Chapter 2. Measuring Political and Policy Preferences Using Item Response Scaling

Joshua D. Clinton
Department of Political Science
Vanderbilt University

It is easier than ever to acquire data on an unprecedented scale. This data is also thought to hold tremendous potential for unlocking previously hard to assess truths about our world. For those interested in questions related to the production and assessment of public policy, the ability to characterize current and past conditions and to assess the consequences and implications of various policies and events using data is unprecedented. Paradoxically, the amount of available data may also be paralyzing; too many measures and indicators may obscure and complicate our ability to extract the core underlying relationships because of uncertainty as to which measure -- or measures -- to use.

If what we are interested in is observable and measurable – e.g., growth in per capita GDP, or the number of riders taking mass transit – then conventional statistical methods are appropriate and often easily used to characterize and analyze the observed variation. If, however, our interest lies in something less tangible – say the extent to which a community's "happiness" is improving ; what the overall policy preferences of a congressional district is (Levundusky, Pope and Jackman 2008); how democratic a country is (Pemstein, Melton, and Meserve 2010) ; how insulated a federal agency is (e.g., Selin 2014) ; or how effective legislators are (e.g., Volden and Wiseman 2014) -- it may be less clear how to make the required comparisons. Moreover, even if we are interested in assessing readily observable actions and outcomes, characterizing the underlying structure of the data in a principled manner may be difficult if there are a multitude of relevant measures.

As our access to data increases and our ability to measure and quantify aspects of interest improves, so too does the number of measures and comparisons that can be made. Insofar as we cannot identify a single best measure of interest, how might we make use of multiple existing measures to best assess the underlying characteristic of interest? For example, how can we identify the relative aptitude of applicants to a public policy school? Or identify which members of a legislature are most pivotal for passing a piece of legislation? Or characterize the policy preferences of bureaucrats and agencies?

Statistical measurement models are increasingly vital for empirically oriented policy analysts who find themselves confronted with an ever-increasing number of measures that they must characterize and evaluate. When correctly and appropriately applied, these models provide a principled way of analyzing multiple measures to identify and characterize the core underlying tendencies.

In the chapter that follows, I briefly describe a statistical measurement model based on item response theory and I provide some illustrative examples of how this approach can be used to characterize features relevant for public policy analysis. Entire books, chapters, and papers have been written on these models (e.g., Johnson and Albert 2000; Baker and Kim 2004; Gelman and Hill 5006; Jackman 2008,2009), but the goal of this chapter is to introduce the main intuition behind the models and to motivate their use by showing how they can be applied to questions that are relevant to the practice and analysis of public policy. In so doing, I do not discuss the computation involved in the analysis and interpretation of the models, but there are many programs devoted to such tasks.[1]

[1] The models in this chapter were implemented in R using either the `pscl` (Jackman 2014) or `MCMCpack` (Martin, Quinn, and Park 2011). In each case, the recovered estimates were normalized to have a mean of 0 and a variance of 1.

Statistical measurement models are useful primarily because they provide a principled way of using multiple measures to characterize a concept that we cannot directly observe – a latent variable – by identifying the common structure of the observed data.

Despite the potential of these models to use multiple measures to learn about hard-to-measure concepts, they are not magic and there is no guarantee that they will necessarily produce the quantity that is most desired by the policy analyst. The ability of statistical measurement models to extract meaningful information from the data depends critically on the validity of the assumptions used in the statistical models, the information contained in the measures being analyzed, and the substantive information used by the analyst to construct the model. While there are great opportunities for such models to help identify trends from a host of related measures, the ability to do so is not unlimited and the applicability of the statistical model and the meaning of the resulting estimates depend heavily on the substantive knowledge of the policy analyst performing the analysis.

In the chapter that follows I first describe the idea of "latent" traits and why they are relevant for the analysis of public policy and how we might think about characterizing such traits using the analogy of administering a standardized test to assess students' aptitude. I then show how this intuition can be applied to three examples – figuring out which senator to lobby in the 2008 Affordable Care Act debate, comparing the policy preferences of career executives relative to elected officials in the US, and using experts to describe the policy preferences of the federal bureaucracy – before offering some concluding thoughts about the use of statistical measurement models for the study of public policy.

The Importance of Measuring Latent Traits For Public Policy

Much of what we are interested in measuring in public policy cannot be directly quantified. This is true if we are observers interested in characterizing political situations (e.g., how conservative are elected officials? How politically aware are citizens? How efficient is a particular government agency?), or if we are government officials interested in assessing conditions in the polity and the consequences of policy interventions (e.g., the "happiness" of citizens, the overall desirability of a policy reform based on several measures). Unlike quantities that can be directly and easily measured such as the outside temperature or the number of cars using a roadway on a particular day, some concepts of interest in public policy cannot be directly measured easily.

Given the rapid growth in the amount of available data, even if we lack a single clear observable measure of the quantity we care about, we sometimes have many measures that we think are related to the unobservable aspect of interest. Statistical measurement models provide us with the ability to use the measures we can observe to learn about features that are thought to be related to the measures we observe.

For example, suppose that we are working for an interest group and we want to identify which elected officials have the most moderate policy preferences. How might we use the set of votes we observe from the elected officials to make this determination? Or imagine you are serving in a newly elected governor's office and you are conducting a survey of supposed experts to determine the extent to which various bureaucratic agencies are more or less ideological in the policies they pursue. How should we combine the various opinions you gather when some experts are probably better than others, but it is not entirely clear which experts are better than others? Or perhaps you are the chief of staff for an elected representative and you are trying to assess the extent to which citizens

are aware of various features of a recently enacted policy, or their propensity to engage in political activism. How might we use a battery of survey questions to measure political interest?

Statistical measurement models are certainly not new and they go by a host of names depending on the particular assumptions and academic disciplines in which they are employed. Various formulations of latent variable models include Structural Equation Modeling (Bollen 1989), Item Response Theory (e.g., Baker and Kim 2004), and Factor Analysis (e.g., Harman 1976), and they have been used in nearly every social science discipline. There are subtle differences between the various statistical measurement models, but they all share a basic underlying structure. All of the models posit that the measures we observe are related to an underlying concept, but that the precise nature of the relationship may vary across the measures with some measures being more or less related to the underlying concept.

Providing Some Intuition for the Approach: Assessing Aptitude

To provide some intuition for statistical measurement models, suppose you are in charge of a government agency and that you are tasked with hiring individuals who are knowledgeable and competent. Or perhaps you are running the admissions of a public policy school and you want to admit students with the greatest aptitude. Competence and aptitude is often hard to assess and it is not something we can directly observe. We think that we can observe aspects that are related to the presence or absence of these traits in an individual, but unlike quantities such as height and weight there is not clear measure of an individual's aptitude. As a result, we often rely on heuristics such as prior experience,

letters of recommendation, and past performance to help identify which applicants possess the most aptitude.

Even with such information, there are non-trivial difficulties with assessing the competence or aptitude of an applicant. Even if an individual's experiences provide an accurate assessment of aptitude or competence, it is difficult to compare applicants because individuals have different experiences and opportunities.

