Chapter 24

Measuring Significant Legislation, 1877 to 1948

Joshua D. Clinton and John S. Lapinski

The authors would like to thank David Brady and Mathew McCubbins, Ted Brader, Brandice Canes-Wrone, Charles Cameron, Gary Cox, Robert Erikson, Alan Gerber, Ira Katznelson, Jeffery Jenkins, Ken Scheve, David Mayhew, Nolan McCarty, Robert Mickey, Keith Poole, Rose Razaghian, Howard Rosenthal; Doug Rivers, Eric Schickler, Rob Vanhowelling, Greg Wawro, and participants at the Center for the Study of Democratic Politics at Princeton University; participants at the History of Congress Conference, Stanford University; and participants in the Columbia, Northwestern and University of Michigan workshops for helpful criticisms of previous versions. We also would like to thank Eldon Porter and Melanie Springer for research assistance.

Introduction

In this paper, we provide a measure of legislative output that can be used to characterize outcomes of the lawmaking process across time. We summarize our efforts to identify the 21,741 public statutes passed between 1877 (45th Congress) and 1948 (80th Congress).[1] In so doing we collect both statute-level descriptive information and the assessments of several scholars and chroniclers who identify enactments of note during this time period. Using this massive data collection we can characterize and assess congressional lawmaking activity across a period spanning the Populist, Progressive, and New Deal Eras as well as the two world wars.

We use a statistical model to integrate the collected information and estimate a statute-level measure of legislative output. Rather than relying on a single scholar's assessment of statutes' significance, we use an item response model to aggregate the information contained in the nine assessments we collect. The statistical model provides the means for integrating the information in these assessments with statute characteristics—for example, the amount of attention the legislation received in Congress. One advantage of incorporating several sources of information is that because our estimates utilize more information than existing determinations, we can make comparisons that are precluded by coarser characterizations. For example, we can

[1] The *United States Public Statutes at Large* lists 21,758 public statutes passed during this period, including amendments to the Constitution. Of these statutes, we were able to identify 21,741 that became public laws. For seventeen statutes listed in *Public Statutes at Large* there is no corresponding information in the *Congressional Record*'s *History of Bills and Resolutions* and they were dropped from the analysis. We are nearly finished extending the dataset through 2002 (107th Congress), covering a total of 39,630 public statutes.

compare the relative notability of the Pure Foods Act of 1906 and the Social Security Act of 1934 instead of treating the two statutes as equally significant. [2]

We develop the measure in several steps. First, we discuss the difficulties associated with defining "significant legislation" to help situate the interpretation of the results of our analysis. Having outlined the importance and the limits of the concept, we then describe the ratings and data we use to estimate legislative significance. The third section argues for a systematic means of combining the many existing assessments of legislative significance and section four presents the item response model we use to do so. Section five assesses the resulting statute-level estimates and characterizes lawmaking in the pre-World War II era. We conclude by discussing possible extensions of the measure and the ways in the measure can be used to address fundamental questions.

Defining Significant Legislation

Defining precisely what is meant by the term "significant legislation" (or "important" or "landmark" legislation) is a very difficult task. Most scholars use established lists of important or landmark enactments for their empirical study, noting that the controversial enterprise of

[2] For example, in his work on critical elections and lawmaking, Ginsberg (1972, 1973, 1976) notes that "clearly, some statutes are more important than others, and statutes vary considerably in the magnitude of the policy objectives they attempt to achieve." However, he does not attempt to differentiate statutes by importance and he concedes that "this analysis, which weights each statute equally, ignores important qualitative differences among statutes…It is, however, difficult to conceive of any procedure which would permit systematic qualitative comparisons of the importance of laws" (Ginsberg 1976, 45). Personal communications with Ginsberg have confirmed that the individual-level data from this project are not available.

defining significant legislation and operationalizing this concept is not theirs.[3] Although several lists of "significant legislation" exist, the bases on which the determinations of significance are made are largely unclear. Cameron (2000, 38) aptly characterizes the situation faced by every scholar seeking to construct such a measure when he notes that the pathways to significance "are so many and so varied that a tired but *apropos* phrase is irresistible: legislative significance is hard to define but easy to recognize."[4]

An attempt to precisely enumerate the conditions under which a statute may rightfully be considered significant quickly leads to a long list of conditions that prove difficult, if not impossible, to operationalize (see, for example, discussions by Chamberlain 1946, Mayhew 1991, Binder 1999, 2003; Coleman 1999; Edwards et al. 1997; Howell et al. 2000; Light 2002; Peterson 2001, Stathis 2003). Furthermore, even if it were possible to specify what constitutes an "important" statute, it is likely impossible to making a determination based on such standards due to the difficult of ascertaining whether the specified conditions are satisfied. The difficulty, if not futility, of such an exercise leads us to define significant legislation explicitly in terms of that which is notable. Consequently, we refer to a statute or constitutional amendment as "significant" if it is identified as noteworthy by a reputable chronicler of the congressional session.[5] Likewise, legislation that was enacted but not mentioned during the period by a rater is

---

[3] For example, Krehbiel (1998, 57) notes: "The dependent variable in the analysis are all forms of Congress-to-Congress *changes in legislative productivity*. Measurement of legislative productivity is a subject of an ongoing controversy about which a neutral stance is adopted here."

