

# Measuring Legislative Accomplishment, 1877–1994

**Joshua D. Clinton** Princeton University

**John S. Lapinski** Yale University

*Understanding the dynamics of lawmaking in the United States is at the center of the study of American politics. A fundamental obstacle to progress in this pursuit is the lack of measures of policy output, especially for the period prior to 1946. The lack of direct legislative accomplishment measures makes it difficult to assess the performance of our political system. We provide a new measure of legislative significance and accomplishment. Specifically, we demonstrate how item-response theory can be combined with a new dataset that contains every public statute enacted between 1877 and 1994 to estimate “legislative importance” across time. Although the resulting estimates and associated standard errors provide new opportunities for scholars interested in analyzing U.S. policymaking since 1877, the methodology we present is not restricted to Congress, the United States, or lawmaking.*

There is a growing literature in political science that seeks to understand the nature of lawmaking. This work has focused primarily on the causes and consequences of legislative “gridlock” (see, for example, Binder 1999; Coleman 1999; Cox and McCubbins 2004; Krehbiel 1998; Mayhew 1991; McCarty, Poole, and Rosenthal 2005). Although this research has increased our understanding of lawmaking, restricted data combined with limited advancement in conceptualizing and measuring legislative output has hindered progress. Most scholars (for example, Coleman 1999; Epstein and O’Halloran 1999; Erikson, MacKuen, and Stimson 2002; Krehbiel 1998) rely heavily on a single source of data—the list of enactments produced by Mayhew for his seminal *Divided We Govern* (1991). Although Mayhew’s data sets the standard for measuring legislative accomplishment, it has limitations: it does not precede World War II, and it captures only landmark legislation, which is a very small

fraction of enacted public statutes. In addressing the need to measure legislative accomplishment, our aim is to construct a measure that spans a very long time period and comprehensively assesses the legislative output produced by Congress.

In assessing legislative accomplishment, we are mindful of the contributions of past work to our knowledge about lawmaking, and we draw on that work extensively. One thing we know from previous work is that Congress passes thousands of inconsequential laws. Cameron makes this point when he writes: “There is a little secret about Congress that is never discussed in the legions of textbooks on American government: the vast bulk of legislation produced by that august body is stunningly banal . . . . The president, like almost everyone else, couldn’t care less about them” (2000, 37). Passing inconsequential laws is an important activity for members of Congress because such laws represent pork politics. In

---

Joshua D. Clinton is assistant professor of politics, Princeton University, 036 Corwin Hall, Princeton, NJ 08544-1012 (Clinton@princeton.edu). John S. Lapinski is assistant professor of political science, Yale University, P.O. Box 208209, New Haven, CT 06520-8209 (John.lapinski@yale.edu).

The authors would like to thank: R. Douglas Arnold, David Brady, Ira Katznelson, David Mayhew, Mat McCubbins, Nolan McCarty, Rose Razaghian, Doug Rivers, Ken Scheve and participants of the Center for the Study of Democratic Politics at Princeton University, participants of the American politics workshops at Columbia University, Northwestern University, and MIT, Cal Tech’s political economy workshop participants, and participants in the History of Congress II Conference at Stanford University for helpful suggestions and reactions. Lapinski thanks the Russell Sage Foundation’s Scholars in Residence Program for a conducive environment for research and writing as well as the National Science Foundation (SES 0318280) and Columbia University’s ISERP and the Institution for Social and Policy Studies at Yale University for financial assistance for the data collection effort. Assistance in data collection has been provided by many individuals, particularly helpful were Eldon Porter, Andrew Roach, Sylvia Wu, Don Phan, Cleve Doty, and Daniel Clemens of Yale University, Melanie Springer and Christine Greer of Columbia University, and Nancy Danch of the Bobst Center for Peace and Justice at Princeton University.

*American Journal of Political Science*, Vol. 50, No. 1, January 2006, Pp. 232–249

©2006 by the Midwest Political Science Association

ISSN 0092-5853

other words, such enactments are usually constituency-oriented.<sup>1</sup> However, testing theories of lawmaking, as well as building new ones, on trivial legislation seems to us to be suspect. We think it is necessary to identify laws that matter as it could be argued that many theories of lawmaking are relevant only for a subset of congressional activity—that activity that could be considered consequential. If the nature of the politics depends on the stakes of the policy being considered, tests of theories relying on measures such as voting behavior and coalition size may be valid only when conducted on significant legislation.

Our effort contributes to the field in several ways. First, the comprehensive dataset we have built of the 37,766 public statutes passed between 1877 and 1994 (the 45th to 103rd Congresses) provides an unparalleled opportunity to characterize congressional policymaking. Our work is unique in that we have collected information, not on a subsample of statutes, but on every public statute enacted during this period. Second, we have collected a large sample of elite evaluations (20) of the importance of public enactments at different points in time. Third, we use an item-response model to integrate the information in these ratings.<sup>2</sup> In contrast to the reliance of existing approaches on the determinations of a single scholar, we use a statistical model to integrate all rater information and account for rater differences. For example, instead of choosing between the determinations of Mayhew (1991) and Howell et al. (2000), we present a method that combines the information of each. The method reconciles disagreements and combines raters who assess different periods of American history to yield a long-term series of legislative accomplishment.

Our statute-level estimates for every enactment stand in contrast to existing measures that treat all significant and nonsignificant legislation as equally significant and perfectly measured. We discriminate between rated legislation, and we use information regarding the amount of attention Congress devotes to each statute, the number of substitute bills, whether a conference committee was

required, and the session in which the legislation was introduced, to extend the ratings to statutes unmentioned by the raters we utilize. Estimating the significance and standard errors of individual statutes allows for different levels of analysis. For example, we can construct a macrolevel analysis of the performance of Congress and the president, or we can use the statute level estimates to conduct microlevel studies on individual bills.

Although we focus on lawmaking activity in the U.S. Congress from 1877 to 1994, the method we present could be employed to identify other important governmental activities, such as executive orders of the U.S. presidency, administrative rule changes of the U.S. bureaucracy, rulings of the National Labor Relations Board, decisions of the U.S. Supreme Court, United Nations resolutions, acts of the European Union, or lawmaking by any legislative body. The method has far-reaching applicability to the study of institutions and offers a means of integrating information in such a way as to estimate decision-level estimates of importance and their associated precision.<sup>3</sup>

## Significant Legislation and the Role of “Raters”

Our work takes as its substantive foundation the pioneering work of David Mayhew. In his *Divided We Govern* (1991), Mayhew tackles the difficult issues of measurement and selection in his analysis of lawmaking and assessment of the role of Congress and the president in the policymaking process. His analysis, which focuses directly on lawmaking rather than on indirect measures such as roll-call votes, provides important information and ideas about measuring the legislative accomplishment of Congress.<sup>4</sup> Mayhew increases our understanding of: what type of data is important for studying lawmaking; what defines “important,” “notable,” or “significant” legislation; and the importance of testing theory and hypotheses

<sup>1</sup>Pork barrel laws, of course, need not be insignificant. Often times large omnibus bills are full of pork. Take for example, Ferejohn’s seminal study on pork politics (1974). His work focused on highly significant Rivers and Harbors legislation between 1947 and 1968. In any case, our own empirical work has revealed many insignificant pork bills. Between 1877 and 1954, 4,899 out of a total of 24,037 enactments are what Katznelson and Lapinski (2005) call quasi-public enactments. Quasi-public enactments are pork bills which include the construction of individual public works projects (e.g., public building or bridges).

<sup>2</sup>The method we present is not new to political science (for example, see Martin and Quinn (2002) and Clinton, Jackman, and Rivers (2004)). However, our application is new and substantively important.

<sup>3</sup>Nothing limits the method to “important” decisions. The method could be used to estimate any latent trait.

<sup>4</sup>Mayhew’s (1991) postwar time series of the count of significant legislative enactments by Congress has been used to test a number of theories dealing with congressional policymaking. In arguing for a distinction between “landmark” and ordinary legislation, Mayhew measures accomplishment by constructing a list of important enactments based on several sources. Using this measure, he contradicts the conventional wisdom that unified party government facilitates legislative productivity and divided government results in gridlock. Evidence of Mayhew’s influence can be found in the work of Binder (1999, 2003), Coleman (1999), and Howell and his coauthors (2000), all of whom address Mayhew’s initial findings and draw upon his methodology for measuring productivity.

against actual policy outputs. Of these three contributions, the second is the most opaque.

### The Concept of Legislative Significance

Although existing lists of legislation purport to identify legislation that is “landmark,” “significant,” “innovative,” or “consequential,” what is meant by “important” legislation is inevitably, if not hopelessly, imprecise. Any characteristics that one could plausibly identify as defining “significant” legislation are themselves plagued by imprecision. For example, Mayhew defines important legislation as that which is “both innovative and consequential—or if viewed from the time of passage, thought likely to be consequential” (1991, 37). What does he mean by “consequential” and “innovative”? Should we measure policy innovation by the percentage of the U.S. Code that is changed by a statute or by how far a policy moves the status quo? Does “consequential” refer to the extent to which a policy changes existing laws, the number of people affected by the legislation, or a measure of the extent of that impact on the affected people? Even if a precise definition were possible, operationalizing the definition appears impossibly complex. For example, although the Civil Rights Act of 1964 and the Voting Rights Act of 1965 are unquestionably two of the most consequential civil rights policies ever enacted in the United States, how are we to assess the consequence and innovation of symbolic legislation such as the act that established Martin Luther King Day as a national holiday or the 1989 Flag Protection Act outlawing flag burning?<sup>5</sup>

Mayhew (1991, 4) utilizes both contemporaneous and retrospective evaluations of the policymaking process—essentially elite evaluations—to assess which legislative enactments meet this standard and qualify as “important.”<sup>6</sup> The resulting list has strong validity and can be rationalized *ex post*, but there is no necessary relationship between the posited criteria of innovation and consequence and whether a statute is sufficiently note-

<sup>5</sup>Not only can symbols be very important to people, but by their very nature they are likely to be controversial and to capture the attention of the media and the mass public. Most observers would agree that allowing or prohibiting flag burning does not affect the day-to-day lives of as many Americans as the Civil Rights Act does, but deciding how to evaluate the comparative consequences of such enactments is not so clear.