Because of this difficulty, we often use standardized formal examinations to provide another measure of traits that can be hard to measure. In fact, one of the first applications of statistical measurement models was an attempt to measure intelligence (Spearman 1904) and these models are still used by the Educational Testing Service to construct the Scholastic Aptitude Test. While acknowledging that it can be difficult to write questions that accurately reveal aptitude or competence, thinking about the properties we would want such questions to have helps motivate the intuition behind statistical measurement models.

First, we would want to ask questions where the ability to provide the correct answer depends primarily on the traits of interest – be it competence or aptitude. For example, if aspects unrelated to a test-taker's aptitude affect the probability of observing a correct response (e.g., personal background), the question will not accurately measure aptitude. The requirement that the probability of correctly answering a question depends primarily on the latent trait of interest is known as *item discrimination* in the item response framework. Questions with high item discrimination are questions where individuals with different aptitudes will have different probabilities of providing a correct answer to the question and questions with low item discrimination are those in which the variation in

answers is unrelated to individual aptitude; a test comprised entirely of questions with low item discrimination will provide little, if any, information about the actual aptitude of the test takers. Giving test-takers a test about chemistry, for example, will not reveal much about their knowledge of politics.

In addition to asking questions whose answers depend primarily on the aptitude or competence of test takers, we may also want to account for the possibility that the probability of observing a correct response may differ across questions and it may be unrelated to the test-takers' aptitude. In the language of item response theory, questions may vary in their *item difficulty*.

To describe more precisely how we might use these concepts to measure a latent trait, let us consider the mathematical relationship between these concepts. Recalling the motivational example, suppose that you are in charge of constructing an entrance exam to determine the students with the highest aptitude for the study of public policy, or a civil service exam designed to identify those with the most competence. In either case you are interested in measuring the latent trait of the test-takers that we will denote as x_i. Because we cannot directly observe x_i, we want to administer a test of T items and use the responses to that test to estimate the latent trait x_i for each test-taker.

For each of the T items, we observe every test-taker *i* providing an answer on item t that can be classified in terms of a binary response (e.g., "correct" (1) vs. "incorrect" (0); "competent" (1) vs. "incompetent" (0)). For the resulting binary outcome y_{it}, we can

express the probability of observing a "correct" or "competent" (i.e., $y_{it} = 1$) response by every test-taker i on every item t as: [2]

$$\Pr(y_{it} = 1) = F(\alpha_t + \beta_t x_i)$$

where F() is a cumulative distribution function (typically either a standard normal or a logistic), x_i is the unknown aptitude of member i that we are interested in characterizing, α_t denotes the probability of providing a correct answer on question t regardless of a test-taker's aptitude (item difficulty), and β_t indicates how much the probability of providing a correct answer varies depending on changes in competence (item discrimination).

If a test item is sufficiently easy and every test-taker is able to provide a correct answer regardless of their actual aptitude (i.e., $y_{it} = 1$ for all individuals i on item t), α_t will be exceptionally high and $\beta_t = 0$. If so, the probability of a correct answer will be unrelated to the aptitude of the test takers – $\Pr(y_{it} = 1) = F(\alpha_t)$ – and we will subsequently be unable to learn about an individual's aptitude x_i because the item is too easy. Conversely, an item may be too hard for the test-takers and if everyone gets the item wrong if will not be useful for ranking the test-takers' aptitude. If so, α_t will be exceptionally low and $\beta_t = 0$. These cases are extreme examples, but they illustrate the concept of item difficulty discussed above.

If the variation in the probability of "correct" responses ($y_{it} = 1$) are perfectly correlated with variation in individual aptitude x_i and test-takers with higher aptitudes are more likely to provide a "correct" response than test-takers with lower aptitudes on the item, β_t will be very high. This is a situation with high item discrimination -- and responses to item t are very useful for ranking test-takers according to their aptitude x_i.

[2] The model has been generalized to allow for continuous and ordered responses as well (see, for example, Quinn 2004).

The benefit of the statistical measurement model is that the model allows for both possibilities to be true – items may vary in both their difficulty (α_t) and their ability to discriminate between various levels of the latent trait (β_t) – and our estimation of the latent trait x_i that uses the items can account for the characteristics of the items being used.

At this point, you may be wondering – why do I need a model? Why not just add up the number of correct answers to assess the aptitude of a test-taker? There are several reasons why a statistical measurement model is often preferable to adding up the number of items that are correctly answered.

First, if different tests are given to different individuals – perhaps because you want to vary the test across time to prevent cheating – and you are interested in comparing scores across tests it is important to account for the possibility that some tests may be harder than others. Adding up the number of correct answers will not account for the fact that some items may be more or less difficult, and different tests may therefore also be more or less difficult depending on the items that are asked. By determining the item difficulty and item discrimination of every item being asked in each test, the statistical measurement model provides an ability to calibrate responses across different tests (so long as a few conditions are satisfied). Because the model explicitly accounts for possible variation in the items being used it is possible to make comparisons across tests.

Second, we can also use the model to learn about the structure of the responses we observe. While the discussion so far presumes that you know what the latent dimension of interest is – i.e., aptitude or competence – suppose that you thought it were possible that several latent traits might structure the pattern of responses you observe. For example, in addition to assessing aptitude, maybe you are also interested in assessing the motivation os

the test-takers by asking several onerous items such that the answers to the test questions might depend on both the aptitude and motivation of the test-takers. If so, we can extend the model to reflect the fact that responses depend on multiple latent traits so long as a few conditions are satisfied. Put differently, we can use statistical measurement models to learn the number of latent traits that are responsible for the observed variation and how each item relates to each of the latent traits. (This is similar to exploratory factor analysis.)

There are many details and nuances involved in the application of these models, but having sketched out the intuition behind such models using the analogy of constructing a test, let us now see how this tool can be used to help us better understand the practice and performance of public policy.

Example 1: Ranking Legislators and the Politics of the 2008 Health Care Debate

Suppose it is 2008 and you are a lobbyist concerned with the upcoming debate on the Affordable Care Act. Or perhaps you are an advisor to President Obama interested in enacting health care reform, or a member of the Office of the Whip in either the majority or minority party and you are trying to identify the impediments to health care reform. Regardless of your position in the debate, you know that passing anything in the U.S. Senate is going to require 60 votes to invoke cloture and end an attempted filibuster. But who do you need to lobby on this vote? Which Senators are most likely to be the 60th most liberal Senator – and thus the Senator whose vote will be pivotal for invoking cloture (or not)? How similar are the policy preferences of these senators to the policy preferences of the average Democrat in the Senate?

The framework provided by item response models discussed above provides a principled way of answering all of these questions. What we seek is a measure of every

member's policy preferences – which we can summarize in terms of their most-preferred ideal point x_i. If we think of political outcomes as being described by a ruler ranging from extremely liberal outcomes to extremely conservative outcomes, the ideal point for member i can be thought of as measuring the location of every member on that scale.

While we cannot directly observe the members' policy preferences because they are a set of beliefs that exist only in their head, we can observe the actions that they take in Congress. One activity that we observe a lot of is roll call voting behavior on issues that come to the floor. Insofar as we think that the votes are being cast in ways that reflect the latent ideology of the members (x_i), we can extend the test-taking example from the prior section to estimate the policy preferences of the members who cast roll call votes. That is, we can treat members of Congress as test-takers who take a test on political ideology whose items are the roll call votes being voted upon.