[4] Cameron (2000, 38) lists some of these pathways as "changing people's lives, redistributing wealth, creating or destroying rights, limning partisan differences."

[5] Note that Mayhew (1991) sometimes substitutes the word "notable" for "significant."

treated as being deemed “insignificant.” Although our measure is more accurately described as measure of notable legislation, we follow others and refer to "notable" legislation as "significant" legislation.

Focusing on notable legislation effectively compiles a list of high-stakes legislation. Although “high-stakes” legislation includes innovative and consequential legislation, it may also include controversial legislation that is neither innovative nor consequential. However, insofar as political conflict is more likely on innovative and consequential legislation, there are probably very few banal or inconsequential statutes that are notable. Although our measure is necessarily because of these caveats, it is difficult to conceptualize a less imperfect measure.

Measuring Legislative Significance

An evaluation of legislation is the result of the work of a “rater”—effectively the criteria used to determine whether a statute is noteworthy. A single individual may count as more than one rater if they use different standards in assessing the legislation. For example, Peterson (2001) uses different sources and standards to produce his Sweep One and Sweep Two lists. Because our measure of legislative significance depends on whether the legislation is noteworthy, it is absolutely critical to identify appropriate raters.

We rely on nine raters who attempt to identify “important” legislation for the period 1877 to 1948. Following Mayhew, we classify the nine raters according to the timing of the assessment: contemporaneous raters assess legislation at or near the time of enactment and retrospective raters rely on hindsight to determine noteworthiness given both prior and subsequent events. Some determinations combine both retrospective and contemporaneous assessments. Although we can certainly compare the assessments that result from the three perspectives to examine how the perceived importance of individual statutes changes over time,

our primary interest is in using the evaluations to construct a measure of legislative significance that reflects both contemporaneous and retrospective perspectives.

We use three contemporaneous raters for the period 1877 to 1948. The legislative wrap-ups of the *American Political Science Review* (*APSR*) and the *Political Science Quarterly* (*PSQ*) provide two sets of ratings. The wrap-up for each journal consists of a summary of the year's proceedings in Congress. The *APSR* covers the years 1919 to 1948 in a section entitled "The Legislative Record," and *PSQ* covers the years 1889 to 1925 in its annual "Record of Public Events." Each wrap-up was read by two coders, and all of the legislation mentioned in the wrap-up was recorded.[6] Peterson's (2001) Sweep One provides the third contemporaneous assessment. Although his replication of Mayhew's coding technique is inexact, Peterson identifies 213 statutes passed from 1881 to 1945 as significant according to contemporaneous sources.[7]

We also use five retrospective evaluations. Peterson's (2001) replication of Mayhew's Sweep Two for the period 1881 to 1947 using several policy area histories constitutes one rating. The nine period-specific volumes of the New American Nation Series provide a second list of significant legislation.[8] These period-specific historical accounts of U.S. political history were

---

[6] Note that sometimes the authors of the records of political events were the same for *PSQ* and *APSR*. For example, Lindsay Rogers wrote the records for the *APSR* from 1919 to 1923 and also wrote the reviews for *PSQ* for 1923.

[7] Note that Peterson's sources vary across Congresses. For example, to identify legislation for the 47th Congress, he uses three sources that, in truth, do not meet Mayhew's criterion. See Peterson (2001) for a complete listing of his sources, which are too numerous to list here.

[8] The series includes Garratty, *The New Commonwealth, 1877-1890* (1968); Faulkner, *Politics, Reform, and Expansion, 1890-1900* (1959); Mowry, *The Era of Theodore Roosevelt, 1900-1912*

read twice in their entirety and content-coded them for mentions of public enactments and policy changes. If a policy change was mentioned but the title of the statute associated with it was not, we searched the index of the *History of Bills and Resolutions* in the *Congressional Record* and the index of *Public Statutes at Large* for the associated statute. When necessary, we also searched the *Congressional Universe* and *Academic Universe* databases, as well as ProQuest's historical archives for the *New York Times*, *Washington Post*, and *Wall Street Journal* using available information and clues (such as the date of enactment or legislative activity) to identify the public statute (if any) associated with the policy.

A third source of retrospective ratings comes from Chamberlain's *The President, Congress, and Legislation* (1946). In chronicling policymaking across ten issue spaces from pre-1877 to 1940, Chamberlain identifies 105 monumentally important enactments. Reynolds's (1995) "mini-research report" culls the *Enduring Vision* textbook for mentions of important legislation to test for the effect of divided government on lawmaking before World War II and he identifies 84 enactments for the period 1877 to 1948. A final source is Sloan's *American Landmark Legislation* (1984). Sloan compiles a selective list of laws based on two factors: the important national significance they had at the time Congress passed them, and the bills' lasting effect on one dimension or another of American life.