<sup>6</sup>Enactments for Mayhew include constitutional amendments and Senate treaty ratifications but exclude “congressional resolutions, appropriations acts, very short-term measures (such as a one-year extension of rent control in 1947), statutory amendments taken by themselves, and extensions or reauthorization acts that seemed to offer little new.”

worthy to appear in a review of the legislative session by major newspapers or a policy history. Although innovative and consequential legislation is likely to be mentioned (and therefore captured by the measure), Mayhew admits that coverage of an enactment may also be affected by other characteristics, such as how controversial it is. Certainly legislation that fundamentally changes the nature of government would be controversial, but controversy may also stem from the political environment rather than the enactment itself. It is not implausible that identical legislation might result in substantially different levels of controversy depending on the political environment. Legislation perceived as controversial during an era of political acrimony might pass with bipartisan support at a time of relative calm. Consequently, Mayhew’s operationalization of importance using notable legislation effectively identifies three types of legislation: legislation controversial enough to warrant coverage; legislation with a great enough impact to warrant coverage; and legislation innovative enough to warrant coverage.

Attempts to enumerate the precise conditions under which a statute may be considered significant lead quickly to a long list of conditions that prove difficult, if not impossible, to operationalize. Even if it were possible to specify precisely what constitutes an “important” statute, the usefulness of such an exercise is doubtful given the likely impossibility of making a determination based on such standards. The difficulty, if not futility, of such an exercise leads us to define significant legislation explicitly in terms of that which is notable. Our working definition of significant legislation is a statute or constitutional amendment that has been identified as noteworthy by a reputable chronicler-rater of the congressional session.<sup>7</sup> This definition is identical to that put forth by Mayhew in his *Divided We Govern*. Although our measure is more accurately described as a way to identify notable legislation, we follow the usage of others, like Mayhew, and refer to “notable” legislation as “significant” legislation. We acknowledge the slippage that results from this terminology, but it is unclear to us that a superior alternative exists.

### Contemporaneous Raters

Following Mayhew (1991), we characterize raters according to whether their assessments are primarily contemporaneous or retrospective. Contemporaneous raters assess legislation at or near the time of enactment. Retrospective

<sup>7</sup>Note that Mayhew (1991) sometimes substitutes the word “notable” for “significant.”

raters rely on hindsight to determine the noteworthiness of legislation in the context of both prior and subsequent events.<sup>8</sup>

Mayhew's Sweep 1 draws on contemporaneous sources in constructing a list of important enactments and represents the first rater we use. Another contemporaneous source comes from Baumgartner and Jones's Policy Agendas Project (2002, 2004), which collects information on every public statute enacted since 1948. Baumgartner and Jones's list of important enactments is based on the number of column lines devoted to each enactment in the *Congressional Quarterly Almanac*.

Additional sources of contemporaneous assessments are provided by the coverage of the *New York Times* and the *Washington Post*. These measures were initially collected and summarized by Mayhew to form his Sweep 1 assessment; subsequently they were independently collected and refined by Cameron (2000) and Howell and his colleagues (2000). The latter combine information collected from these two newspapers with story coverage in CQ and Mayhew's series to construct a four-category ordinal measure of significance for the period 1948 to 1993. We use an indicator of whether the legislation passes the lowest threshold for nontrivial legislation ("A" through "C" in the coding of Howell et al. 2000).<sup>9</sup>

We also use assessments made in the legislative wrap-ups of the *American Political Science Review* (APSR) and *Political Science Quarterly* (PSQ). Each journal summarized the year's proceedings in Congress for the years 1889 to 1925 (PSQ) and 1919 to 1948 (APSR).<sup>10</sup> We also use Peterson's (2001) attempts to replicate Mayhew's coding technique for 1881 to 1945 using contemporaneous sources that vary from Congress to Congress.

In each case, we recorded and content-coded all mentions of public enactments and policy changes. If the title of the statute was not present in the text, we searched through the index of the History of Bills and Resolutions in the *Congressional Record* as well as the index of the *Pub-*

*lic Statutes at Large* to derive a dichotomous measure of whether the statute was mentioned.<sup>11</sup>

## Retrospective Raters

A second type of rating data we use is retrospective evaluations. Mayhew's Sweep 2 for the 1948 to 1994 period and Peterson's (2001) replication of Mayhew's Sweep 2 for the 1881 to 1947 period are two sources of retrospective ratings. In addition to the retrospective ratings of Mayhew and Peterson, which aggregate the determinations of several different policy histories, we also rely on several general political history and policymaking series. We use 11 period-specific volumes of the *New American Nation* series—a "comprehensive, cooperative history of the area now embraced in the United States, from the days of discovery to the present" (Matusow 1984, ix) and the *American Presidency Series* (see appendix).<sup>12</sup> In the *American Presidency Series*, "each book treats the then-current problems facing the United States and its people and how the president and his associates felt about, thought about, and worked to cope with these problems. In short, the authors in this series strive to recount and evaluate the record of each administration and to identify its distinctiveness and relationships to the past, its own time, and the future" (Greene 2000, ix). A complete listing of the individual books of these series is listed in the appendix.

A fifth source of retrospective ratings is Chamberlain's *The President, Congress, and Legislation* (1946) which contains a detailed history of U.S. policymaking across 10 distinct spheres. A sixth source of retrospective ratings is Reynolds's (1995) "mini-research report" which culled the *Enduring Vision* textbook for mentions of important legislation between 1877 and 1948 to test for the effect of divided government on policymaking before World War II. A seventh source is Sloan's *American Landmark Legislation* (1984) which compiles a selective list of laws based on: the "important national significance they had at the time Congress passed them" and their lasting effect on "one dimension or another of American life." Yet another retrospective rater is Light's *Government's Greatest Achievements: From Civil Rights to Homeland Defense* (2002). A ninth set of retrospective raters are several of the public affairs history books identified in Mayhew's *America's*

<sup>8</sup>Our treatment of raters who use both retrospective and contemporaneous assessments as being of the retrospective type has no substantive or analytical consequence.

<sup>9</sup>A rating of "C" or above means that a given enactment is mentioned in the *Congressional Quarterly Almanac*, the *New York Times*, or the *Washington Post*. Although the higher thresholds of Howell et al. (2000) could be used, Baumgartner and Jones's (2002, 2004) measure already essentially accounts for this information.

<sup>10</sup>Although the *Western Political Quarterly* (WPQ) continued the legislative wrap-ups after the APSR stopped—even using the same author (Floyd Riddick) the APSR had used for a number of Congresses—the content of WPQ changed significantly, making these wrap-ups uninformative for our purposes.

<sup>11</sup>*Congressional Universe* and *Academic Universe* were also used to identify public laws associated with the policy using available information (such as the date of the enactment or legislative activity).

<sup>12</sup>Although Bornet's *The American Presidency of Lyndon B. Johnson* is a part of the American Presidency series, it differs substantially in its treatment of legislation. We consequently treat it as a separate rater.

TABLE 1 Descriptive Statistics of Raters

Rater	Type	Time Period	Enactments Rated	Enactments Mentioned
Mayhew Sweep 1 (including updates)	Contemporaneous	80 <sup>th</sup> –103 <sup>rd</sup>	16933	216
Baumgartner & Jones	Contemporaneous	81 <sup>st</sup> –103 <sup>rd</sup>	16026	550
Howell et. al.	Contemporaneous	79 <sup>th</sup> –103 <sup>rd</sup>	17667	2282
APSR	Contemporaneous	66 <sup>th</sup> –80 <sup>th</sup>	11893	483
PSQ	Contemporaneous	50 <sup>th</sup> –68 <sup>th</sup>	9935	286
Peterson Sweep 1	Contemporaneous	47 <sup>th</sup> –79 <sup>th</sup>	20220	213
Mayhew Sweep 2(including updates)	Retrospective	80 <sup>th</sup> –101 <sup>st</sup>	15878	197
Peterson Sweep 2	Retrospective	47 <sup>th</sup> –79 <sup>th</sup>	20220	144
<i>New American Nation series</i>	Retrospective	45 <sup>th</sup> –90 <sup>th</sup>	29802	340
<i>American Presidency Series</i>	Retrospective	45 <sup>th</sup> –88 <sup>th</sup> ; 91 <sup>st</sup> –102 <sup>nd</sup>	35371	674
Chamberlain	Retrospective	45 <sup>th</sup> –76 <sup>th</sup>	18682	105
Reynolds	Retrospective	45 <sup>th</sup> –80 <sup>th</sup>	21741	84
Sloan	Retrospective	45 <sup>th</sup> –80 <sup>th</sup>	21741	16
Light	Retrospective	81 <sup>st</sup> –103 <sup>rd</sup>	16026	164
Grantham	Retrospective	81 <sup>st</sup> –99 <sup>th</sup>	13608	126
Barone	Retrospective	81 <sup>st</sup> –100 <sup>th</sup>	14321	63
Blum	Retrospective	87 <sup>th</sup> –93 <sup>rd</sup>	4954	30
Dell & Stathis	Retrospective	45 <sup>th</sup> –96 <sup>th</sup>	33589	582
Stathis	Retrospective	45 <sup>th</sup> –103 <sup>rd</sup>	37767	739

*Congress* (2000) which we treat as separate raters. Two final retrospective raters come from the work of the Congressional Research Service. Dell and Stathis (1982) and a *Landmark Legislation 1774–2002* (2003), expands, extends, but does not strictly restate the assessments of Dell and Stathis (1982). Table 1 summarizes the raters we use, including the number of statutes covered and the number of statutes mentioned. Statutes passed during the time period but not mentioned by a rater are counted as being unmentioned.

### Rater Differences

Given the inability to identify precisely how the legislation mentioned by raters corresponds to some acceptable definition of policy significance, we believe the task of constructing the “best” measure is impossible. Consider the task of resolving the differences between the lists compiled by Mayhew (1991) and Stathis (2003) for statutes passed in 1981. Mayhew identifies just two important statutes: the Economic Recovery Tax Act (PL 97-34) and the Omnibus Budget Reconciliation Act (PL 97-35). Stathis mentions these and adds the Veterans’ Health Care, Training,

and Small Business Loan Act (PL 97-72), Restrictions on Military Assistance and Sales to El Salvador (PL 97-113), Fiscal 1982 Department of Defense Appropriations (PL 97-114), and the Social Security Act Amendments (PL 97-123). Are we to conclude that two significant statutes were enacted in 1981, or six? Or some intermediate number? The justification for choosing any of these numbers is unclear. Are the Social Security Act Amendments important for having restored the minimum benefit eliminated by the Omnibus Budget Reconciliation Act and enabling the Old-Age and Survivors Insurance fund to borrow from the Hospital Insurance and Disability Insurance trust funds? Is it important or unimportant that the Veterans’ Health Care, Training, and Small Business Loan Act extended GI Bill eligibility for Vietnam veterans by two years? It is difficult to resolve such differences because the basis for these distinctions is unclear.