By thinking of members of Congress as test-takers and by thinking of the votes that they cast as casting votes for policy outcomes that are more or less conservative as test-takers who are answering questions on a survey about their conservativeness we can use the item response model presented above to estimate the policy preferences of members of congress and answer questions such as: which members have the highest probability of being pivotal for invoking cloture (i.e., who needs to be lobbied?), and how dissimilar are those members from the Democratic caucus in the Senate (i.e., how hard do they need to be lobbied)?

To make the connection explicit, we observe every member of the Senate *i* in 1…100 voting either "yea" (1) or "nay" (0) on each vote t (out of T total observed votes). If we assume that members vote for policies that are closest to their ideal point x_i – which is

where senator i is located along a liberal-conservative scale - the probability of observing a yea vote by member i on vote t (i.e., $y_{it} = 1$) can be expressed as:

$$\Pr(y_{it} = 1) = F(\alpha_t + \beta_t x_i)$$

where the item difficulty parameter α_t denotes the probability of voting yea on vote t regardless of a member's ideal point, and β_t describes how much the probability of voting yea depends on the member's ideal point. Even though we no longer have "correct" answers as we did when applying the model to a test-taking situation, we can accomplish an analogous interpretation by defining whether conservative policy outcomes are greater than or less than zero. For example, we can define conservative policy outcomes to be associated with positive ideal points by constraining the ideal point of then Sen. Minority Leader Mitch McConnell (R, KY) to be greater than 0. If we do so, then $\beta_t < 0$ if vote t involves a vote on a liberal proposals which means that $\beta_t x_i < 0$ for conservative members with ideal points $x > 0$ and they will therefore have a lower probability voting "yea" on the liberal proposal than liberal members with $x < 0$, it will be the case that $\beta_t x_i > 0$. Similarly, votes on conservative proposals will generate votes with $\beta_t > 0$ which will flip the relative probabilities, and votes that are unrelated to ideology – such as so-called "hurrah" votes on which all members agree – will produce estimates such that $\beta_t = 0$.

Figure 1 presents the results from analyzing all 696 roll calls cast in the 112th U.S. Senate that enacted the Affordable Care Act using the model of Clinton, Jackman, and Rivers (2003) as implemented using the `ideal` function in the `pscl` package for R (Jackman 2014). The resulting ideal point estimates are normalized to have a mean of 0 and a variance of 1 so the estimated policy preferences will roughly range from -2 (a point associated with the voting behavior of more liberal members) to +2 (a point associated with the voting

behavior of more conservative members. Senators' ideal points are plotted along the bottom axis, along with the density of ideal points by party. Figure 1 reveals that every Democrat votes in a more liberal manner than the most moderate Republican, and whereas Democrats are fairly evenly spread out between -1.25 and -0.5 on the liberal side of the spectrum whereas most Republicans were located in the neighborhood of 1.25. The vertical lines denote the median Senator (i.e., the 50th Percentile) as well as the location of the 60th Percentile – which is where the Senator needed to overcome a conservative filibuster is located. Ending debate in the U.S. Senate in order to vote on the policy requires the support of at least 60 senators so those located near the right-most vertical line are critical for enacting policy.

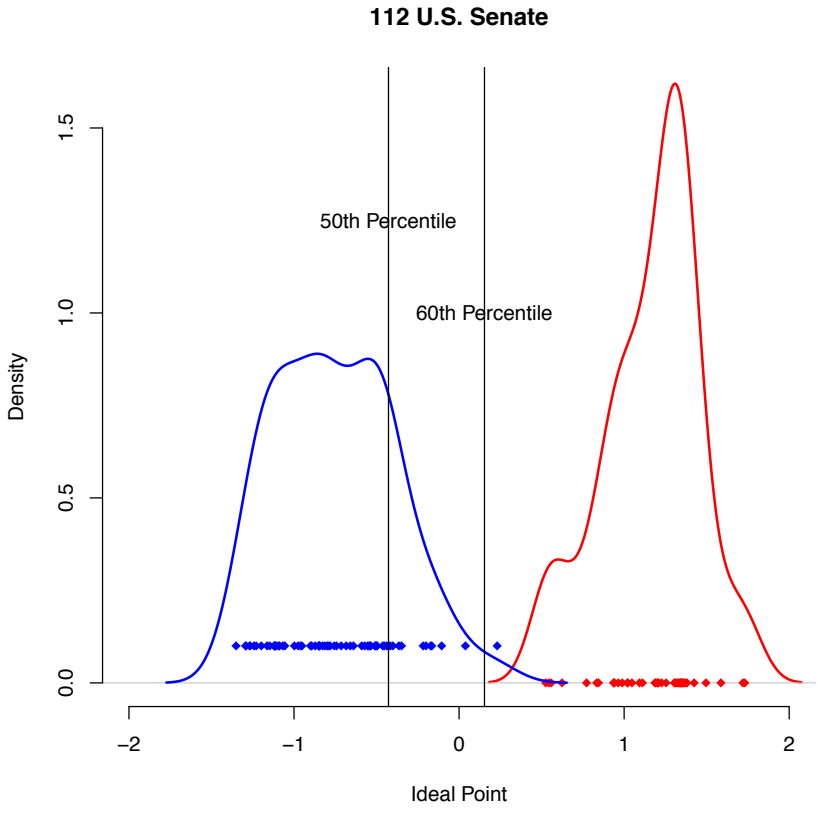

Figure 1: Distribution of Ideal Points in the 112th U.S. Senate

Figure 1 presents the overall distribution of policy preferences in the U.S. Senate, but if you are interested in figuring out who should be lobbied to ensure that cloture is invoked and a filibuster does not occur you need to determine which senator are most likely to be critical for that vote – you do not want to spend your time lobbying members who will almost surely oppose (or support) the cloture vote. Figure 2 uses the ideal points summarized in Figure 1 to identify the members who are most likely to be the 60th most liberal member responsible for invoking cloture by plotting the estimated rank of each senator along with the 95% regions of highest (posterior) probability. These are the members that are required to end debate in the Senate and proceed to a vote.