(1958); Link, *Woodrow Wilson and the Progressive Era, 1910-1917* (1954); Ferrell, *Woodrow Wilson and World War I, 1917-1921* (1985); Hick, *Republican Ascendancy, 1921-1933* (1960); Leuchtenburg, *Franklin D. Roosevelt and the New Deal, 1932-1940* (1963); Buchanan, *The United States and World War II*, vol. 1 (1964); and Buchanan, *The United States and World War II*, vol. 2 (1964).

The third type of rating, the hybrid of contemporaneous and retrospective sources, is based on previously compiled lists of important enactments that are drawn from both types of sources. We use one hybrid rater in this chapter: Congressional Research Service employees Dell and Stathis (1982), chronicle two hundred years of congressional effort—from the 1st Congress in 1789 to the 96th in 1980—and identify 271 significant public enactments during the period from 1877 to 1948.

In addition to these nine assessments, we also measure several characteristics of the statutes themselves. We collect information on when the legislation was introduced to control for the possibility that legislation introduced early in a session is more likely to be significant than legislation introduced later in the session (Neustadt 2001; Rudalevige 2002).[9] To proxy for Congressional attention to the legislation we count of the total number of pages—including debate and general bill activity (such as committee referral or discharge)—relating to the statute in the *History of Bills and Resolutions* of the *Congressional Record*. Although imperfect, the page count measure–which ranges from ranges from 4 to 4,024 pages--is available for the entire period and attempts to assess the importance of legislative debate and rhetoric (Bessette 1994). Our decision to include members' extended remarks is inconsequential; the page counts in randomly selected years correlate in excess of .95.

In every year we consider, between five and nine raters assess the enactments. Not surprisingly, the vast majority of legislation—nearly 96 percent (20,788 of 21,741)—is unmentioned by any rater. Furthermore, there is significant variation in the number of mentioned statutes due to the extensiveness of the rating criteria employed. For example, although both the

[9] The norm is two congressional sessions; however, several Congresses have had extra "special" sessions.

New American Nation Series and Sloan's *American Landmark Legislation* implicitly consider every public statute between 1877 and 1948, the former deems 305 enactments to be significant whereas the latter lists only 16.

Further evidence of rater heterogeneity is revealed by examining ratings from the 66th, 67th, and 68th Congresses. All nine raters evaluate the 1,832 enactments of this period: 1,729 statutes are unmentioned by any rater, 49 statutes are mentioned by one rater, 26 by two raters, 8 by three raters, 11 by four raters, 1 by five raters, 4 by six raters, and 4 by seven raters. No statute is identified by more than seven raters. The four statutes rated as significant by seven of the nine raters are: the National Prohibition Act of 1919 (not mentioned by Sloan or Chamberlain), the Transportation Act of 1920 (not mentioned by Sloan or Reynolds), the Immigration Act of 1921 (not mentioned by Sloan or Reynolds), and the Immigration Act of 1924 (not mentioned by Sloan or Peterson, Sweep Two). Although heterogeneity is clearly evident, unanimously agreement exists in periods covered by fewer raters. For example, during the years when only eight raters are active, all eight mention the Securities Act of 1934, the Sherman Anti-Trust Act of 1890, the Pure Foods Act of 1906, and the Federal Trade Commission Act of 1906.

The above discussion demonstrates that any measure of legislative significance must confront the issue of rater heterogeneity: How should we interpret rater disagreement, and how should a measure of significance account for this heterogeneity?

There are three interpretations of rater disagreement. First, raters may differ because they use different criteria. For example, one rater may identify only the most important legislation sponsored by Southern legislators, whereas other lists attempt to identify all important

enactments. In this case, the selection criteria are not the same. We consciously selected raters to minimize this possibility; the raters we use all aim either explicitly to identify significant legislation or the major legislative events of the year or time period.

Second, raters may employ different standards—that is, the threshold that establishes legislative significance may differ among raters. For example, in its annual wrap-up the *American Political Science Review* may be more likely to identify a statute as an important policy change than a rater such as Sloan, who surveys an entire time period before making a determination. In other words, two raters may agree as to the dimension of interest but differ on the level that establishes significance.

A third, but related, possible difference is the issue of rater consistency across time. Because not every rater rates every statute, it is difficult to know how to compare ratings from different eras. For example, although Peterson attempts to extend Mayhew's Sweep One and Sweep Two ratings to an earlier era, 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 because Peterson and Mayhew use completely different sources in making their determinations and never rate a common set of statutes.

The second source of temporal inconsistency is the potential difference between contemporaneous and retrospective assessments. Being charged with publishing a review of the activity of a congressional session may cause a rater to artificially inflate the perceived significance of some legislation—particularly during periods of relative inactivity. For instance, a newspaper that summarizes every session may feel obliged to designate at least some legislation as significant in each session in order to provide content. Perhaps the standard used to identify significant legislation is not absolute (across all time, was any policy produced in this

session particularly newsworthy?), but rather relative (of all the legislation passed in the session, which was the most newsworthy?). Depending on the amount of important legislation produced by a session and the standard used by the observer, standards may change from session to session due to the pressures of publishing a story (and therefore mention legislation).

