.1093/pan/mpm021 Published online by Cambridge University Press

legislative gridlock can be sustained without relying upon W-NOMINATE to measure preferences.

However, this effort to buttress earlier findings falls short, as the measure is almost certainly not measuring preference differences exclusively. Even if results from the analysis with conference votes are the same as those generated using W-NOMINATE, they do not provide additional support.

First, there is a *prima facie* case that the conference vote measure is tapping something besides bicameral difference. The correlation coefficients of the conference vote measure with the W-NOMINATE and the common space measures of bicameral difference are only 0.135 and −0.011. If either of these earlier measures is capturing true differences, then it is hard to maintain that this newer measure is doing the same.

Second, and related to these low correlations, including voice votes as unanimous votes underestimates bicameral differences and produces measurement error biasing scores toward zero if the unanimity assumption is wrong. For example, if voice votes occur because either minority coalitions know the final outcome and do not wish to publicly lose, or, particularly before electronic voting, members prefer that valuable time not be wasted, there is a downward bias. Furthermore, there are not enough relevant votes, particularly before 1970, to drop voice votes. Except for the 81st Congress, there are less than six nonvoice conference votes in each of the Congresses between the 80th and 88th Congresses, and the mean number of nonvoice conference votes per Congress from the 80th through the 106th Congresses is 15.7.

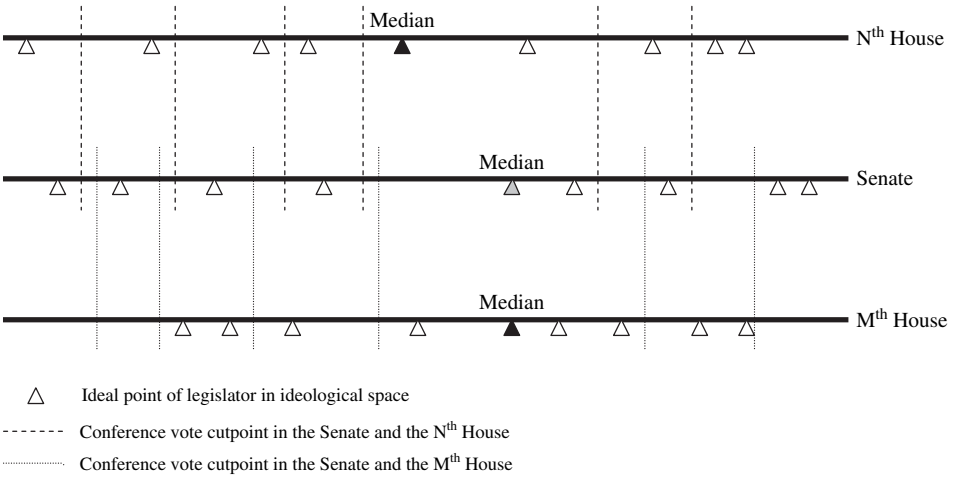
Additionally, unlike W-NOMINATE or common space scores, using the conference vote measure necessarily combines first and second dimension preferences. Distinguishing which votes are driven by the first dimension and which votes are driven by the second, even if possible, would further restrict the already small number of votes available to measure preferences.

Finally, for average differences to approximate true bicameral differences for the entire period being studied requires very strong conditions being met.[18] For equivalence with true bicameral differences, defined as preference differences between the House and the Senate medians, very strong assumptions about the curvature of preference distributions by Congress and the distribution of cut points capturing the difference between proposals and the status quo must be made for every Congress being compared. As such, equivalence between true preference differences and the conference vote measure is a knife-edged case.