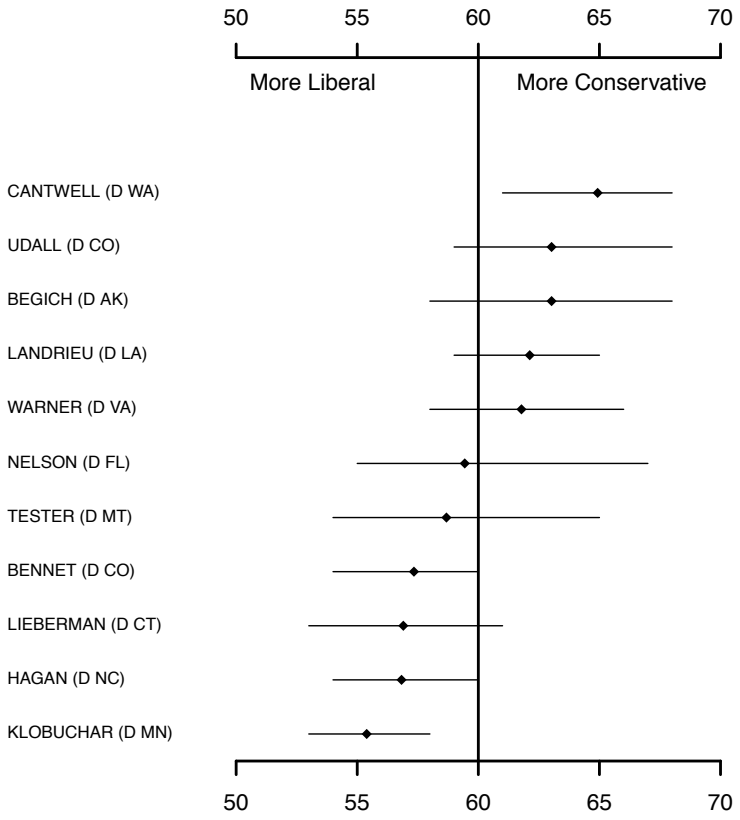

Figure 2: Senators Most Likely to Be Critical for Invoking Cloture on Health Care

An absolutely critical aspect of statistical analysis is the ability to quantify our uncertainty about a measured quantity. The National Institute for Standards and Technology is quite explicit about the need to quantify our uncertainty, noting that: "A measurement result is complete only when accompanied by a quantitative statement of its uncertainty. The uncertainty is required in order to decide if the result is adequate for its intended purpose and to ascertain if it is consistent with other similar results." A benefit of the statistical measurement model we present is that it is possible to explore how certain we are of our estimates.

Figure 2 reveals, for example, that Sen. Nelson (D, FL) is the Senator whose ideal point is closest to the 60th percentile in the U.S. Senate but there is a considerable amount of uncertainty in this assessment -- we cannot be sure that he isn't actually either the 55th or the 65th most liberal. Figure 2 therefore reveals that there is likely a need to target multiple members to be sure that the pivot senator is lobbied. While we can be confident that Sen. Klobuchar (D, MN) is more liberal than the 60th most liberal senator and that Sen. Cantwell (D, WA) is more conservative than the 60th most liberal, there are a range of senators who might plausibly be the 60th most senators. Any senator whose range of possible ranks includes 60 is a possible candidate for being pivotal.

This framework could be easily extended to examine other political incidents – e.g., the impeachment trial of President Clinton (Bertelli and Grose 2006) – or other political elites such as U.S. Supreme Court Justices (e.g., Martin and Quinn 2002) as well as assessing important questions such as: How well do elected officials reflect the views of their constituents (e.g., Jessee 2009; Bafumi and Heron 2010)? How much political polarization is there among elected elites (McCarty, Poole, and Rosenthal 2006; Poole and Rosenthal

2007)? And how do policy preferences affect lawmaking behavior (e.g., Wawro and Schickler 2007)?

Example 2: Analyzing the Opinion of the Public and Elites: Measuring the Political Preferences of Career Executives in the U.S. Bureaucracy

To illustrate the model more concretely, suppose that you are a member of the Executive Office of the President and you are trying to assess the political beliefs of career bureaucrats and how those beliefs compared to members of Congress. Particularly given the very different experiences that career bureaucrats may have with politics and political elites depending on both their position and their location in the federal bureaucracy, it is not obvious how best to do so.

One possibility would be survey bureaucrats (e.g., Aberbach and Rockman 2000; Maranto and Hull 2004) – possibly using a questions common to many surveys given to the general public. That is, to ask them a variant of the following – "In general, would you describe your political views as: (1) Very conservative, (2) Conservative, (3) Somewhat conservative, (4) Moderate, (5) Somewhat liberal, (6) Liberal, (7) Very liberal, or (8) Don't Know." While this question may be acceptable for the general public, you may worry about the usefulness of such a question when applied to individuals who are more closely attuned to politics than the general public. What exactly does "strong conservative" mean for an employee in the Department of Homeland Security and how does that meaning compare to an employee in the Department of Agriculture? Are bureaucrats thinking of the same policy areas when responding to this broadly worded question? Moreover, what is the policy difference between "strong conservative" and "weak conservative" (or "strong liberal" and "weak liberal") and are these differences meaningful across individuals? To

make any inferences using responses to this question requires us to assume that these aspects are equivalent, but we do not know for certain whether this is true (Brady 1985).

A further limitation is that even if bureaucrats have a shared conception of ideology and what it means to be "strong" or "weak," how should we compare the survey results to the ideology of members of Congress? How do the political views of a bureaucrat who thinks of themselves as "liberal" compare to the positions of President Obama? Or Rep. Nancy Pelosi? While some have tried to ask members of Congress about their ideology – see, for example, the Political Courage Test administered by Project Vote Smart (e.g., Ansolabehere, Stewart, and Snyder 2001a, 2001b; McCarty and Shor 2011) – it is unclear how the responses of legislators compare to generic political questions that are asked of bureaucrats.

Given the importance of federal bureaucrats for the implementation of public policy, some scholars have attempted to locate the policy preferences of career bureaucrats relative to those of elected officials by asking bureaucrats how they would have voted on fourteen issues that were considered in the previous Congress (Clinton, et. al. 2012). By collecting measures on how bureaucrats would have voted on these issues we can better relate the political ideologies of these two groups and avoid the ambiguities of interpreting what it means to indicate that a respondent is "conservative." For example, bureaucrats were asked: "We are also interested to know your personal opinion about several key votes in Congress in the last few years. Specifically, would you have supported the following measures? A bill to permanently reduce estate taxes: Yes, No, or Don't Know." Table 1 presents the full set of questions that were asked on the *Survey on the Future of Government Service*.

In addition to the general political background of executive officials, we are also interested to know your personal opinion about several key votes in Congress in the last few years. Specifically, would you have supported the following measures?

	Yes	No	Not Sure
A bill to authorize electronic surveillance of suspected terrorists without obtaining court approval (502/HR5825).	○	○	○
A bill to ensure access to federal courts for individuals who challenge government use of eminent domain to take their property (511/HR4772).	○	○	○
Efforts to amend the Constitution to prohibit desecration of the U.S. flag. (189/SJRes12)	○	○	○
A bill to require photo identification and proof of citizenship for voters in a federal election. (459/HR4844).	○	○	○
A bill to create federal grants to support sex education programs (214/S403)	○	○	○
A bill to halt deployment of space-based missile defense systems (142/HR5122).	○	○	○
A bill to increase the minimum wage to $7.25 per hour in two years (179/S2766)	○	○	○
A bill to prohibit funds for contracts with companies that incorporate offshore to avoid U.S. taxes (275/HR5576).	○	○	○
A measure to amend the Constitution to define marriage as the union of a man and a woman (378/HJRes88)	○	○	○
A bill to permit federal funds for embryonic-stem-cell research (206/HR810)	○	○	○
Confirmation of Samuel Alito as an associate justice on the Supreme Court (1/.)	○	○	○
A bill to make it a federal crime to take a minor across state lines to obtain an abortion without parental notification or consent. (216/S403)	○	○	○
A bill to establish English as the national language and require immigrants to pass proficiency tests (131/S2611)	○	○	○
A bill to permanently reduce estate taxes (315/HR5638)	○	○	○

Table 1: Questions asked of Executive Officials in the U.S. Federal Bureaucracy.