Although we cannot rule out this possibility, available evidence suggests that this is not likely to be the case. Contrary to what we would expect if publishing needs artificially inflate the number of mentioned statutes relative to their true significance, the number of mentioned enactments varies considerably across years among the contemporaneous raters. For example, whereas *PSQ* mentions only three statutes in the 50th Congress and five in the 60th, it mentions thirty and thirty-one statutes in the 65th and 67th Congresses, respectively. The need to fill up column space does not result in a similar number of enactments being mentioned across time.

A final source of rater variation is the possibility that raters may make mistakes in their assessments. The process of culling through the legislative record is both exhausting and time-consuming. Raters may miss legislation that would have qualified as noteworthy according to their own standards.

The most prevalent response to rater heterogeneity is to adopt a "replication" strategy and determine whether substantive conclusions are robust to alternative measures. Howell and his colleagues (2000) employ this strategy in their replication of Mayhew's work on divided government. Although the replication strategy is used quite commonly in political science and is particularly pervasive in studies of lawmaking (see, for example, Krehbiel 1998; Coleman 1999), it only effectively checks whether the ratings are close enough to produce similar coefficient estimates when used in a regression structure.

A replication strategy is problematic because it fails to account for the information contained in the multiple ratings. Replication treats the four statutes identified as significant by seven raters (during the 66th, 67th, and 68th Congresses) identically to the forty-nine statutes rated significant by a single rater; it ignores the plausible conclusion that rater agreement constitutes evidence of increased significance and certainty about the enactment's assessment. Likewise, rater disagreement suggests that we should be less certain about a bill whose significance is disputed relative to an enactment with complete rater agreement. Rater agreement and disagreement conveys information about the magnitude and certainty of our assessment of legislative significance—information ignored by methods employing a replication strategy.

Two additional related problems arise. First, the dichotomous classification of landmark legislation versus everything else is extremely coarse. It is unlikely that that every statue mentioned by Dell and Stathis (for example) is equally significant. Because we use the information in the multiple ratings to distinguish between more and less significant legislation, our measure allows us to make such distinctions. Second, existing measures of legislative significance tell us nothing about the certainty we have about the designated significance. The fact that raters disagree in their determinations of the significance of legislation implies that uncertainty exists; although we may be quite confident that the four statutes identified as significant by seven raters during the 66th, 67th, and 68th Congresses are important, we may be less certain about the significance of the forty-nine statutes mentioned by a single rater.

It may be tempting to account for the information contained in the multiple ratings by aggregating the ratings using the sum or mean rating. However, aggregation requires strong assumptions. First, an additive aggregate measure assumes that significance increases linearly--a statute rated significant by nine raters is nine times more significant than a bill so rated by only

one. Rater agreement may serve to increase our certainty that legislation is significant instead of providing us with the magnitude of the significance.

Second, it is not obvious how to account for the fact that not every rater evaluates every statute in such a framework. If only two raters evaluate a particular Congress and both designate a statute as significant, how should we compare the resulting significance to a statute in another Congress that is rated as significant by the same two raters but not rated as significant by the other raters? Is the first statute as significant as the second, as the summation of the number of significant ratings suggests? Or is the first statute more significant than the second, as the average significance rating suggests?

A third difficulty is that such an aggregation strategy assumes that rater determinations are equally informative—no rater is better able to determine the true significance than others. The presence of rater disagreement suggests that we should account for differences in how raters evaluate in determining the proper weight to give their determinations. There is no reason to assume that if two raters disagree in their assessment of the significance of two pieces of legislation they must be equally significant; it is equally plausible that one rater is more likely to err in this determination.

Fourth, using a statistic from the ratings themselves does not provide an obvious way to incorporate additional information into the determination of legislative significance. Specifically, measuring significance using aggregated ratings ignores additional available information about legislation. For example, significant legislation may be more likely to be introduced in the first congressional session, or it may be that the more attention is paid to the legislation (as measured through pages of coverage in the index of the *Congressional Record*), the more likely it is to be

significant.[10] A convincing model of legislative significance would be one that exploits all available information.

A Statistical Model of Legislative Significance

Given the concerns noted in the previous section, we employ a statistical model to aggregate the information contained in the multiple ratings and the statutes' characteristics. The intuition of the model is that raters' perceptions of the true underlying significance of legislation are subject to error and that their decision to rate legislation as significant or not depends on whether their perception of the legislation's true significance exceeds their threshold of what constitutes significance. Although we assume that all raters agree on what constitutes significant legislation (that is, they agree on what the significance dimension is), raters may differ in their determinations either because they employ different significance thresholds (that is, they use higher or lower thresholds) or because idiosyncratic perceptual differences may obscure the true importance of the legislation to the rater. In the statistical literature we employ what is known as the multi-rater model (see Albert 1999; Johnson and Albert 1999), which we estimate using Bayesian methods.

The measurement model we present improves on existing work in at least three ways. First, employing a statistical rather than a counting model yields estimates and standard errors of legislative significance that account for the information contained in the ratings of every rater while simultaneously recovering estimates of the relationship between the raters' determinations.