Figure 2 illustrates this point graphically. Suppose that we have a bicameral legislature, with the House and the Senate each having nine members (it does not matter that, in reality, the U.S. House is larger than the Senate). Assume also that there are six, nonvoice, conference votes. Now, consider what happens when we look at the relationship between bicameral difference and average difference for two realizations of House preference distributions, which we label as the *N*th and *M*th Houses; for the *N*th House and the Senate, the cut points are represented by long-dashed lines, whereas for the *M*th House and the Senate, short-dashed lines are employed. Despite the House median in the *N*th House being more liberal than its Senate counterpart, the absolute difference between the two chambers' average approval percentage in this case is zero. By contrast, if we

[18]This problem is roughly analogous to Snyder's (1992a) analysis of how employing interest group ratings to measure legislative ideal points can artificially produce extreme rankings because groups focus on close votes with cut points around the center of the ideological spectrum. However, there are fairly straightforward cures for the problem that Snyder identifies, such as overweighting lopsided votes. There is no analogous quick fix for the measurement problem here, as there are few conditions where the average difference measure will even approximately capture true preference differences.

https://doi.org/10.1093/pan/mpm021 Published online by Cambridge University Press



Note: Bicameral distance is measured by the absolute difference between the Senate and House medians; in our illustration, this distance is 0 if the N^{th} House preference distribution is realized and it is positive given the M^{th} House preference distribution.

Fig. 2 Effects of preference and cut point distributions on conference measure of bicameral differences—an illustration.

substitute the *M*th House, so that the Senate and House medians now have the same preference, this absolute difference is 0.148. In other words, not only is the absolute difference not conceptually equivalent to the bicameral difference but also larger bicameral differences may correspond to smaller approval differences.

Hence, although creative, using common conference vote differences is not a solution. There is measurement error, an inability to separate preference dimensions, and a need to meet extremely strict assumptions about preference distributions within and across legislatures. Furthermore, these problems are interrelated: for example, the use of voice votes virtually ensures that the curvature assumptions are not met.

7 Discussion

To summarize, Binder (1999) offers an important perspective on policy gridlock. Despite this contribution, her adoption of W-NOMINATE produces results that are not borne out when measurement error is addressed by substituting more appropriate common space scores. Contrary to this initial analysis, we find that the partisan/electoral (percentage of moderates, ideological diversity, time since the current majority has had majority status) and institutional variables (severity of filibuster threat and bicameral distance) that Binder emphasizes do not explain legislative gridlock. This is true whether we use Binder's (1999) gridlock measure or other available alternatives. We also show that these partisan/electoral and institutional variables better explain legislative productivity than gridlock and that results for the absolute number of legislative outputs are not as sensitive to preference measurement as those for gridlock. Furthermore, although somewhat intuitively appealing, we demonstrate that one cannot sidestep problems of scaling by backing out bicameral difference using conference votes (Binder 2003).

Do these results indicate that the kind of partisan/electoral and institutional forces that Binder stresses are simply unimportant for understanding how legislators overcome gridlock? And, with our use of better measurement, have we settled much of the debate

https://doi.org/10.1093/pan/mpm021 Published online by Cambridge University Press

about when legislative initiatives are produced? Although we believe that our analysis provides important evidence about the factors that do and do not impact gridlock, it also provides reasons for caution in rushing to such judgments in answering these and similar questions.

We see that, besides paying careful attention to theory, addressing important questions that require comparing legislatures over time demands great sensitivity to the kinds of measurement concerns highlighted here. Otherwise, we may end up drawing wrong conclusions, be they false positives or false negatives. As we have seen, a seemingly defensible, innocuous decision, such as choosing to use W-NOMINATE scores for a purpose for which they are not intended, can have dramatic results. Hopefully, as work on legislative outputs and congressional macropolitics continues to progress, our analysis will serve to make scholars more aware of the assumptions they are making regarding measurements and the pitfalls that might be created. Even more generally, our analysis shows just how many roadblocks there can be for addressing many concerns of contemporary institutional scholars as such issues frequently involve comparing political decision makers in multiple institutions over time. Without proper measurement, even the most creative analysis may yield incorrect inferences.

References

Achen, Christopher. 1985. Proxy variables and incorrect signs on regression coefficients. *Political Methodology* 11:299–316.

Bailey, Michael. 2005. Bridging institutions and time: Creating comparable preference estimates for presidents, senators, representatives and justices, 1950–2002. Unpublished manuscript, Georgetown University.

Binder, Sarah A. 1999. The dynamics of legislative gridlock, 1947–96. *American Political Science Review* 93:519–33.