Even if we ask career executives these questions, how should we analyze the resulting responses? One possibility would be to simply add the responses together to create a scale for the number of "liberal" (or "conservative") responses. While this may at first seem like the obvious solution, there are several complications that quickly emerge. First, it may not be obvious which outcomes are "liberal." Consider for example, responses to "A bill to ensure access to federal courts for individuals who challenge government use of eminent domain to take their property." Which is the liberal outcome? Second, even if it were possible to identify the liberal and conservative responses associated with each question, is ideology additive such that someone who responds with the liberal outcome seven times is seven times a liberal as a respondent who responds with the liberal outcome only once?

We can use the statistical measure model presented above to help answer this task. Recall that the basic model assumes that:

$$\Pr(y_{it} = 1) = F(\alpha_t + \beta_t x_i)$$

where in this case y_{it} indicates that bureaucrat *i* indicates an agreement with item t, x_i indicates bureaucrat i's most-preferred policy outcome, α_t measures the probability that bureaucrats support the policy being asked about in item t regardless of their policy preferences, and β_t reflects the extent to which the support for the policy varies depending on their most-preferred policy. Note that unlike the case of trying to construct an additive scale based on the number of "correct" answers, when we use a statistical measurement model we do not need to determine which answers are "liberal" or "conservative" beforehand. Given the model for a "yes" response, the model will use the observed responses to find the variation that best differentiates between the responses. Both "yes" and "no" answers can be associated with conservative views because the model can adjust the sign for α_t and β_t to ensure that positive values of x are associated with conservative views. (To be clear, this also illustrates a potential issue with any model – we are assuming that the variation that we recover is based on ideological differences, but nothing ensures that this is the case. An important job of the analyst is to think carefully about what the model is actually estimating when interpreting the results.)

An important limitation of the statistical measurement model is that because we are estimating everything – the only data we are using is data on the set of responses – the meaning of the scale that we recover can sometimes be difficult to interpret. The model is able to extract the characteristic that is most responsible for the variation that we observe and it is able to rank-order individuals according to that characteristic, but it is up to the analyst to interpret the meaning and importance of the characteristics that are estimated. While in this case we think it is reasonable to assume that bureaucrat's answers are largely

structured by their policy preferences, nothing in the model technically ensures that we can equate x with policy preferences as the model is interested in finding the latent trait that best differentiates between the responses of different individuals. Relatedly, the scale that is recovered does not have any real meaning apart from measuring the extent to which individuals are differentiated. While we can be certain that a policy preference of "1" differs from a policy preference of "-1", what policies someone with a policy preference of "1" would support is unclear. (In principle, we can use the statistical model to evaluate the likelihood of supporting various policy outcomes given a set of item parameters to help interpret the meaning of the scale, or we can use the estimates from a prominent political leaders to help define what differences in the estimates imply about the nature of political conflict).

If we analyze the 1,889 bureaucrats who answer at least 2 out of the 14 questions alongside members of the U.S. House and Senate by treating bureaucrats' responses as being equivalent to congressional roll call votes we get the estimates plotted in Figure 3. For each member of the U.S. House (top), U.S. Senate (middle) or career executive in the federal bureaucracy (bottom) we can use the statistical measurement model to estimate not only what their most-preferred policy position in (black dot) on a scale that ranges from very liberal (-3) to very conservative (3), but also how certain we are about each characterization (the grey horizontal line around each point).[3]

[3] Because a Bayesian methodology is used, what is graphed is actually the posterior mean for each bureaucrat as well as the region of highest posterior density (see Jackman (2009), for example, for more discussion on the use and interpretation of Bayesian statistics).

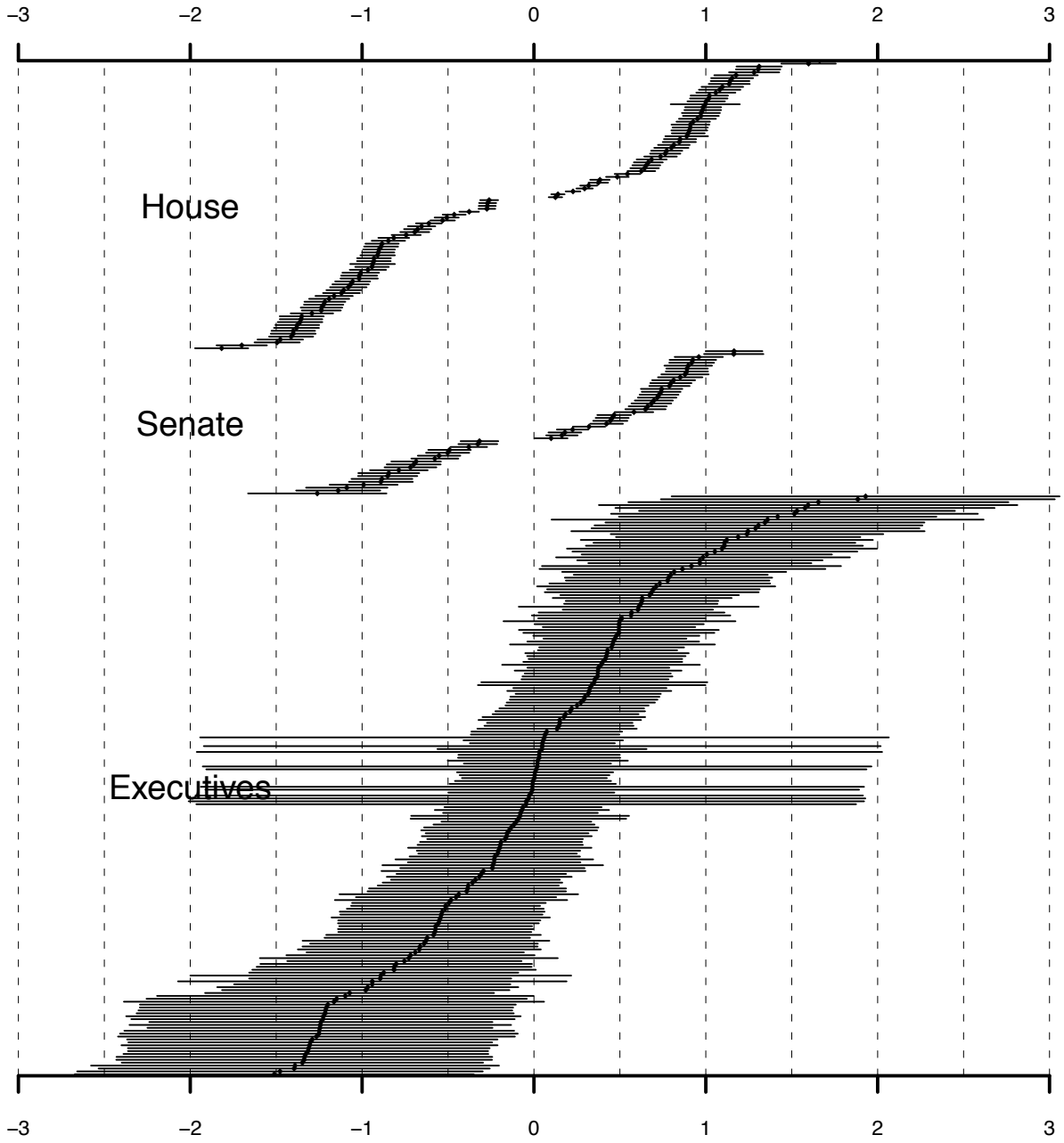

Figure 3: A Random Sample of Individual Ideal Point Means (and 95% Credible Intervals): A random sample of Representatives, Senators and executives was selected and their ideal point (along with the associated 95% credible interval).