A benefit of our statistical method is that since a principled adjudication between competing significance lists is difficult, we can remain agnostic and let the data determine the appropriate resolution. Relationships between raters determine the relative contribution of each rater to the composite significance estimate. Since we use raters who explicitly or implicitly identify significant legislation, the resulting list is likely to consist of “high-stakes” legislation on “important,” “innovative,” or “consequential”

policy. Despite this ambiguity, for purposes of exposition we refer to our estimates as measuring “significance.”<sup>13</sup>

The number of available raters and the differences evident in their ratings raise three additional questions that are largely avoided in the literature on measuring legislative significance. How should we interpret rater disagreement? How should such heterogeneity affect a measure of significance? And how can we compare the assessments of raters of different time periods?

There are three ways in which we can interpret rater heterogeneity. First, raters may differ because they use different criteria. For example, it may be that the *American Presidency Series* focuses on legislation related to presidential programs and the *New American Nation* series focuses on statutes that are retrospectively notable and “stand the test of time.” If this is the case, then a composite measure is problematic—the aggregate measure reflecting disparate concerns is less meaningful than the individual ratings. We consciously select raters to minimize this possibility: the raters we use explicitly seek to identify significant legislation or to identify the major legislative events of the time period. Furthermore, our statistical model can test whether rater assessments are based on different underlying dimensions.

Second, raters may employ the same criteria, but different thresholds for designating statutes significant. For example, the *American Political Science Review* may be more likely to identify a statute as constituting an important policy change in its annual wrap-up than a rater such as Sloan, who surveys the entire time period before making such an identification. In other words, two raters may agree as to the dimension of interest but differ on the threshold that defines significance. Contemporaneous assessments are most vulnerable to this possibility, since publishing annual reviews may result in inflated assessments during periods of relative inactivity.<sup>14</sup> Finally, raters may make mistakes in their assessments. The

<sup>13</sup>We believe existing measures are most properly conceived of as lists reflecting raters’ judgments as to which statutes are most deserving of mention in their chronicle of legislative activity. Although legislation is presumably notable because it is “consequential,” “influential,” or “novel,” it is not necessarily so. Several of the raters we employ explicitly claim to identify consequential legislation, but nothing ensures either that all consequential legislation is mentioned or that inconsequential legislation is unmentioned.

<sup>14</sup>There is no evidence that contemporaneous ratings are inflated by the need to publish as the number of statutes mentioned by contemporaneous raters exhibits considerable variation. For example, whereas *PSQ* mentions only three statutes in the 50th Congress and five in the 60th, it mentions 30 statutes in the World War I 65th Congress. Such variation is consistent with the interpretation that the chroniclers are responsive to some statute characteristics rather than the need to mention a relatively constant number of statutes in each issue or article.

process of culling through and assessing legislation is exhausting and time-consuming. Raters may miss legislation that would have qualified as noteworthy according to their own standards.

The literature takes two approaches in accounting for rater disagreement. First, some argue that one list is preferable to another (see, for example, Howell et al. 2000; Kelly 1993; Mayhew 1993). Although such comparisons raise important points, it is difficult to assess them objectively and conclusively in light of the difficulties already noted of determining the relationship between any measure and a statute’s “true” significance. A second way in which competing lists are employed is to ignore the measurement debates and instead determine whether the results of interest depend on the list of significant legislation. In other words, researchers replicate their analyses using different measures (for example, Coleman 1999; Krehbiel 1998). Although pervasive, this approach fails to account for rater heterogeneity and examines only whether coefficients of interest (for example) depend on the measure used.

Accounting for rater heterogeneity and unanimity is important because such an accounting conveys information about both the magnitude and certainty of our assessment of legislative significance. It makes little sense to treat the twelve statutes that receive unanimous mention by active raters equivalently to the 2,164 statutes mentioned by a single rater between 1877 and 1994. Rater agreement may indicate increased significance or increased certainty about the probability that the legislation is significant relative to the significance of a bill that experiences rater disagreement.

Two additional concerns plague existing lists. First, the classification of landmark legislation is extremely coarse. It is hard to believe that the Voting Rights Act of 1982 and the Voting Rights Act of 1965 are equally significant. However, that is the conclusion suggested by the dichotomous ratings of Stathis (2003). Second, existing measures fail to quantify the precision of the measures. Rater disagreement suggests that our assessments of legislative significance are uncertain. Although we may be quite confident of the importance of the Civil Rights Act of 1964, which is mentioned by each of the 12 raters covering the period, we may be less certain of the significance of a statute mentioned by only two of the 12 raters. It is implausible that we are equally certain of the significance of the Civil Rights Act of 1964 and PL-379, which passed the same year and established water resource research centers at land-grant colleges and state universities.<sup>15</sup>

<sup>15</sup>PL-379 is mentioned in the Johnson volume of the *American Presidency Series* and in Howell et al. (2000).

A final difficulty results from over-time comparisons. Since not every rater evaluates every statute, it is difficult to know how to compare ratings from different eras. For example, Peterson (2001) attempts to extend Mayhew's (1991) Sweep 1 and Sweep 2 ratings to an earlier era. However, because Peterson and Mayhew use different sources and never rate a common set of statutes, it is impossible to determine whether they employ the same significance threshold and consequently how the two periods compare in terms of the number of significant enactments. It is also impossible to compare legislative productivity and test theories of lawmaking using legislative output both before and after World War II unless we restrict ourselves to one of the few lists that extend across time (for example, Stathis 2003).

To account for the information contained in all of the various ratings and recover a more finely grained measure of legislative significance along with standard errors, we use an item-response model. This provides a means of leveraging all available information, facilitating over-time comparisons, and testing some underlying assumptions (for example, that all raters' assessments are determined by a common latent dimension). A statistical model also provides a means of integrating the rater data described in this section with statute characteristics potentially correlated with the statute's importance (for example, the amount of *Congressional Record* coverage, the number of substitute bills, the session in which it is passed, whether a conference committee was required, and whether the statute was an omnibus bill).

## The Basic Model

The item-response model was developed and is widely used in educational testing research. This is the statistical model we use to integrate the ratings discussed in the prior section. The model is commonly used in educational testing because it models how "items"—commonly questions on a test—discriminate between individuals on the basis of a latent trait such as ability, aptitude, or intelligence (see, for example, Baker's (1977) review, Patz and Junker's (1999) overview, or the textbook accounts provided by Hambleton, Swaminathan and Rogers (1991) and Johnson and Albert (1999)). In other words, the item-response model is a model of how observable responses reflect an underlying latent dimension. The model is specifically designed to address the complication that the parameters of interest are unobserved: the latent ability of the test takers, and the ability of the test questions to discriminate between test takers. Prior work in political science

has recognized the similarities between students answering questions and legislators (or judges) casting votes in applying the models (e.g., Clinton, Jackman, and Rivers 2004; Jackman 2001; Martin and Quinn 2002). Poole and Rosenthal's (1997) seminal work on ideal point estimation uses a very similar model. In applying this model, we exploit a natural comparison between students being rated by examiners and statutes being assessed by congressional chroniclers.

We assume that associated with all legislation is a true and unobservable value of legislative significance. For legislation  $t \in 1 \dots T$  we denote the true latent significance as  $z_t$ . We assume that raters  $i \in 1 \dots N$  of legislative significance (for example, Mayhew's Sweep 1, *Political Science Quarterly*, *New American Nation* series) agree on what constitutes legislative significance and that  $z_t$  entirely determines a statute's true significance relative to other legislation. If the true significance were observable and measurable, then all raters would agree on the relative legislative significance ranking even though they might employ different thresholds in determining the significance of a bill—no rater would think that statute  $i$  is more important than statute  $j$  if  $z_j > z_i$ . Although raters may differ in the methods they use to assess significance, they all fundamentally agree on the unobservable determinants of legislative significance.<sup>16</sup> Each rater "votes" whether legislation  $t$  is significant (1) or insignificant (0) based on the latent trait  $z_t$ . Recall that a "vote" in this context is a mention. Let  $\mathbf{Y}$  be the  $T \times N$  matrix of ratings, with element  $y_{ti}$  containing the dichotomous choice of rater  $n$  with respect to legislation  $t$ .<sup>17</sup> For the period we examine (the 45th through 103rd Congresses),  $T = 37,766$  and  $N = 20$ . Consistent with the observation that rater assessments may differ, we allow raters to imperfectly observe the true significance value  $z$ . Raters rely on the proximate measure  $\mathbf{x}$ , where  $x_{ti} = z_t + \varepsilon_{ti}$ . In other words, for reasons such as the inherent difficulty of assessing legislative significance, differences in selection methodologies, and differences in effort level, the true significance of legislation is only imperfectly observed by raters. Although we assume that rater perceptions are unbiased (i.e.,  $E[\varepsilon] = 0$ ), we allow for the possibility that legislators differ in the precision of their perceptions—perhaps because of different degrees of ambiguity in raters' standards of determining significance. We denote the variance of  $\varepsilon_{ti}$  for rater  $i$  by  $\delta_i^2$ .

<sup>16</sup>This assumption is tested in the next section.

<sup>17</sup>See Jackman (2004) and Quinn (2004) for work extending the framework to ordinal and continuous measures, respectively. An ordinal coding scheme for rater data is possible, but it introduces additional coder discretion. We attempt to avoid coder discretion and consequently used dichotomous measures.