Second, although not every rater covers the entire time period, the overlap among the raters provides the required "bridging observations" to ensure that the scale is comparable across

[10] Other sources of information of this type include the number of changes in the U.S. Code for each new public enactment (Young and Heitshusen 2002).

raters and across time. As long as the standards employed by each rater remain constant, we can use the relationship between the ratings on the 1,832 statutes passed in the 66th, 67th, and 68th (1919 to 1925) Congresses to recover rater characteristics that are directly comparable. [11] A second source of comparability results from relating the determinations of non-overlapping raters to a rater overlapping both. For example, if raters A and B contemporaneously rate legislation, then we can determine how their ratings relate using the set of commonly rated legislation. Similarly, the ratings of contemporaries B and C can also be determined. As long as B does not change the way he or she determines significance, knowing how the ratings of A and C relate to B's permits us to relate the ratings of A to C's even if A and C do not overlap.

Third, the statistical model we use allows us to incorporate information such as the amount of page coverage each statute receives in the *Congressional Record* and the session of Congress in which the legislation was proposed. Accounting for information besides the actual ratings is beneficial not only because the resulting measures accounts for this additional information but also because, if we are interested in the regression specification itself—as we would be if we were testing accounts predicting whether a statute is likely to be notable—then the estimated regression equation "controls" for the considerable uncertainty in the estimates of legislative significance.

The Basic Model

[11] The argument is analogous to the means by which Poole and Rosenthal (1997a) estimate "common space" scores across institutions. Assuming that individuals do not change their behavior across institutions (or across time in our instance) provides "bridging" observations that can be used to relate estimates in the two institutions (or across time).

We assume that a true and unobservable value of legislative significance is associated with all legislation. For legislation $t \in 1...T$ we denote the true latent significance as z_t. We assume that raters $i \in 1...N$ of legislative significance (for example, *Political Science Quarterly* and New American Nation) agree on what constitutes legislative significance and that z_t entirely determines a piece of legislation'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. Although there is no reason why significance must be a unidimensional concept, our investigation reveals no evidence of multiple dimensions—perhaps because we selected certain raters precisely because they share a common aim of identifying important legislation.

Each rater "votes" on whether legislation t is significant (1) or insignificant (0) based on the latent trait z_t. Let **Y** be the T x N matrix of ratings, with element y_{ti} containing the dichotomous choice of rater n with respect to legislation t. For the period we examine (the 45th through 80th Congresses), T = 21,741 and N = 9. Consistent with the observation that rater assessments may differ, we allow raters to imperfectly observe the true significance value z_t. Raters rely on the proximate measure $x_{ti} = z_{ti} + \varepsilon_{ti}$. In other words, for reasons such as the inherent difficulty of assessing legislative significance, differences in the selection methodology, and differences in effort level, the true significance of legislation is only imperfectly observed by raters. Although we assume that rater perceptions are unbiased (that is, $E[\varepsilon_{ti}]=0$), we allow for

the possibility that raters differ in the precision of their perceptions, perhaps because of different amounts of ambiguity in raters' standards. We denote the variance of ε_{ti} for rater i by δ_i^2.

Denoting the common distribution of ε 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. We assume that a rater's determination of whether an enactment is significant depends only on whether the rater perceives the latent value of the legislation x_{ti} as exceeding the rater's threshold γ_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. Consequently, raters may differ in their significance determinations either because there are differences in the threshold they use to determine significance or because they differ in the threshold they use. The model assumes that the discrepancy between the number of statutes identified by Dell and Stathis (1982) (241) and the number identified by Sloan (1984) (16) is due to the fact that Sloan employs a higher threshold rather than their evaluations on different dimensions of legislation.[12]

Given this rating rule, the probability of observing rater i designating legislation t as significant is given by the probability that the rater's observed trait for t is greater than his or her threshold. Given the assumption that $\varepsilon_{ti} \sim F(0, \varepsilon_{ti}/\delta_i)$, this implies that:

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

[12] We believe it is reasonable to assume that this is the basis for the difference for two reasons. First, we include only those raters who are consciously attempting to identify significant legislation. Second, estimating a two-dimensional model of significance that permits rater assessments to correspond to different standards reveals no statistical support for this possibility.

$$= \quad 1 - F((\gamma_i - z_t)/ \delta_i)$$

$$= \quad F((z_t - \gamma_i)/ \delta_i)$$

The expression $(z_{ti} - \gamma_i)/ \delta_i$ can be re-expressed as: $\beta_i z_t - \alpha_i$, where $\beta_t = 1/ \delta_i$ and $\alpha_i = \gamma_i/\delta_i$. β_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 the 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 β_i is very high, then the determination of significance is determined almost entirely by the underlying true quality of the legislation. The estimate of α_i is a function of the threshold used by rater i in determining significance. High (low) values of α_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; if a rater fails to evaluate an enactment, this absence is uninformative about both the quality of the legislation and the rater. This assumption is unproblematic given that the missing data result not from the raters' conscious choice but rather from the fact that not all raters evaluate for the entire time period. Given this structure, the Bernoulli probability of y_{ti} is: $\Pr(y_{ti} = 1 \mid z_t, \beta_i, \alpha_i) = F(\beta_i z_t - \alpha_i)^{yti} \times [1 - F(\beta_i z_t - \alpha_i)]^{(1-yti)}$. Assuming that the ratings are independent across raters conditional on the true latent quality of legislation z_t and the rater parameters β_i and α_i implies that $\Pr(y_{t1}, y_{t2},\ldots, y_{tN} = \mathbf{1} \mid z_t, \beta_i, \alpha_i) = (F(\beta_1 z_t - \alpha_1)^{yt1} \times [1 - F(\beta_1 z_t - \alpha_1)]^{(1-yt1)}) \times (F(\beta_2 z_t - \alpha_2)^{yt2}[1 - F(\beta_2 z_t - \alpha_2)]^{(1-yt2)}) \times \ldots \times (F(\beta_N z_t - \alpha_N)^{ytN}[1 - F(\beta_N z_t - \alpha_N)]^{(1-ytN)}) =$