Substantively, we can conclude several things. First, most bureaucrats are estimated to have policy preferences located near 0 based on the responses they provide, but the preferences of elected officials are typically considerably more extreme – note the absence

of any ideal points in either the House or the Senate near 0. While there are certainly some bureaucrats who are as ideologically extreme as members of Congress, most bureaucrats express views on policy that place them between the two parties in Congress regardless of whether we consider the House or the Senate. Second, while career bureaucrats appear to have more centrist policy views in general, we are far less certain about the preferences of bureaucrats than we are about members of Congress. This is largely because bureaucrats are far more likely to answer that they "Don't know" what they think than members of Congress are to miss a vote. For example, for a couple bureaucrats we are entirely uncertain as to what their most-preferred policy is and it could be anything from -2 to 2 because they answered so few questions.

Besides simply comparing how the policy views of bureaucrats compare to those of elected officials, there are a host of interesting questions and analyses that the ability to measure bureaucrats' policy preferences allows related to congressional oversight (e.g., Clinton, Lewis and Selin 2014), bureaucratic performance, and other issues that are centrally related to policymaking and government activity (e.g., Bertelli and Grose 2009). Approaches based on similar ideas can also be used to assess critical issues related to the separation of powers (e.g., Bailey and Maltzman 2011) and representation (e.g., Bafumi and Herron 2010).

Example 3. Accounting for "Expertise": Analyzing Expert Opinion

Because the statistical measurement model allows for some items to be more or less related to the underlying concept of interest – e.g., some votes may be more ideological than others and some test questions may be more related to aptitude than others - the model can also be used is to help combine and evaluate the available information. Suppose,

for example, that you are involved in a project that requires collecting the opinions of several policy experts and you must decide how to use and analyze the experts' opinions. It is certainly likely that even if the experts all agree on what is being measured, that there are still likely to be differences in the ratings that they provide. If so, you are confronted with the following critical question: do the differences in expert ratings reflect genuine differences in the outcome of interest, or are the differences due to differences in the quality (or taste) of the experts themselves? Does expert disagreement reflect differences in the quality of raters, or does it primarily reflect uncertainty about what is being rated?

To illustrate how expert responses can be analyzed using a statistical measurement model to account for the consequences of differences between raters, consider an example from Clinton and Lewis (2007). Clinton and Lewis were interested in characterizing the ideological leanings of government agencies using the knowledge of 37 experts working in universities, think tanks, and the media. Each expert was asked: "Please see below a list of United States government agencies that were in existence between 1988– 2005. I am interested to know which of these agencies have policy views due to law, practice, culture, or tradition that can be characterized as liberal or conservative. Please place a check mark (O) in one of the boxes next to each agency—''slant Liberal, Neither Consistently, slant Conservative, Don't Know.'' 26 experts responded to the inquiry and the goal was to use the experts' ratings of the 82 departments and agencies in existence between 1988 and 2005 to construct an estimate of agency ideology while allowing for the possibility that some experts may provide more useful determinations than others.

Whereas the focus on the prior section was on estimating senators' ideal points – the x's in the statistical model of section 1 – the focus here is on the estimates related to

how well the rater's determination corresponds to the relationship that is suggested by the other raters. While the actual statistical model is slightly different because the outcome is an ordered variable with 3 categories -- slant liberal, neither consistently, slant conservative – the intuition is the same as the binary model discussed above (Quinn 2004).

With several ratings and the assumption that a majority of the raters are able to correctly identify the variation of interest, the same model that is used to locate legislators can also be used to combine expert ratings and effectively "rate" the raters (see, for example, Peress and Spirling's (2010) work evaluating movie critics or Clinton and Lapinski's (2006) work identifying notable legislation). Building on the intuition of the prior example, instead of thinking about senators casting votes and using the votes that are cast to make inferences about the senators, we can think about the experts' views as the "votes" that are being voted upon by the objects we are interested in evaluating. That is, we can think of the agency being rated by the experts as the elite official in the earlier examples indexed by i and the rating of a particular expert t as the observed vote. In terms of the measurement model:

$$\Pr(y_{it} = "Slants\ Liberal") = F(\alpha_t + \beta_t x_i)$$

where in this example each expert t is equivalent to a test question and we are interested in learning about the agency's ideology x_i from these "test questions" while accounting for the fact that raters may differ in their standards and quality even if they are all responding to the same latent input.

As before, the model produces a series of estimates (and estimates about how certain we are about those estimates for each agency). Table 2 reports some of those estimates arranged from most liberal (negative) to most conservative (positive). Some

agencies are clearly thought to be more liberal in their policy mandate than others – e.g., the Peace Corps is unambiguously more liberal than the Department of Defense – and other agencies (e.g., the Department of Transportation and NASA) are thought to be non-ideological.

Agency	Mean	2.5% CI	97.5% CI
Peace Corps	-1.72	-2.49	-1.02
Consumer Product Safety Commission	-1.69	-2.42	-0.99
Equal Employment Opportunity Commission	-1.58	-2.28	-0.97
Occupational Safety and Health Review Commission	-1.52	-2.25	-0.82
Department of Labor	-1.43	-2.03	-0.81
Department of Housing and Urban Development	-1.33	-1.93	-0.80
Department of Health and Human Services	-1.32	-1.91	-0.78
Department of Education	-1.22	-1.78	-0.75
Environmental Protection Agency	-1.21	-1.74	-0.72
National Foundation on the Arts and the Humanities	-1.00	-1.52	-0.54
Social Security Administration	-0.45	-0.78	-0.10
National Labor Relations Board	-0.27	-0.58	0.05
Department of State	-0.27	-0.58	0.04
National Aeronautics and Space Administration	-0.07	-0.38	0.24
Federal Election Commission	0.05	-0.27	0.35
Department of Transportation	0.07	-0.23	0.36
Federal Emergency Management Agency	0.08	-0.19	0.37
Federal Trade Commission	0.12	-0.19	0.47
Department of Agriculture	0.16	-0.16	0.50
Department of Veterans Affairs	0.23	-0.12	0.55
Federal Communications Commission	0.32	0.02	0.62
Department of Energy	0.35	0.01	0.68
Department of Justice	0.37	0.05	0.67
Trade and Development Agency	0.40	-0.34	1.18
Department of the Interior	0.47	0.14	0.81
Securities and Exchange Commission	0.73	0.38	1.11
Department of Homeland Security	0.88	0.51	1.26
Department of the Treasury	1.07	0.68	1.48
Small Business Administration	1.17	0.72	1.67
Department of Commerce	1.25	0.80	1.75
National Security Council	1.40	0.91	1.87
Office of National Drug Control Policy	1.77	1.04	2.47
Department of Defense	2.21	1.49	3.06

Table 2: Expert Ratings of Average Agency Policy Preferences, 1988-2005: Agencies are ordering from most liberal to most conservative according to (posterior) mean estimate. 95 % HPD Interval also reported. Estimates reflect the ratings of 26 experts.