Denoting the distribution of  $\varepsilon$  by  $F(\bullet)$ , these assumptions imply that  $\varepsilon_{ti} \sim F(0, \varepsilon_{ti}/\delta_i)$ . To model the significance of legislation in a statistical model we impose some structure on raters' decisions. Specifically, we assume that a rater's determination of whether a piece of legislation is significant depends only on whether the rater perceives the latent value of the legislation  $x_{ti}$  as exceeding the rater's threshold  $\gamma_i$ . Thus,  $y_{ti} = 1$  if and only if  $x_{ti} > \gamma_i$ . In addition to allowing raters to differentially and imperfectly perceive the true significance of legislation, we also permit each rater to employ a different threshold when determining legislative significance.<sup>18</sup>

Given this rating rule, the probability of rater  $i$  mentioning statute  $t$  is the probability that the rater's observed trait for  $t$  exceeds her threshold. Given the assumption that  $\varepsilon_{ti} \sim F(0, \varepsilon_{ti}/\delta_i)$ , this implies that:

$$\begin{aligned} \Pr(y_{ti} = 1) &= \Pr(x_{ti} > \gamma_i) \\ &= \Pr(z_t + \varepsilon_{ti} > \gamma_i) \\ &= \Pr(\varepsilon_{ti} > \gamma_i - z_t) \\ &= 1 - F((\gamma_i - z_t)/\delta_i) \\ &= F((z_t - \gamma_i)/\delta_i) \end{aligned} \tag{1}$$

The expression  $(z_{ti} - \gamma_i)/\delta_i$  in equation 1 can be reexpressed as:  $\beta_i z_t - \alpha_i$ , where  $\beta_t = 1/\delta_i$  and  $\alpha_i = \gamma_i/\delta_i$ .  $\beta_i$  measures the "discrimination" of rater  $i$ —how much the probability of a rater classifying a piece of legislation as significant changes in response to changes in the latent quality of legislation. For example, if  $\beta_i = 0$ , a rater's determination of legislative significance is entirely invariant with respect to the quality of legislation  $z_t$ . If  $\beta_i$  is very high, then legislative significance is determined almost entirely by the underlying true quality of the legislation. The estimate of  $\alpha_i$  is a function of the threshold used by rater  $i$  in determining significance. High (low) values of  $\alpha_i$  indicate that a rater is more (less) likely to classify the legislation as significant for a given value of  $z_t$ .

Since not every rater evaluates every piece of legislation, we assume that the decision to not rate a piece of legislation is independent of both the significance of the legislation and rater qualities; a rater's failure to evaluate an enactment is uninformative about both the quality of the legislation and the rater. This assumption is unproblematic given that the missing data results from the fact that not all raters evaluate for the entire time period rather

than a conscious choice of the raters. An important question concerns the extent to which ratings of early and recent statutes can be compared if the sources used to assess the statutes differs. The analogous problem in roll-call analysis lies in comparing voting behavior across Congresses (e.g., DW-NOMINATE) or chambers (e.g., Common Space scores; Poole 1988). The solution is to identify "bridging" observations that can be used to orient the two scores. Our bridging observations are the assessments of raters who overlap with nonoverlapping raters.<sup>19</sup> For example, even though Mayhew (1991) and Peterson (2001) rely on different sources and rate different time periods, we use raters who overlap each such as Stathis (2003), Dell and Stathis (1982), and the *American Presidency Series*, to bridge their ratings. The intuition is that since Stathis and Mayhew rate legislation concurrently, we can relate their ratings to one another. We can similarly relate the ratings of Stathis and Peterson. Since the ratings of Mayhew and Peterson are each calibrated to Stathis's ratings, we can therefore relate the assessments of Mayhew and Peterson.

Given this structure, the Bernoulli probability of  $y_{ti}$  is:  $\Pr(y_{ti} = 1 | z_t, \beta_i, \alpha_i) = F(\beta_i z_t - \alpha_i)^{y_{ti}} \times [1 - F(\beta_i z_t - \alpha_i)]^{(1-y_{ti})}$ . Assuming that the ratings are independent across raters conditional on the true latent quality of legislation  $z_t$  and the rater parameters  $\beta_i$  and  $\alpha_i$  implies that  $\Pr(y_{t1}, y_{t2}, \dots, y_{tN} = \mathbf{1} | z_t, \beta_1, \alpha_1) = (F(\beta_1 z_t - \alpha_1)^{y_{t1}} \times [1 - F(\beta_1 z_t - \alpha_1)]^{(1-y_{t1})}) \times (F(\beta_2 z_t - \alpha_2)^{y_{t2}} [1 - F(\beta_2 z_t - \alpha_2)]^{(1-y_{t2})}) \times \dots \times (F(\beta_N z_t - \alpha_N)^{y_{tN}} [1 - F(\beta_N z_t - \alpha_N)]^{(1-y_{tN})}) = \prod_{i=1}^N F(\beta_i z_t - \alpha_i)^{y_{ti}} \times [1 - F(\beta_i z_t - \alpha_i)]^{1-y_{ti}}$ . Further assuming that ratings are independent across legislation yields the likelihood

$$\begin{aligned} L(z, \alpha, \beta) &= \prod_{i=1}^N \prod_{t=1}^T F(\beta_i z_t - \alpha_i)^{y_{ti}} \\ &\quad \times [1 - F(\beta_i z_t - \alpha_i)]^{1-y_{ti}} \end{aligned} \tag{2}$$

where only  $\mathbf{Y}$  is observable. Despite a different motivation for and interpretation of the estimated parameters, the likelihood function in equation 2 is identical to that used in roll-call analysis (see Clinton, Jackman, and Rivers 2004).<sup>20</sup> This equivalence is most evident if we conceptualize legislation as being "voted" on by each rater as to its significance (each of the 37,766 "legislators" cast up to 20 "votes").

<sup>18</sup>One might extend the model by permitting the rater parameters to vary over time according to a specified parameterization rather than assuming a constant ability. For example, the ability of raters to rate contemporaneous enactments may differ from the ability to rate enactments passed before the birth of a rater.

<sup>19</sup>The analogous assumption typically made in roll-call analysis is that members serving in both Congresses or chambers keep the same ideal point (Poole 1988).

<sup>20</sup>See also Jackman (2004), which applies the methodology to graduate admission decisions.



## The Integrated Model

Although the item-response model represents a valuable way of incorporating information from several raters, it faces two limitations. First, there is additional nonrater information available which is plausibly correlated with the significance of statutes. Second, even employing 20 raters results in a relatively small percentage of enactments being rated (3,591 out of 37,766). Absent additional information we have no ability to distinguish between the significance of nonrated legislation.

To address these concerns we use an integrated model that estimates a hierarchical prior for  $\mathbf{z}$ . Instead of assuming that the prior distribution of legislation significance is identical and uninformative for all statutes (e.g.,  $z_t \sim N(0,1)$ ), we assume that the prior distribution of  $\mathbf{z}$  is  $N(\boldsymbol{\mu}_z, \boldsymbol{\tau}^2)$ , where the prior mean  $\boldsymbol{\mu}_z$  is a function of additional information. This provides a means of incorporating information plausibly correlated with legislative significance and determining how such characteristics relate to legislative significance. The importance of this is that it allows us to use the additional information as a means of “bridging” rated and unrated legislation. For example, if we believe that as the number of pages in the *Congressional Record* increases the more likely it is that the legislation is significant we can account for such possibilities. Let  $\mathbf{W}$  denote the  $T \times d$  matrix of covariates. If we assume an additive and linear relationship, then we assume that  $\boldsymbol{\mu}_z = \mathbf{W}\boldsymbol{\kappa} + \boldsymbol{\zeta}$  where  $\boldsymbol{\kappa}$  is a  $d \times 1$  matrix of regression coefficients and  $\boldsymbol{\zeta}$  is an iid error term. Note that this integrated model is almost identical to the exchangeable item-response model discussed by Johnson and Albert (1999) with the important distinction that our model concerns the prior of  $\mathbf{z}$ , not the item parameter priors.<sup>21</sup>

When we impose this linear regression structure and estimate an integrated model, three benefits arise. First, the estimator of  $\mathbf{z}$  is able to “borrow strength” from the additional information contained in the covariate matrix  $\mathbf{W}$ . Since legislation with identical characteristics is assumed to share the same prior mean (although the prior variance permits legislation with identical covariates to differ in significance), the integrated model can distinguish between identically rated legislation on the basis of their characteristics (assuming at least some characteristics covary with legislative significance). Second, the regression specification permits a way to leverage the relationship between the covariates and rated legislation to generate

<sup>21</sup> $\boldsymbol{\tau}^2$  describes the contribution of the regression structure relative to the rater data. As  $\boldsymbol{\tau}^2$  approaches 0 the regression structure becomes more important (at  $\boldsymbol{\tau}^2 = 0$  the structure entirely determines the significance estimate) and as  $\boldsymbol{\tau}^2$  gets large the influence of the regression structure relative to the rating data decreases

significance estimates for the unrated legislation. Since we observe the characteristics of all 37,766 statutes but only 3,591 statutes are mentioned by one of the 20 raters, the relationship between the covariates and the rated statutes essentially generates estimates for the unrated legislation. Third, if we are interested in the covariates of legislative significance, the integrated approach recovers estimates  $\boldsymbol{\kappa}$  that account for the substantial uncertainty in estimates of  $\mathbf{z}$ .

To measure the amount of attention given to the legislation by Congress, we count the total number of pages—including debate and general bill activity (such as committee referral or discharge)—relating to the statute in the *Congressional Record’s* History of Bills and Resolutions. Although imperfect, the page count measure is appealing because it attempts to assess the importance of legislative debate and rhetoric (Besette 1994) and is available for the entire period.<sup>22</sup> We also control for the number of conference committees required, the number of substitute bills associated with the public law, the session of Congress in which the legislation was introduced (Rudalevige 2002), and whether the bill was omnibus (Baumgartner and Jones 2004; Krutz 2001). The regression we estimate for  $\boldsymbol{\mu}_{zi}$  is given by:

$$\begin{aligned} \boldsymbol{\mu}_{zi} = & \kappa_0 + \kappa_1 \log(\text{Pages}) + \kappa_2 I_2 + \kappa_3 I_3 + \kappa_4 I_4 \\ & + \kappa_5 \text{Omnibus} + \kappa_6 \text{Conference Committee} \\ & + \kappa_7 \text{Number of Substitutes} + \zeta_t \end{aligned} \quad (3)$$

where  $I_2$ ,  $I_3$ , and  $I_4$  are indicator variables for whether legislation  $t$  was introduced in the second, third, or fourth sessions.

Since everything except for  $\mathbf{Y}$  in equation 2 is unobserved, the scale and rotation of the parameter space are not identified (Rivers 2003). This means that  $\mathbf{z}$  and  $-\mathbf{z}$  yield equivalent values for the likelihood, as does  $\mathbf{z}$  and  $A \times \mathbf{z}$  where  $A$  is an arbitrary constant. Since we estimate the model using Bayesian methods and since we lack prior information about the discrimination and difficulty of the raters, we adopt uninformative priors for  $\mathbf{z}$ ,  $\boldsymbol{\beta}$ ,  $\boldsymbol{\kappa}$ ,  $\boldsymbol{\tau}$  and  $\boldsymbol{\alpha}$ .<sup>23</sup>

<sup>22</sup>Although deciding whether to include the extended remarks of individual members would seem to be problematic, our decision to include those remarks is inconsequential, since the page count totals in randomly selected years correlate in excess of .95 (see Lapinski 2000 for a discussion).