$\prod_{i=1}^{N} F(\beta_i z_t - \alpha_i)^{yti} \times [1 - F(\beta_i z_t - \alpha_i)]^{1-yti}$ Further assuming that ratings are independent

across legislation yields the likelihood function:

$$L(z,\alpha,\beta) = \prod_{i=1}^{N} \prod_{t=1}^{T} \mathrm{F}(\beta_i \; z_t - \alpha_i)^{yti} \times [1 - \mathrm{F}(\beta_i \; z_t - \alpha_i)]^{1-yti}$$

where only **Y** is observable.[13]

One might question whether the possibility that raters rely on similar sources in their evaluations violates the assumption of independent errors. It does not because the assumption of independence is conditional on the rater parameters and the true significance of the legislation. In other words, that raters may employ similar sources means that the rater parameters (that is, the significance threshold used and the amount of error in rater perceptions of true significance) are similar, not that raters' errors are correlated. An analogy to roll call analysis might be helpful. Even though several roll call votes may deal with a similar issue, we would not say that, conditional on the legislators' ideal points and the proposals being voted upon, the errors in legislative voting are correlated. Instead, the fact that votes on similar issues will be similar is reflected in the similarity of the resulting estimates for the bill item parameters.

Despite a different motivation and interpretation of the estimated parameters, the likelihood is identical to that used in roll call analysis (see Clinton, Jackman, and Rivers 2004). This equivalence is most evident if we conceptualize pieces of legislation as "voting" on whether or not they are significant according to each rater (in other words, each of the 21,740 "legislators" may cast up to nine "votes").[14] Since everything except for **Y** is unobserved, the scale and rotation of the parameter space are not identified by the likelihood (Rivers 2003). This

[13] To estimate the model we assume that the errors are distributed logistically.

[14] In other words, the same technology that is used to measure legislative inputs (legislator-induced preferences) can also be employed to measure legislative outputs.

means that **z** and -**z** yield equivalent values for the likelihood, as does **z** and A × **z** where A is an arbitrary constant. Since we estimate the model using Bayesian methods, we must specify prior distributions for the estimated parameters vectors **z**, **β** , and **α**. Since we have no prior information about the discrimination and difficulty of the raters, we adopt uninformative priors.[15]

The Hierarchical Model

We can expand the basic model to accommodate additional information related to the latent significance of legislation using a hierarchical item response model. A hierarchical model specifies a regression structure between observed legislative characteristics and the latent significance **z** instead of assuming that we possess no additional information (that is, the prior distribution of legislation significance is identical and uninformative [for example, $z_t \sim N(0,1)$]). In particular, we assume that the prior distribution of **z** is $N(\mu_z, \tau^2)$, where the prior mean μ_z is a function of additional observable information.

In our case, it is plausible that the length of time members of Congress spend on a statute (as measured by the number of pages devoted to it in the *Congressional Record*) is correlated with legislative significance. Some have also argued that legislation introduced in the first session of a Congress is more likely to be significant than legislation introduced in later sessions (Neustadt 2001; Rudalevige 2002).[16] A hierarchical item response model allows us to account for such possibilities. Let **W** denote the T × d matrix of covariates. If we assume an additive and linear relationship, then estimating a hierarchical model assumes that $\mu_z = \mathbf{W}\kappa + \varsigma$, where **κ** is a d × 1 matrix of regression coefficients and **ς** is an iid error term. For example, in this paper we

[15] The likelihood surface has many "flat spots" because of the small number of raters.

[16] The norm is two congressional sessions; however, several Congresses have had extra "special sessions."

posit that $\mu_z = \kappa_0 + \kappa_1 \log(\text{Pages}) + \kappa_2 I1 + \kappa_3 I2 + \kappa_4 I3 + \varsigma_t$, where I1, I2, and I3 are indicator variables for whether legislation t was introduced in the first, second, or third session, respectively.

Imposing this linear regression structure and estimating a hierarchical model yields two benefits relative to the nonhierarchical model of the previous section. First, the hierarchical estimator of **z** i "borrows strength" from the additional information. In other words, estimates of **z** account for information contained in the covariates **W**. Since legislation with identical characteristics and covariates is assumed to share the same prior mean (although the prior variance permits legislation with identical covariates to differ in significance), the hierarchical estimator can distinguish between identically rated legislation on the basis of their characteristics if those characteristics covary with legislative significance. This is particularly important when the problems associated with "chunky data" (Londregan 2000) result in few ratings.