In this case, however, we are also likely interested in assessing the performance of the individual experts that were used to construct the rating. To do so, we can compare the how the probability that each expert rates an agency as "slants Liberal," "neither," or "slants Conservative" for every possible value of agency ideology relative to the rating given by other experts. This allows us to see how the standards of different experts compare to the average rating and whether some experts are behaving differently from the others. So-called "characteristic curves" describe the probability that a rater will place an agency with a given true ideology into each of the possible classifications. Note that similar analysis can be performed in any of the previous examples – we can see how well each item/vote/question relates the latent quantity being studied into the probability of observing a positive response (e.g., a "correct" vote, a "yea" vote, or a "liberal" survey response").

Figure 4 plots the characteristic curves for 4 selected experts.

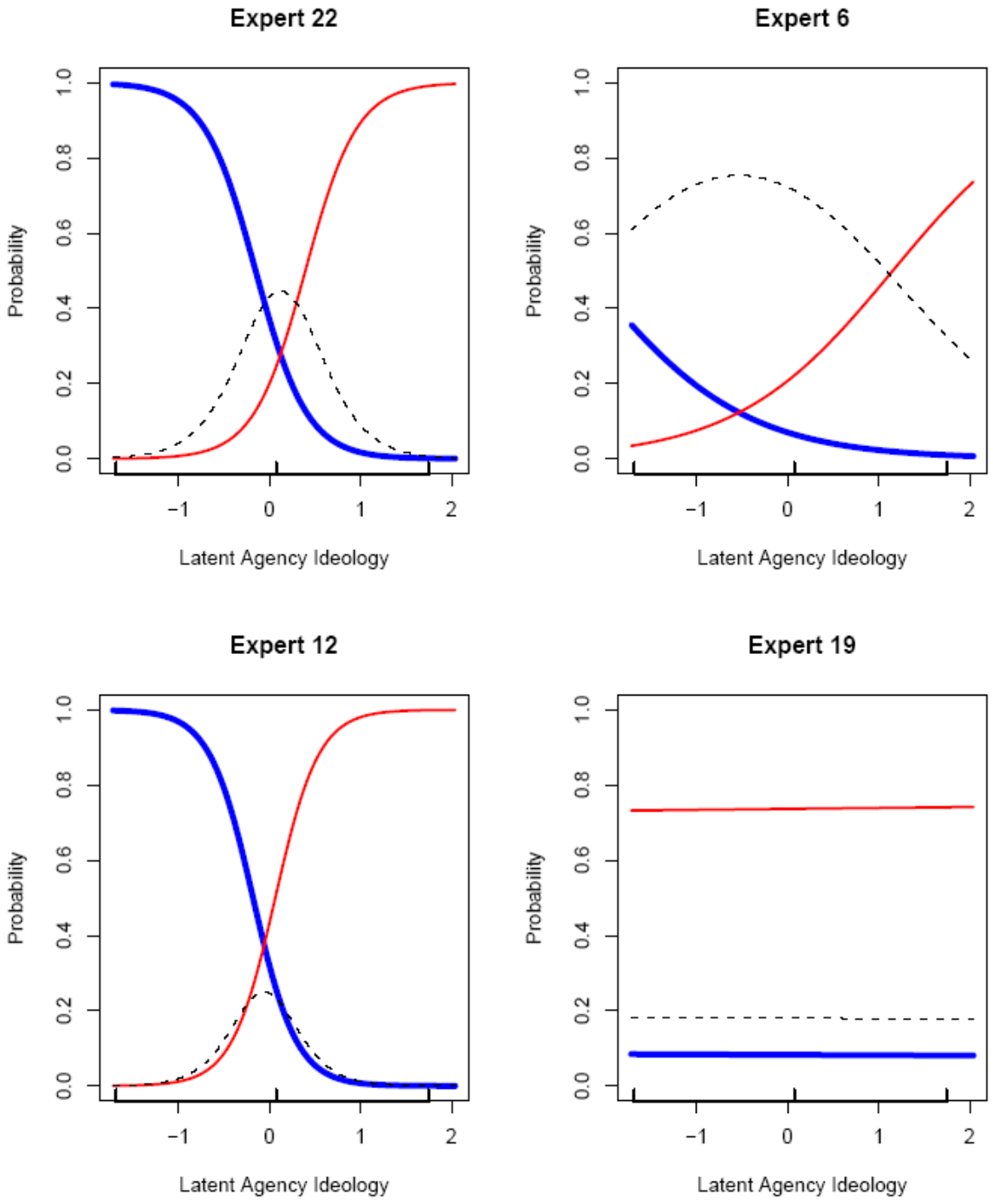

Figure 4: Selected Characteristic Curves for Experts: the thickest line denotes the probability of the expert designating an agency with the given policy preference as "slants Liberal", the dashed line denotes the probability of the expert indicating "Neither" and the slender solid line denotes the probability of a "slants Conservative" rating.

Figure 4 reveals significant variation across the various experts in terms of their scoring of agency ideology. The x-axis describes the ideology of a true agency, and the y-axis describes the probability of that expert providing each of the 3 possible rankings (which must sum to 1 for a selected agency ideology). For example, if we consider the rankings of Expert 22, if the agency's true ideology is -1, the expert has roughly a 95 percent chance of rating the agency as ``slants liberal", roughly a 5 percent chance of ``neither consistently" and a fraction of a percentage of rating the agency "slants conservative." However, if the agency's true ideology is 0, then the expert is most likely to rate the agency as "neither consistently" (roughly a 40-percent chance), and they are equally less likely to rate the agency as either "slants liberal" or "slants conservative' (30 percent chance of each).

Expert 12's rankings are very similar to those of Expert 22, but they are always more likely to choose either "slants liberal" or "slants conservative" than "neither consistently" – even if the true agency ideology is 0, "neither consistently" is never the most probable ranking for this expert. In contrast to Expert 12 and Expert 22, Expert 6 is most likely to respond "neither consistently" for every agency whose true ideology is less than 1. Finally, Expert 19 is most likely to respond "slants conservative" regardless of the agency ideology – their ranking is completely unresponsive to the true agency ideology.

Collectively, given the set of rankings we receive, the views of expert 19 stand apart from the others in that the ratings of expert 19 appear to have no relationship to the quantity we are interested – not only is there no systematic variation in the rankings they report, but there is also no relationship between the ideology of an agency in the views of the other experts and expert 19's rankings. In contrast, for the other three experts as the

true ideology of the agency increases, so too does the probability of observing a "slants conservative" response.