———. 2003. *Stalemate: Causes and consequences of legislative gridlock.* Washington, DC: Brookings Institution.

Brady, David W., and Craig Volden. 2006. *Revolving gridlock: Politics and policy from Jimmy Carter to George W. Bush.* 2nd ed. Boulder, CO: Westview Press.

Chiou, Fang-Yi, and Lawrence S. Rothenberg. 2003. When pivotal politics meets partisan politics. *American Journal of Political Science* 47:503–22.

Cole, Rosanne. 1969. Data errors and forecasting accuracy. In *Economic forecasts and expectations*, ed. Jacob Mincer, 40–73. New York: National Bureau of Economic Research.

Coleman, John J. 1999. Unified government, divided government, and party responsiveness. *American Political Science Review* 93:821–35.

DeVaro, Jed L., and Jeffrey M. Lacker. 1995. Errors in variables and lending discrimination. *Federal Reserve Bank of Richmond Economic Quarterly* 81(3):19–32.

Edwards, George C., III, Andrew Barrett, and Jeffrey Peake. 1997. The legislative impact of divided government. *American Journal of Political Science* 41:545–63.

Greene, William H. 2003. *Econometric analysis.* 5th ed. Upper Saddle River, NJ: Prentice-Hall.

Hashimoto, Masanori, and Levis Kochin. 1980. A bias in the statistical estimation of the effects of discrimination. *Economic Inquiry* 18:478–86.

Herron, Michael C. 2004. Studying dynamics in legislator ideal points: Scale matters. *Political Analysis* 12:182–90.

Hodges, S. D., and P. G. Moore. 1972. Data uncertainties and least squares regression. *Applied Statistics* 21: 185–95.

Howell, William, Scott Adler, Charles Cameron, and Charles Riemann. 2000. Divided government and the legislative productivity of Congress, 1945–1994. *Legislative Studies Quarterly* 25:285–312.

Katsouyanni, K., J. Schwartz, C. Spix, G. Touloumi, D. Zmirou, A. Zanobetti, B. Wojtyniak et al. 1996. Short-term effects of air pollution on health: A European approach using epidemiological time series data. The APHEA protocol. *Journal of Epidemiology and Community Health* 50(Suppl 1):S12–8.

Krehbiel, Keith. 1996. Institutional and partisan sources of gridlock: A theory of divided and unified government. *Journal of Theoretical Politics* 8:7–40.

Krehbiel, Keith. 1998. *Pivotal politics: A theory of U.S. Lawmaking.* Chicago, IL: University of Chicago Press.

https://doi.org/10.1093/pan/mpm021 Published online by Cambridge University Press

Levi, Maurice D. 1973. Errors in the variables bias in the presence of correctly measured variables. *Econometrica* 41:985–6.

Mayhew, David. 1991. *Divided we govern.* New Haven, CT: Yale University Press.

Poole, Keith T. 1998. Recovering a basic space from a set of issue scales. *American Journal of Political Science* 42:964–93.

———. 2003. Changing minds! Not in Congress. Unpublished manuscript, University of Houston.

Poole, Keith T., and Howard Rosenthal. 1997. *Congress: A political economic history of roll call voting.* New York: Oxford University Press.

Rao, Potluri. 1973. Some notes on the errors-in-variables model. *American Statistician* 27:217–8.

Rothenberg, Lawrence S., and Mitchell Sanders. 2004. Much ado about very little: Reply to Herron. *Political Analysis* 12:191–5.

Schwartz, J., C. Spix, G. Touloumi, L. Bachárová, T. Barumamdzadeh, A. le Tertre, and T. Piekarksi et al. 1996. Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions. *Journal of Epidemiology and Community Health* 50(Suppl 1):S3–11.

Snyder, James M. 1992a. Artificial extremism in interest group ratings. *Legislative Studies Quarterly* 17:319–45.

———. 1992b. Long-term investing in politicians; or, give early, give often. *Journal of Law and Economics* 35:15–43.

https://doi.org/10.1093/pan/mpm021 Published online by Cambridge University Press