Second, if we are interested in the determinants and covariates of legislative significance, adopting a hierarchical approach recovers estimates of $\boldsymbol{\kappa}$ that account for the (substantial) uncertainty in estimates of **z**. The intuition is that since the current estimates **z** are regressions on **W** at every iteration, as the estimates of **z** fluctuate across iterations so too will the coefficients $\boldsymbol{\kappa}$ measuring the relationship.

Characterizing Legislative Significance

This section assesses the different significance scores we estimate for the period between 1877 and 1948. We compare the recovered parameter estimates from the nonhierarchical and hierarchical models and we examine the resulting statute-level estimates of significance and estimates of aggregate lawmaking activity.

Comparing the nonhierarchical and hierarchical estimates to the mean rating for the 21,741 public statutes passed from the end of Reconstruction to the conclusion of World War II yields several conclusions.[17] First, if we compare the mean estimate to those of the nonhierarchical estimator the estimated change in significance in moving from an average rating of 0 to .2 is much greater for the nonhierarchical estimates than the estimated change of moving from an average rating of .2 to .4; the marginal change of our estimate of legislative significance for the first rater's determination of significance is much larger than that of subsequent raters. Consequently, legislative significance is not simply a linear function of the number of raters who evaluate the legislation.

Second, incorporating additional information about the legislation using a hierarchical model refines our estimates of legislative significance. While the mean rating scores every nonrated statute with a significance of 0 (and the nonhierarchical item response model recovers as small an estimate as possible given the lack of any information about those statutes), the hierarchical estimates permit covariates of significance to influence the recovered significance rating. The primary contribution of the hierarchical estimator is its incorporation of information that can be used to distinguish among unrated legislation—we recover significance estimates for unrated statutes that range from 0 to .5 depending on the characteristics of the statute. [18] Since the vast majority of statutes are unrated, and since it is clear that unrated enactments are not equally unimportant, this is a valuable benefit.

---

[17] For purposes of comparison, the posterior means of the hierarchical and nonhierarchical estimators have been rescaled to lie between 0 and 1. Given the scale invariance noted earlier, this rescaling is unproblematic.

As an illustrative example, consider public statutes authorizing bridge building. Passing legislation to build bridges is typical "pork barrel" politics. Although the appetite of members of Congress for pork is a well-established fact (Ferejohn 1974), not all pork is equivalent. From 1877 to 1948, Congress passed over 3,100 bridge bills—a little more than 14 percent of all passed statutes. If we rely on the mean rating, all of these bills would receive the identical significance estimate of 0. The nonhierarchical estimator recovers similar estimates. However, the hierarchical estimator provides a means for distinguishing between important and trivial bridge bills on the basis of the additional information that the hierarchical estimator employs. Since page coverage in the *Congressional Record* for bridge bills ranges from four to fifty-seven pages, the hierarchical model distinguishes between significant and insignificant bridge bills using this information. For example, the bridge bill with the highest page count, passed in 1908, was the first omnibus bridge legislation ever enacted into law. [19] This bill included twenty-three bridges and faced unsuccessful attempts at obstruction. The posterior mean of the hierarchical estimate for this bill is .36, placing it in the top 8 percent of the distribution of all public bills according to the hierarchical estimator.

As further evidence, consider the top 1,000 significant statutes according to the hierarchical model. Of the top 1,000, 113 are not rated by a single rater as significant. Of these 113 statutes, over 90 address appropriation or tax-related measures. While these bills are clearly not as important as some of the legislation passed during the period (as evidenced by their lack

[19] One might think that important projects, including the building of the Golden Gate and George Washington Bridges, would have been ranked as significant. Neither building project, however, was the direct result of action taken by Congress. The Reconstruction Finance Corporation played an important role in purchasing revenue bonds to finance the building of the Golden Gate.

of mention), they are certainly more important than most of the other unmentioned bills passed by Congress. These unrated enactments score high because several hundred pages of debate in Congress are recorded in the *Congressional Record* for each of them, a characteristic that corresponds to the likelihood of being rated given the regression structure we employ. These examples are typical of many other unmentioned statutes that are far from trivial.

Assessing Face Validity

Although the evidence thus far suggests that the hierarchical estimator performs reasonably well relative to available alternatives, it is useful to establish face validity for the hierarchical estimates by comparing the recovered estimates to conventional wisdom. The enactment with the highest estimated significance score is the establishment of the Federal Trade Commission (FTC) in the 63rd Congress.[20] The FTC dealt with the business trust issue first addressed by the Sherman Act of 1890.[21] Although we cannot be certain that the establishment of the FTC is *the* most important statute once we account for the (considerable) estimation

[20] The need for the FTC did not gain prominence as an issue until President Theodore Roosevelt repeatedly included messages about the dangers of trusts in his State of the Union addresses. Roosevelt made no progress, however, in regulating trusts during his administration (Chamberlain 1946). The trust issue was addressed through legislation during the 63rd Congress under unified Democratic control. Spurred on by the Supreme Court decision over Standard Oil on May 15, 1911, Congress passed two bills that were signed by President Wilson, including the Clayton Act and the act establishing the FTC.