<sup>23</sup>Specifically,  $z_t \sim N(0,1)$ ,  $\alpha_i \sim N(0,25)$ , and  $\beta_i \sim N(0,25)$ . For the two-dimensional model we follow Jackman’s (2001) suggestion and constrain the parameters to lie on certain dimensions. Unlike the case of Clinton and Meirowitz (2004) in which theory and substantive knowledge provide useable information that is accounted for via informative prior distributions, in this application there is no additional information. If external validation regarding

We employ Bayesian methods because they provide a coherent means of integrating all available information while simultaneously propagating the error through the specification. The intuition underlying the method for the integrated method is as follows. First, generate significance scores for the 3,591 rated using some procedure (e.g., Poole's (2000) Optimal Classification). Second, regress the estimated significance scores on the covariates using equation 3. Third, use the recovered coefficient estimates and the covariate values for the unrated statutes to generate predicted values for the unrated legislation.<sup>24</sup> Employing Bayesian methods provides a coherent and tractable means of simultaneously accomplishing these tasks while accounting for the error in each.<sup>25</sup>

## Significance Measures Compared

Our method yields four sets of estimates: estimates about rater assessments; statute-level estimates of significance; measures of legislative accomplishment resulting from the aggregation of the statute-level estimates; and the regression coefficients of the relationship between legislative characteristics and statute significance. Although of less substantive interest than the statute-level estimates and the resulting measures of legislative accomplishment, rater estimates can be used to assess several potential concerns with the statistical model.

### Comparing Raters

To examine whether raters' assessments can be interpreted as reflecting a common evaluative dimension we determine the ratings' dimensionality. For example, it could be that legislation mentioned in the *American Presidency Series* reflects both the legislation's importance and whether it was related to a president's legislative program. If so, legislation mentioned by this rater might include legislation that is insignificant but central to a president's program (unlikely but plausible), legislation that is significant but

rater quality was available one could incorporate this information through the choice of informative priors.

<sup>24</sup>Strictly speaking, difficulty emerges in the second step because only rated legislation is used to generate the estimates of  $\kappa$  in equation 3 rather than all legislation.

<sup>25</sup>While it is possible to use classical methods, this would require accounting for: the errors generated in step 1, the selection problem resulting from the fact that only rated legislation is used in step 2, and the prediction error in step 3. The sparse data matrix (i.e., the high degree of missing data) may also cause problems (see, for example, Londregan 2000).

unrelated to a president's program, and legislation that is both. We investigate this possibility when we estimate a multidimensional measure of significance  $z$  and constrain the item parameters of the *American Presidency Series* and Mayhew's Sweep 1 and Sweep 2 to lie on different dimensions. Our findings suggest that a single dimension works well. Fitting a two-dimensional model and estimating an additional 37,766 parameters (20 additional item parameters and 37,766 additional statute significance estimates) fails to noticeably improve the fit of the unidimensional model, which correctly predicts 98.6% of the mentions.

Given the high naive classification rate and the very large standard errors for the two-dimensional estimates, we also test whether there is sufficient evidence to conclude that two-dimensional significance estimates do not lie on the 45-degree line. Failure to reject the null hypothesis that the first- and second-dimension estimates are equivalent suggests that there is only a single relevant dimension. The posterior indicates a nonzero difference in only 3% of the sample—a level below conventional significance levels.<sup>26</sup>

Having dismissed via empirical testing the possibility that the raters we use rely on different evaluative dimensions, we turn to the question of the consequences of different rater thresholds. We estimate two parameters for each rater: the extent to which each rater's determinations discriminate between legislation based on the latent significance  $z$  ( $\beta_i = 1/\sigma_i$ ) and the extent to which raters differ in their assessment of legislative significance holding constant the true value of legislative significance ( $\alpha_i = \gamma_i/\sigma_i$ ). Table 2 denotes the estimated discrimination and difficulty parameters for each of the 20 raters used in the analysis and described in the previous section.

Some retrospective raters (for example, the *American Presidency Series*) have comparatively low thresholds, and some contemporaneous raters (Mayhew Sweep 1, for instance) have high thresholds. More importantly, even though contemporaneous raters appear to employ lower thresholds, this is accounted for by the statistical model. A rater with a low threshold is comparatively less informative for distinguishing between significant legislation. The appropriate analogy is to an exam question testing for subject material knowledge: if it is so easy that every student answers it correctly, then it is comparatively less informative in determining student knowledge and distinguishing between student ability than a more difficult

<sup>26</sup>If we restrict the analysis to the 3,591 statutes mentioned at least once (although the statistical basis for this restriction is unclear), the percentage rises to 25%. Some of this rise is attributable to the fact that 169 statutes are mentioned only by either Mayhew or the *American Presidency Series* and therefore are essentially constrained by the required identification restriction to differ.

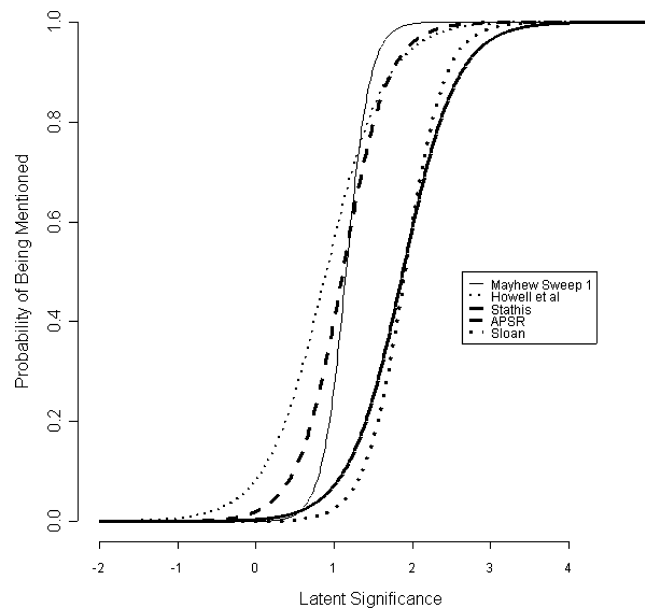
TABLE 2 Rater Ratings

Rater	Retro (Contemp)	$\alpha$ (Std. Err.)	$\beta$ (Std. Err.)
American Presidency Series	R	2.39 (.38)	2.74 (.18)
American Presidency (Johnson)	R	-0.08 (.41)	2.77 (.26)
New American Nation series	R	3.64 (.48)	3.41 (.24)
Stathis	R	2.47 (.56)	4.02 (.26)
Dell & Stathis	R	2.74 (.57)	4.07 (.28)
Mayhew Sweep 2	R	5.38 (1.36)	7.57 (.80)
Peterson Sweep 2	R	4.24 (.53)	3.54 (.30)
Barone	R	4.52 (.52)	3.39 (.34)
Blum	R	5.42 (.87)	4.59 (.73)
Chamberlain	R	4.90 (.62)	3.91 (.39)
Grantham	R	3.93 (.60)	4.09 (.36)
Light	R	4.02 (.70)	4.73 (.43)
Sloan	R	8.32 (1.13)	4.04 (.74)
Reynolds	R	6.30 (.77)	4.64 (.50)
Mayhew Sweep 1	C	6.35 (1.43)	9.37 (1.15)
Peterson Sweep 1	C	3.58 (.53)	3.65 (.29)
Baumgartner & Jones	C	1.51 (.46)	3.24 (.23)
Howell et. al.	C	-2.50 (.75)	5.26 (.43)
PSQ	C	1.52 (.44)	3.06 (.25)
APSR	C	.99 (.42)	2.99 (.22)

question. However, just as hard questions are useful for distinguishing between the smarter set of students, easier questions are useful for distinguishing between students who are less so. Rater heterogeneity is desirable precisely because it allows us to make these determinations.

There is significant variation in the thresholds used by the various raters. Reassuringly, Sloan—who mentions only 16 statutes in his survey of 72 years of lawmaking—employs the highest threshold, and the “A to C” categorization of Howell and his colleagues (2000) employs the lowest threshold. Furthermore, it is not the case that all raters are equally responsive to the true underlying significance of legislation ( $z$ ). The fact that Mayhew’s Sweep 1 has a large estimate of  $\beta$  indicates that Mayhew’s assess-

FIGURE 1 Item Response Curves for Selected Raters



ment is more likely to change in response to changes in the true underlying significance ( $z$ ) than will the assessment of a rater with a lower discrimination parameter, such as the APSR.

These differences can be illustrated by inspecting the item response curves implied by the parameter estimates reported in Table 2. Figure 1 plots the curves for selected raters. The disparity between the item discrimination parameters  $\beta$  is evident in the differences in the slopes. The flatter the line, the less responsive the rater is to the underlying latent significance  $z$  (plotted along the horizontal axis). The item difficulty parameter  $\alpha$  determines the location of the curve in reference to the latent scale. The further to the right a curve is, the higher the threshold employed by the rater; for a given  $z$ , the rater is less likely to rate the statute as significant. In terms of the plotted raters, the measure of Howell and his colleagues (2000) is the least responsive to true significance, whereas Mayhew’s Sweep 1 is the most responsive (steepest).<sup>27</sup>

A related issue concerns inter-rater comparability. For example, since the volumes of the *New American Nation* and *American Presidency Series* are written by different scholars, should they be treated as a single rater or as 20 and 11 raters, respectively? A consequence of employing an item-response model is that this question is empirically

<sup>27</sup>Given the assumption of logistic errors, the item curves are given by:  $\exp(z\beta_i - \alpha_i) / (1 + \exp(z\beta_i - \alpha_i))$  where  $z$  is evaluated over a range of hypothetical values.

resolvable. The question as to whether writers working within a multivolume series organized under specified goals are sufficiently similar can be answered by testing whether the item parameters of the various raters are statistically distinguishable. For example, we treat the Johnson volume of the *American Presidency Series* as a separate rater because it was clear in reading the work that its treatment of legislative accomplishment is systematically different from that of the other volumes. This is confirmed by the estimates reported in Table 2; the item parameters for the Johnson volume differ significantly from the parameters for the other volumes.

### Statute-Level Estimates

Having established that the evaluations are based on a single evaluative dimension and illustrated the consequences of rater heterogeneity on the recovered estimates, we turn now to the statute estimates. Of most interest are the estimates and standard errors of legislative significance  $z$  for the 37,766 laws we analyze. We compare three estimates: the mean rating, the item-response estimator described in an earlier 3, and the integrated estimator also described earlier. Figure 2 graphs the relationship between the recovered estimates. Recall that the scale is defined only relative to the identifying restrictions employed.

Figure 2 illuminates several points. First, the relationship between the mean rating and the nonintegrated estimator in the top graph gives us reason to be suspicious of the assumption that significance is an additive function of the ratings: legislation with higher means are not always estimated to be more significant. This is a consequence of the fact that raters employ different thresholds: being mentioned by three raters with low thresholds may indicate less importance for a statute than being mentioned once by a rater with a very high threshold. Second, the change in estimated significance resulting from the first mention is more than the change resulting from the third mention (for example).

The impact of employing an integrated estimator is made apparent in the bottom graph. The integrated estimator uses the regression structure to distinguish between legislative enactments with identical mean ratings. This is most apparent in the unmentioned legislation. Whereas the normal item-response estimator locates all unmentioned statutes around 0, including statute characteristics distinguishes between the significance of unrated legislative enactments.<sup>28</sup>

<sup>28</sup>The estimates of the nonintegrated estimator are slightly less than 0 because of the identifying restriction employed:  $z \sim N(0, 1)$ .

FIGURE 2 Comparing Estimates of Legislative Significance

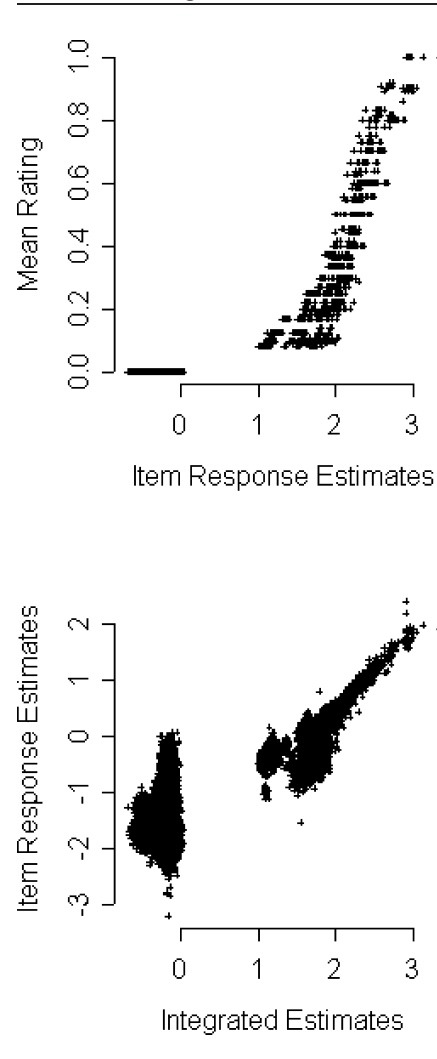


Table 3 provides an example of the power and face validity of our estimates for seven civil rights bills drawn from our list of public laws. Although examining the issue of civil rights is limiting in that most legislating took place from the 1950s onward, it is an illustrative example given its long and important history in our country and the fact that many are familiar with this legislation. Similar patterns are found in other policy areas.

The highest score for a civil rights enactment, not surprisingly, is accorded to the Civil Rights Act of 1964. This bill massively expanded federal power over voting rights, outlawed discrimination in federally funded projects, and gave the attorney general new powers to prosecute state and local authorities who did not desegregate public accommodations. This bill was truly a landmark

**TABLE 3** Examples of Civil Rights Enactments

Title	Congress	Estimate	S.E.
Civil Rights Act of 1964	88 <sup>th</sup> (HR 7152, P.L. 352)	1.860	.428
Voting Rights Act of 1965	89 <sup>th</sup> (S 1564; PL 110)	1.516	.328
Voting Rights Act Amendments of 1982	97 (HR 3112; PL 205)	1.184	.284
Civil Rights Act of 1957	85 (HR 6127; PL 315)	1.102	.254
Civil Rights Act of 1960	86 (HR 8601; PL 449)	1.039	.232
Civil Rights Commission Extension	90 (HR 10805; PL 198)	-.145	.360
Relief of Mrs. Elizabeth G Mason; Civil Rights Commission Extension	88 (HR 3369; PL 152)	-1.491	.598

achievement, and we think it deserves a place atop our list of civil rights legislation. The second highest scoring civil rights bill on our list is the Voting Rights Act of 1965. Building on the 1964 act, this enactment made it illegal to use literacy tests and voter qualifications to screen out voters, brought in federal examiners to supervise registration in states where such requirements had existed, and established criminal penalties for those who violated the act. Below these two very influential landmark laws fall three important enactments passed in 1957, 1960, and 1982. The 1982 enactment, passed during the presidency of Ronald Reagan, was popularly called the Voting Rights Act Amendments of 1982. This law extended the key components of the Voting Rights Act of 1965 for 25 years and gave private parties standing in federal court to sue to overturn any law or procedure resulting in de facto discrimination. The Civil Rights Act of 1957, marshaled by then-senator Lyndon B. Johnson, was a major enactment, but historical accounts tell us that the act that became law was very much watered down. Still, it was the first act of its kind in nearly a century, and therefore even a diluted enactment is quite important. Quite near the 1957 act in terms of significance is the 1960 Civil Rights Act. This act gave federal judges authority to assist people in voting and established criminal penalties for attempting to obstruct voting through violence.

Two other enactments are less significant in comparison to these five enactments. Public Law 90–198, which passed in 1967, extended the life of the Civil Rights Commission by three years and made a nontrivial appropriation of \$2.65 million to support it. Clearly this enactment, though important, is qualitatively different from major legislation such as the 1964 and 1965 acts. The second of these lesser enactments, Public Law 88–152, received a score well below the other laws in our table. This law provided relief for Mrs. Elizabeth Mason for the death of her husband, who died in combat in Belgium during World War II, and extended the Civil Rights Commission for one

year. The law is a one-sentence amendment that extended the Civil Rights Commission without an appropriation and is attached to what appears to be a private relief bill.

These seven enactments provide strong face validity to our measure. The scores associated with these enactments are sensible and provide us with considerable information about the relative importance of the enactments. Our measures account for the fact that the 1964 and 1965 enactments differ from the acts passed in 1957, 1960, and 1982. In contrast, most existing lists treat these enactments as identically “notable” or “significant.”

Since we employ an integrated structure, we also recover an estimate of the extent to which observable covariates covary with legislative quality  $\kappa$ . Table 4 presents these estimated quantities. Recall that with the integrated model the regression parameter estimates  $\kappa$  account for the considerable uncertainty in the estimated significance of legislation  $z$ . To highlight the importance of accounting for the uncertainty present in the statute-level estimates we present the results from running OLS on the posterior mean of the integrated statute estimates. Although the coefficient estimates are (as expected) almost identical, the OLS standard errors for the model ignoring uncertainty in the statute estimates are noticeably smaller.<sup>29</sup>

Although the regression structure used in the estimation does not exhaust the set of possible specifications, we can nonetheless draw several substantive conclusions. Consistent with expectations, legislative significance is increasing in the number of pages. Legislation whose page count differs by one standard deviation (55 pages) differs in predicted significance by 2.52. Important legislation is also more likely to require a conference committee to resolve House and Senate differences, to have several substitute bills, and be an omnibus bill. The purpose

<sup>29</sup>The standard errors are too small because the variance of the error is correlated with the covariates—the most and least significant statutes are measured with the most error.

TABLE 4 Regression Coefficients

Variable	Integrated Coefficient (Std. Error)	OLS
Constant	-3.21 (.20)	-3.21 (.01)
log(Pages)	.63 (.04)	.63 (.004)
Session 2	-.08 (.02)	-.08 (.003)
Session 3	-.27 (.05)	-.27 (.006)
Session 4	-.40 (.15)	-.40 (.02)
Omnibus	.63 (.06)	.62 (.02)
Conference Committee	.49 (.04)	.49 (.006)
Number of Substitutes	.12 (.02)	.12 (.004)
N	37767 $\tau^2 = 2.04 (.24)$	37767 $R^2 = .70$

of the regression structure is not so much to suggest a causal explanation (although correlates of causal explanations could certainly be incorporated into the model), but rather to enable the estimating of the significance of unrated statutes using the plausible correlates of legislative significance.<sup>30</sup> In terms of the relative impact on the estimated legislative significance of being mentioned by Mayhew's Sweep 1 (for example) and statute characteristics such as the length of legislation, note that a one-standard deviation change in the logged page length results in a predicted significance change of .34. In contrast, since the implied threshold for Mayhew's Sweep 2 is 1.41, being mentioned in Sweep 2 implies a significance estimate of at least 1.41.<sup>31</sup>

### Congressional Accomplishment

Although useful for describing the scope of policymaking from 1877 to 1994, the true contribution of the statute

<sup>30</sup>For example, although extensive coverage of a statute in the *Congressional Record* presumably indicates that it relates to an issue that is one of some importance and worthy of congressional attention, verbiage in and of itself does not necessarily result in increased importance.

<sup>31</sup>Using the result that  $\beta_i = 1/\sigma_i = 5.38$  and  $\alpha_i = \gamma_i/\sigma_i = 7.57$  implies that  $\gamma_i = 1.41$ .

significance measure is the opportunity to uncover the determinants of lawmaking. By measuring policy output, we enable investigations into the conditions under which lawmaking does and does not occur and assessments of the impact of institutional or preference changes. Although explanations of *why* institutional change occurs exists (for example, Schickler 2001), work assessing the impact on policy has been limited by the lack of measures comparable to the ones we offer.

To measure legislative accomplishment we follow Mayhew (1991) and count the number of statutes passed each year. This productivity measure represents a *descriptive*, not *normative*, characterization. Although a year in which 15 significant public laws were passed was indeed more productive than a year in which three public laws were passed, the interpretation differs if the 15 statutes were passed in a year in which there were 40 opportunities for legislating and the three public laws represented the only opportunities for legislating change in the year they were passed. This is the "denominator" problem noted by Mayhew (1991).<sup>32</sup>

One difficulty raised by our measure is determining how to aggregate and summarize the statute-level estimates. Existing dichotomous measures such as those of Mayhew (1991) and Stathis (2003) summarize productivity by counting the number of significant statutes enacted, treating every statute as being of equal significance. Since we estimate the significance for every public law, the appropriate measure for us is more ambiguous. Consequently, we follow Baumgartner and Jones (2002, 2004) and construct a measure focusing on statutes whose significance rank exceeds an exogenous threshold. Since the threshold is arbitrary—it is unclear whether analyzing the top 500 (of 37,766) enacted statutes provides a better measure of accomplishment than analyzing the top 2,000—we examine the consequences of several thresholds.