[21] Interestingly, the Sherman antitrust bill receives one of the highest significance scores in our sample. The significance of this bill is certainly captured by the retrospective importance of the act.

uncertainty, what is notable is that the statutes we would expect to receive the highest significance scores actually do so. For example, the Pendleton Civil Service Reform Act (1883) receives a score of .866, while the National Industrial Recovery Act receives a statistically indistinguishable score of .865. Both statutes were tremendously influential, even though their policy content is distinct and they were passed in different eras. Also encouraging is the result that some Congresses do not produce legislation as significant as legislation produced by others. For example, the 45th Congress produced the Bland-Allison Silver Act (1878), which receives a score of .855, and the Timber and Stone Act (1878), which receives a score of .674, while the most significant statutes of the 46th Congress consisted of a sundry civil appropriation (.486) and a military appropriation (.471). The 54th and 58th Congresses were also not particularly memorable in terms of landmark legislation. The New Deal era, by contrast, especially the 73rd and 74th Congresses, produced loads of important enactments.

In addition to examining whether the statute-level estimates appear reasonable, we make a second check on validity: do the estimates aggregate to produce a time series of significant enactments that matches with historical understanding? To measure productivity we follow Mayhew and count the number of significant statutes enacted during each year. A question arises: what gets counted? For a dichotomous measure of significance such as that employed by Mayhew, answering this question is straightforward: sum the identified bills. However, since we can estimate a much finer measure, the question as to what is counted as a significant enactment is slightly more complicated. Consequently, we rank-order the 21,741 statutes according to their estimated significance level and count the number of statutes each Congress passed that lie in some restricted set. Given this ambiguity, figure 24.2 presents the number of statutes passed each year that are among the top 1,000 (upper graph), top 500 (middle graph), and top 250 (lower

graph) statutes.[22] The horizontal line denotes the number of statutes that would be passed if every Congress were equally productive.

A measure with face validity would capture important and known trends, such as the surges in legislative productivity during the Progressive and New Deal Eras and the spikes during crises such as war. The expected surges, including solid bumps during World War I and World War II, are easy to identify. Capturing moments of non-activity is also important. The infamous 80th "Do Nothing" Congress produced only a single top 100 enactment (the Taft-Hartley Labor-Management Relations Act) and a well-below-average twelve top 500 enactments, but surprisingly, it also produced an above-average forty in the top 1,000. Consequently, although the "Do Nothing" Congress was inactive relative to the Congresses of the New Deal and World War II, it is actually near the middle of the distribution in terms of overall productivity. Such findings and trends are interesting in and of themselves in describing congressional productivity, but the true importance of the measure lies in offering a way to address substantive questions, such as the determinants of legislative productivity.

Conclusions

Improving our ability to identify high-stakes legislation is critical to assessing many of the questions that are fundamental to representative government. For example, our measure of significance can be used to explore how institutional rules affect collective choice outcomes. As Schickler (2001, 3) argues: "Congressional politics has depended crucially on such innovations

[22] Note that the uncertainty in the significance estimates necessarily produces uncertainty in our estimate of the number of significant statutes passed each year. Accounting for this uncertainty is straightforward, but since this uncertainty does not affect our assessment of face validity, we have omitted it for presentational clarity.

as the 'Reed rules' of 1890, the Senate cloture rule adopted in 1917, the creation of congressional budget committees in 1974, the House breakaway from seniority rights to committee chairmanships in 1975, and the package of reforms adopted by the Republicans when they took over the House in 1995."

Since many of the most important reforms occurred prior to existing measures, scholarship has focused largely on the causes of institutional reform—for example, Binder's (1997) examination of the creation of minority and majority rights within Congress—rather than on the impacts of such reform on legislative productivity and lawmaking.[23] Given the number of important institutional (and electoral) reforms that occurred prior to World War II—such as Reed's rules, cloture reform, committee reorganizations, and the direct election of senators—to the extent that the difficulty in measuring legislative output prevents investigations of the consequences of reform, the data and methodology we present offer the opportunity to examine the consequences of institutional changes. Although important and difficult questions are clearly raised by any such investigation (for example, how to control for both the demand for and supply of high-stakes legislation), our data and measure enable us to begin to confront such questions.

By helping us understand how the configuration of preferences within an institutional structure affects legislative productivity, our measure gives us a way to begin answering questions such as: did "divided control" of the government make Congress unable to pass important legislation? Did the realigning elections of 1896 and 1932 produce periods of increased lawmaking? What has been the impact of a strong Speaker of the House (such as Reed and Cannon)? How has party polarization influenced Congress? Although all of these factors

---

[23] Dion's *Turning the Legislative Thumbscrew: Minority Rights and Procedural Change in Legislative Politics* (1997) is another important book in this field.

have received considerable attention, our ability to determine the actual impact of each on lawmaking has been inhibited by the difficulty to date of measuring policy output. Our estimates make it possible to conduct precisely such investigation.