The number of statutes passed is most similar to existing measures of accomplishment (for example, Baumgartner and Jones 2002, 2004; Mayhew 1991; Stathis 2003). Although it does capture differences in the quantity of legislation passed, this number cannot account for differences in quality. In counting the number of statutes passed that lie in the top 500 (for example), differences in the significance of enacted legislation in the top 500 are ignored and all legislation in the top set is treated equally. A second measure which arguably account for both quantity and quality considerations results if we sum the significance of every statute surpassing a given threshold. All else equal, Congresses that pass more legislation

<sup>32</sup>Since the agenda is endogenous, it is unclear whether an appropriate denominator is even possible.

are judged to have accomplished more, and the higher the significance of the statutes passing the threshold, the higher the measure. Although in principle superior, the second measure correlates with the count measure at very high levels given the small significance differences among top-rated statutes.

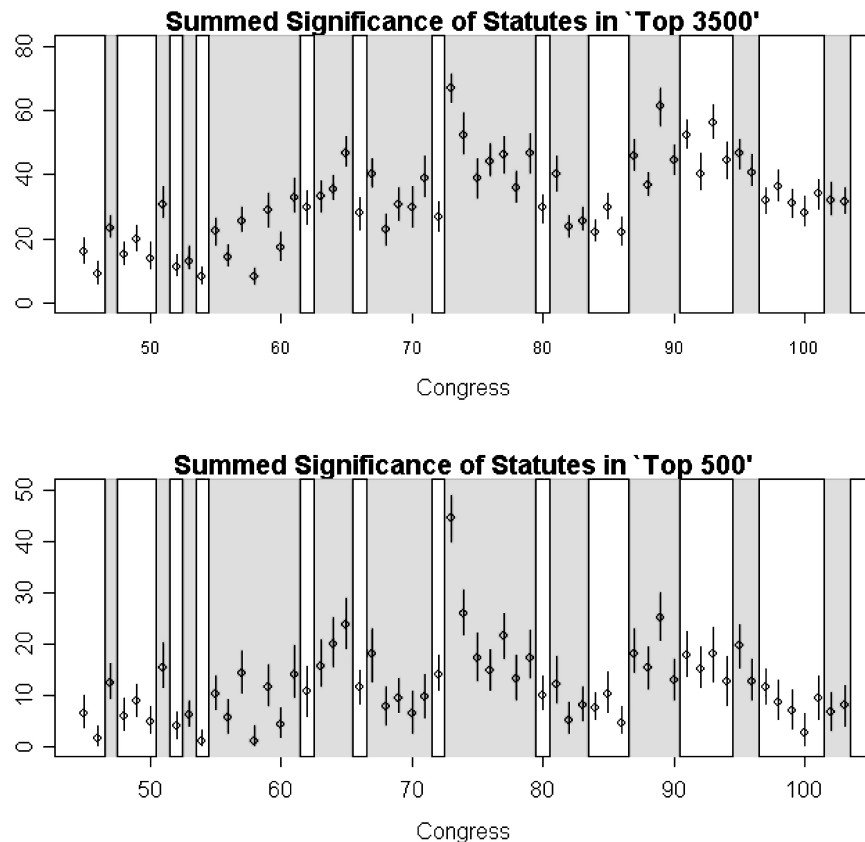
One consequence of our statistical method is that our measure of accomplishment can account for the error in the statute-level estimates. There are two sources of estimation error. First, since the statute-level estimates being summed are measured with error, the summation will clearly contain error. The posterior mean is not perfectly measured, and a measure of accomplishment should reflect this imprecision. Second, because the significance of each statute is measured with error, the number (and identity) of statutes passing a given threshold during a particular Congress is subject to error. Each statute has a probability of lying in the top set that should be accounted for.

A consequence of using a Bayesian estimator is that it is straightforward to calculate the posterior for any

function of estimated parameters. Consequently, when summing the significance of the number of top statutes enacted, we account for both uncertainty in the probability that a statute lies in the top set as well as uncertainty in each statute's significance estimate. Figure 3 graphs the summed significance of the statutes passed by each Congress among the 3,500 and 500 most notable statutes passed from 1877 to 1994 and the 95% highest posterior density region for each.

Regardless of the exogenous threshold used, the measure exhibits considerable similarity (the measure using the top 3,500 and top 500 statutes correlate at .87). Consequently, at least for nonnegligible thresholds, there is no evidence that the threshold used to define accomplishment matters. Reassuringly, the trend graphed in Figure 3 possesses strong face validity—the frenzied lawmaking of the “New Deal” and the “Great Society” are easily evident. For context, the shaded areas denote Congresses with unified party control of government. Although this measure, like all measures, is imperfect, the fact that it is responsive to both the quantity and quality of enacted legislation

**FIGURE 3 Congressional Accomplishment, 1887–1994**



represents a notable advance over existing measures. Despite a healthy correlation with existing measures (for example, the correlations between the top 500 estimate plotted in Figure 3 and Mayhew's Sweep 1 and Sweep 2 are .94 and .94, respectively, and the top 3500 estimates correlate at .69 and .74, respectively), the ability to estimate statute-level measures of significance can lead to important improvements in how we characterize legislative accomplishment. In addition to covering a longer time period than most ratings, our measure of accomplishment also accounts for statute-level variation in the significance of enacted statutes.

## Conclusion

Studying policymaking—what we refer to as legislative accomplishment—is critical if we are to assess how well a political system is working. Unfortunately, theoretical progress in studying policymaking has not been matched by comparable empirical progress, and we believe that the lack of empirical progress can be attributed to the relative lack of attention paid to defining measures and deriving appropriate estimators.

We present a new and more expansive measure of legislative significance that exploits existing and new data sources, including the assessments of legislative importance constructed by historians, policy experts, and the media and information on the characteristics of legislation. This measure uses recent statistical advances in item response modeling to address important limitations of existing measures. The end result is a significance score for every legislative enactment between 1877 and 1993—37,766 enactments in all as well as an assessment of the uncertainty of the legislative significance estimates. We also demonstrate how these scores can be aggregated to yield congressional (or yearly) legislative accomplishment scores and how these scores might be used to select important enactments.

Although our larger project is oriented toward understanding lawmaking in the United States, it is important to reiterate the generality of the estimation method we present. An analogous procedure could be used to identify the relative importance of presidential executive orders, bureaucratic rule changes, and Supreme Court decisions. Furthermore, nothing restricts the method's application to policymaking in the United States; the accomplishments of any institution can, in principle, be assessed using the method we outline. All that is required is a set of assessments by chroniclers of the institution which are then analyzed using an item-response model.

## Appendix

### New American Nation

Arthur S. Link's *Woodrow Wilson and the Progressive Era, 1910–1917* (1954), George E. Mowry's *The Era of Theodore Roosevelt, 1900–1912* (1958), Harold U. Faulkner's *Politics, Reform, and Expansion, 1890–1900* (1959), Eric F. Goldman's *The Crucial Decade and After: America, 1945–1960* (1960), John D. Hick's *Republican Ascendancy, 1921–1933* (1960), William E. Leuchtenburg's *Franklin D. Roosevelt and the New Deal, 1932–1940* (1963), Russell A. Buchanan's *The United States and World War II*, volumes 1 and 2 (1964), John A. Garratty's *The New Commonwealth, 1877–1890* (1968), Allen J. Matusow's *The Unraveling of America: A History of Liberalism in the 1960s* (1984), and Robert H. Ferrell's *Woodrow Wilson and World War I, 1917–1921* (1985).

### American Presidency Series

Paolo E. Coletta's *The Presidency of William Howard Taft* (1973), Eugene P. Trani and David L. Wilson's *The Presidency of Warren G. Harding* (1977), Lewis L. Gould's *The Presidency of William McKinley* (1980) and *The Presidency of Theodore Roosevelt* (1991), Justus D. Doenecke's *The Presidencies of James A. Garfield and Chester A. Arthur* (1981), Donald R. McCoy's *The Presidency of Harry S. Truman* (1984), Martin L. Fausold's *The Presidency of Herbert C. Hoover* (1985), Homer B. Socolofsky and Allan B. Spetter's *The Presidency of Benjamin Harrison* (1987), Ari Hoogenboom's *The Presidency of Rutherford B. Hayes* (1988), Richard E. Welch Jr.'s *The Presidencies of Grover Cleveland* (1988), Chester J. Pach Jr. and Elmo Richardson's *The Presidency of Dwight D. Eisenhower* (1991), James N. Giglio's *The Presidency of John F. Kennedy* (1991), Burton I. Kaufman's *The Presidency of James Earl Carter Jr.* (1993), Kendrick A. Clements's *The Presidency of Woodrow Wilson* (1992), John Robert Greene's *The Presidency of Gerald R. Ford* (1995) and *The Presidency of George Bush* (2000), Robert H. Ferrell's *The Presidency of Calvin Coolidge* (1998), Melvin Small's *The Presidency of Richard Nixon* (1999), and George T. McJimmsey's *The Presidency of Franklin Delano Roosevelt* (2000).

We supplement the series by including Michael Schaller's *Reckoning with Reagan: American and Its President in the 1980s* (1992).

**Selected Public Affairs Texts.** Dewey W. Grantham's *Recent America: The United States Since 1945* (1987), Michael Barone's *Our Country: The Shaping of America from Roosevelt to Reagan* (1990), and John Morton Blum's



*Years of Discord: American Politics and Society, 1961–1974* (1991).

## References

- Baker, Frank B. 1977. "Advances in Item Analysis." *Review of Educational Research* 47(1):151–78.
- Barone, Michael. 1990. *Our Country: The Shaping of America from Roosevelt to Reagan*. New York: Free Press.
- Baumgartner, Frank R., and Bryan D. Jones. 2002. *Policy Dynamics*. Chicago: University of Chicago Press.
- Baumgartner, Frank R., and Bryan D. Jones. 2004. "The Evolution of American Government: Information, Attention, and Policy Punctuations." Unpublished paper. Pennsylvania State University.
- Bessette, Joseph. 1994. *The Mild Voice of Reason: Deliberative Democracy and American National Government*. Chicago: University of Chicago Press.
- Binder, Sarah. 1999. "The Dynamics of Legislative Gridlock, 1947–1996." *American Political Science Review* 93(3):519–33.
- Binder, Sarah. 2003. *Causes and Consequences of Legislative Gridlock*. Washington: Brookings Institution Press.
- Blum, John Morton. 1991. *Years of Discord: American Politics and Society, 1961–1974*. New York: W. W. Norton.
- Bornet, Vaughn Davis. 1983. *The American Presidency of Lyndon B. Johnson*. Lawrence: University Press of Kansas.
- Boyer, Paul S., and Clifford E. Clark, Jr. *The Enduring Vision*, 5<sup>th</sup> ed. Boston: Houghton Mifflin.
- Buchanan, Russell A. 1964. *The United States and World War II*, vols. 1 and 2. New York: Harper & Row.
- Cameron, Charles. 2000. *Veto Bargaining: Presidents and the Politics of Negative Power*. New York: Columbia University Press.
- Chamberlain, Lawrence. 1946. *The President, Congress, and Legislation*. New York: Columbia University Press.
- Clements, Kendrick A. 1992. *The Presidency of Woodrow Wilson*. Lawrence: University Press of Kansas.
- Clinton, Joshua D., Simon Jackman, and Douglas Rivers. 2004. "The Statistical Analysis of Legislative Behavior: A Unified Approach." *American Political Science Review* 98(2):355–70.
- Clinton, Joshua D., and Adam Meirowitz. 2004. "Testing Accounts of Legislative Strategic Voting: The Compromise of 1790." *American Journal of Political Science* 48(4):675–89.
- Coleman, John J. 1999. "Unified Government, Divided Government, and Party Responsiveness." *American Political Science Review* 93(4):821–35.
- Coletta, Paolo E. 1973. *The Presidency of William Howard Taft*. Lawrence: University Press of Kansas.
- Cox, Gary W., and Mathew D. McCubbins. 2004. "Setting the Agenda: Responsible Party Government in the U.S. House of Representatives." Unpublished Manuscript. University of California San Diego.
- Dell, Christopher, and Stephen W. Stathis. 1982. "Major Acts of Congress and Treaties Approved by the Senate, 1789–1980." Congressional Research Service Report 82–156 GOV.
- Doenecke, Justus D. 1981. *The Presidencies of James A. Garfield and Chester A. Arthur*. Lawrence: Regents Press of Kansas.
- Epstein, David, and Sharyn O'Halloran. 1999. *Delegating Powers: A Transaction Cost Politics Approach to Policy Making Under Separate Powers*. New York: Cambridge University Press.
- Erikson, Robert S., Michael B. MacKuen, and James A. Stimson. 2002. *The Macro Polity*. New York: Cambridge University Press.
- Faulkner, Harold U. 1959. *Politics, Reform, and Expansion, 1890–1900*. New York: Harper & Row.
- Fausold, Martin L. 1985. *The Presidency of Herbert C. Hoover*. Lawrence: University Press of Kansas.
- Ferejohn, John A. 1974. *Pork Barrel Politics: Rivers and Harbors Legislation, 1947–1968*. Palo Alto: Stanford University Press.
- Ferrell, Robert H. 1985. *Woodrow Wilson and World War I, 1917–1921*. New York: Harper & Row.
- Ferrell, Robert H. 1998. *The Presidency of Calvin Coolidge*. Lawrence: University Press of Kansas.
- Garratty, John A. 1968. *The New Commonwealth, 1877–1890*. New York: Harper & Row.
- Giglio, James N. 1991. *The Presidency of John F. Kennedy*. Lawrence: University Press of Kansas.
- Goldman, Eric F. 1960. *The Crucial Decade and After: America, 1945–1960*. New York: Random House.
- Gould, Lewis L. 1980. *The Presidency of William McKinley*. Lawrence: University Press of Kansas.
- Gould, Lewis L. 1991. *The Presidency of Theodore Roosevelt*. Lawrence: University Press of Kansas.
- Grantham, Dewey W. 1987. *Recent America: The United States Since 1945*. Arlington Heights, IL: Harlan Davidson.
- Greene, John Robert. 1995. *The Presidency of Gerald R. Ford*. Lawrence: University Press of Kansas.
- Greene, John Robert. 2000. *The Presidency of George Bush*. Lawrence: University Press of Kansas.
- Hambleton, Ronald K., H. Swaminathan, and H. Jane Rogers. 1991. *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications.
- Hick, John D. 1960. *Republican Ascendancy, 1921–1933*. New York: Harper & Row.
- Hoogenboom, Ari. 1988. *The Presidency of Rutherford B. Hayes*. Lawrence: University Press of Kansas.
- Howell, William, E. Scott Adler, Charles Cameron, and Charles Riemann. 2000. "Divided Government and the Legislative Productivity of Congress, 1945–1994." *Legislative Studies Quarterly* 25(2):285–312.
- Jackman, Simon. 2001. "Multidimensional Analysis of Roll Call Data via Bayesian Simulation: Identification, Estimation, Inference, and Model Checking." *Political Analysis* 9(3):227–41.
- Jackman, Simon. 2004. "What Do We Learn from Graduate Admissions Committees? A Multiple-Rater, Latent Variable Model with Incomplete Discrete and Continuous Indicators." *Political Analysis* 12(4):400–24.
- Johnson, Valen, and James Albert. 1999. *Ordinal Data Modeling*. New York: Springer-Verlag.

- Katznelson, Ira, and John S. Lapinski. 2005. "Studying Policy Content and Legislative Behavior." In *Macropolitics of Congress*, ed. E. Scott Adler and John S. Lapinski. Princeton: Princeton University Press. xx–xx.
- Q1 Kaufman, Burton I. 1993. *The Presidency of James Earl Carter Jr.* Lawrence: University Press of Kansas.
- Kelly, Sean. 1993. "Divided We Govern? A Reassessment." *Polity* 25(Spring):475–84.
- Krehbiel, Keith. 1998. *Pivotal Politics: A Theory of U.S. Lawmaking*. Chicago: University of Chicago Press.
- Krutz, Glen. 2001. *Hitching a Ride: Omnibus Legislation in the U.S. Congress*. Columbus: The Ohio State University Press.
- Lapinski, John S. 2000. *Representation and Reform: A Congress Centered Approach to American Political Development*. Ph.D. Dissertation. Columbia University.
- Leuchtenburg, William E. 1963. *Franklin D. Roosevelt and the New Deal, 1932–1940*. New York: Harper & Row.
- Light, Paul. 2002. *Government's Greatest Achievements: From Civil Rights to Homeland Defense*. Washington: Brookings Institution Press.
- Link, Arthur S. 1954. *Woodrow Wilson and the Progressive Era, 1910–1917*. New York: Harper & Row.
- Londregan, John. 2000. "Estimating Legislator's Preferred Points." *Political Analysis* 8(1):35–56.
- Martin, Andrew D., and Kevin M. Quinn. 2002. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999." *Political Analysis* 10(2):134–53.
- Matusow, Allen J. 1984. *The Unraveling of America: A History of Liberalism in the 1960s*. New York: Harper & Row.
- Mayhew, David. 1991. *Divided We Govern*. New Haven: Yale University Press.
- Mayhew, David. 1993. "Let's Stick with the Longer List." *Polity* 25(Spring):485–88.
- Mayhew, David. 2000. *America's Congress: Actions in the Public Sphere, James Madison Through Newt Gingrich*. New Haven: Yale University Press.
- McCarty, Nolan, Keith Poole, and Howard Rosenthal. 2005. *Polarized America: The Dance of Political Ideology*. Unpublished manuscript. Princeton University.
- McCoy, Donald R. 1984. *The Presidency of Harry S. Truman*. Lawrence: University Press of Kansas.
- McJimsey, George T. 2000. *The Presidency of Franklin Delano Roosevelt*. Lawrence: University Press of Kansas.
- Mowry, George E. 1958. *The Era of Theodore Roosevelt, 1900–1912*. New York: Harper & Row.
- Pach, Chester J., Jr., and Elmo Richardson. 1991. *The Presidency of Dwight D. Eisenhower*. Lawrence: University Press of Kansas.
- Patz, Richard J., and Brian W. Junker. 1999. "A Straightforward Approach to Markov Chain Monte Carlo Methods for Item Response Models." *Journal of Educational and Behavioral Sciences* 24(2):146–78.
- Peterson, Eric. 2001. "Is It Science Yet? Replicating and Validating the *Divided We Govern* List of Important Statutes." Presented to the annual meeting of the Midwest Political Science Association.
- Poole, Keith. 1988. "Recent Developments in Analytical Models of Voting in the U.S. Congress." *Legislative Studies Quarterly* 13(1):117–33.
- Poole, Keith. 2000. "Non-Parametric Unfolding of Binary Choice Data." *Political Analysis* 8(3):211–37.
- Poole, Keith T., and Howard Rosenthal. 1997. *Congress: A Political-Economic History of Roll Call Voting*. New York: Oxford University Press.
- Quinn, Kevin M. 2004. "Bayesian Factor Analysis for Mixed Ordinal and Continuous Responses." *Political Analysis* 12(4):338–53.
- Reynolds, John. 1995. "Divided We Govern: Research Note." Unpublished paper. University of Texas, San Antonio.
- Rivers, Douglas. 2003. "Identification of Multidimensional Spatial Voting." Unpublished paper. Stanford University.
- Rudalevige, Andrew. 2002. *Managing the President's Program: Presidential Leadership and Legislative Policy*. Princeton: Princeton University Press.
- Schaller, Michael. 1992. *Reckoning with Reagan: America and Its President in the 1980s*. New York: Oxford University Press.
- Schickler, Eric. 2001. *Disjointed Pluralism: Institutional Innovation and the Development of the U.S. Congress*. Princeton: Princeton University Press.
- Sloan, Irving. 1984. *American Landmark Legislation*, 2nd series. Dobbs Ferry, NY: Oceana Publications.
- Small, Melvin. 1999. *The Presidency of Richard Nixon*. Lawrence: University Press of Kansas.
- Socolofsky, Homer B., and Allan B. Spetter. 1987. *The Presidency of Benjamin Harrison*. Lawrence: University Press of Kansas.
- Stathis, Stephen W. 2003. *Landmark Legislation: 1774–2002*. Washington: Congressional Quarterly Press.
- Trani, Eugene P., and David L. Wilson. 1977. *The Presidency of Warren G. Harding*. Lawrence: University Press of Kansas.
- Welch, Richard E. 1988. *The Presidencies of Grover Cleveland*. Lawrence: University Press of Kansas.