Thus, the statistical measurement model provides one way to aggregate the opinions of experts in a way that allows us to determine if some experts are more informative than others. In hindsight, for example, we can see that Expert 22's determinations are far more informative than those of Expert 19 (or even Expert 6) in terms of providing rankings that vary in ways that reflect the true ideology of agencies. This is important because if we are interested in using the ratings to assess agency ideology we may wish to account for the differential ability of experts.

To be clear, we can only infer what agency ideology is likely to be based on what underlying pattern best explains the variation we observe in the ratings themselves and then characterizing experts according to whether or not their individual ratings vary along with the agency ideology that we are inferring. As a result, if most of the ratings themselves are completely unrelated to the concept we are interested in the model, we will not be able to magically recover what we are interested in. The statistical model is aimed at extracting a parsimonious structure from the observed data, but if the observed data is not meaningful then so too will be estimates that are extracted. Put differently, while the model provides an ability to determine which experts may be better positioned than others, this assessment is based on an assessment of how well the expert's ratings compare to the underlying structure based on the collective responses – not some objective criteria. As a result, if most of the ratings are ill-informed or even mistaken, the statistical model will be unable to identify the few experts who have it right. If, however, most of the experts

are able to get it right but some may be better than others, then the model can identify which experts are best able to characterize the variation.

Conclusion:

The availability of data has had a tremendous impact on both the study and practice of public policy. Individuals interested in characterizing outcomes or the circumstances surrounding the adoption, implementation and alteration of public policies now frequently have a large number of observable features than can be used to reach data-driven conclusions and assessments. Despite the availability of data, however, it is often still the case that either we have multiple measures of the feature of interest and it is unclear as to which measure is unambiguously "best," or else we have multiple measures that are related to the concept we are interested in but we lack a direct measure of that concept.

In either of these circumstances, a statistical measurement model can be useful for helping to analyze the structure of the observed data and extracting the core tendencies. Moreover, the model is suitably flexible so that the same underlying model can be readily applied to a variety of different situations; the examples discussed in this chapter, for example, show how it can help in analyzing: questions on an exam, a series of elite decisions, an opinion survey, and the ratings provided by experts. While there are many details involved in the implementation of statistical measurement model that the chapter omits, hopefully you get a sense of what such models can offer to those interested in analyzing data relevant for public policy.

That said, it is important to emphasize that while the statistical models are powerful in terms of being able to provide insights that may otherwise be either elusive or unwieldy, the results are only as good as the assumptions of the statistical measurement model. The

assumptions made by the model are certainly less stringent than those that are implied by attempts to aggregate measures by treating every item as equally informative and taking an average or a sum, but we can only learn about traits that are already in the data structure. Relatedly, while the statistical measurement models are able to generate estimates of the latent trait (labelled *x* in the examples above), it does not reveal what the actual meaning of *x* is and it is incumbent on the analyst to interpret what *x* means substantively. For example, does the x estimated from elite behavior in example 1 above reflect the personal policy preferences of the elite or does it reflect other factors (e.g., the policy preferences of voters)? Does the *x* being measured by responses to a series of test items measure just aptitude or does it also reflect educational opportunities (or aptitude given the educational opportunities of the test-taker)? Using a statistical measurement model is only the first step in thinking through what the results do or do not reveal and the critical issues related to the interpretation of what is found are issues that cannot be "solved" statistically.

If careful attention is given to what the estimated parameters actually mean, statistical measurement models can provide a powerful tool to those who are interested in analyzing the wealth of data that now often surround issues involving public policy. When appropriately estimated and interpreted, these models provide the ability to summarize underlying features of the data and to characterize aspects that are important but unobserved to make novel conclusions and characterizations about the policymaking process and its outcomes.

Works Cited

Ansolabehere, Stephen, James M. Snyder Jr., and Charles Stewart III. 2001. "Candidate Positioning in U.S. House Elections," *American Journal of Political Science* 45(1): 136-159.

Ansolabehere, Stephen, James M. Snyder Jr., and Charles Stewart III. 2001. "The Effects of Party and Preferences on Congressional Roll-Call Voting," *Legislative Studies Quarterly*, 26(4): 533-572

Aberbach, Joel D., and Bert A. Rockman. 2000. *In the Web of Politics*. Washington, DC: Brookings.

Bafumi, Joseph, and Michael C. Herron. 2010. "Leapfrog Representation and Extremism: A Study of American Voters and their Members of Congress." *American Political Science Review* 104(3):519-42.

Bailey, Michael A. and Forrest Maltzman. 2011. *The Constrained Court: Law, Politics, and the Decisions Justices Make*. Princeton, NJ: Princeton University Press.

Baker, Frank B. and Seock-Ho Kim, 2004. *Item Response Theory: Parameter Estimation Techniques*, Second Edition. CRC Press.

Bertelli, Anthony M., and Christian R. Grose. 2006. "The Spatial Model and the Senate Trial of President Clinton." *American Politics Research* 34:4:535-559.

Bertelli, Anthony M. and Christian R.Grose. 2009. "Secretaries of Pork? A New Theory of Distributive Politics", *Journal of Politics* 71(3): 926-45.

Bertelli, Anthony M., and Christian R. Grose. 2011. "The Lengthened Shadow of Another Institution? Ideal Point Estimates for the Executive Branch and Congress." *American Journal of Political Science* 55:4 (October).

Bollen, Kenneth A. 1989. *Structural Equations and Latent Variables*. Wiley: NY,NY.

Brady, Henry E. 1985. "The Perils of Survey Research: Inter-Personally Incomparable Responses," *Political Methodology*, 11(3/4): 269-291.

Clinton, Joshua D., David E. Lewis, and Jennifer L. Selin. 2014. "Influencing the Bureaucracy: The Irony of Congressional Oversight." *American Journal of Political Science* 58(2): 387-401.

Clinton, Joshua D., Anthony Bertelli, Christian Grose, David E. Lewis and David C. Nixon. 2012. "Separated Powers in the United States: The Ideology of Agencies, Presidents and Congress." *American Journal of Political Science* 56(2): 341-354.

Clinton, Joshua D., and John S. Lapinski. 2006. "Measuring Legislative Accomplishment, 1877-1994," *American Journal of Political Science* 50(1): 232-249.

Clinton, Joshua D., Simon Jackman and Doug Rivers. 2004. "The Statistical Analysis of Roll Call Voting: A Unified Approach." *American Political Science Review* 98(2) 355-70.

Clinton, Joshua D., David E. Lewis. 2008. "Expert Opinion, Agency Characteristics, and Agency Preferences." *Political Analysis* 16(1):3-20.

Gelman, Andrew and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. NY, NY: Cambridge University Press.

Harman, Harry H. 1976. *Modern Factor Analysis*, 3rd Edition. University of Chicago Press.

Jessee, Stephen. 2009 "Spatial Voting in the 2008 Presidential Election," *American Political Science Review* 103(1): 59-81.

Jackman, Simon. 2008. "Measurement," in The Oxford Handbook of Political Methodology, Eds: Janet M. Box Steffensmeier, Henry E. Brady, and David Collier. Oxford University Press: NY,NY.

Jackman, Simon. 2009. *Bayesian Analysis for the Social Sciences*, NY,NY: Wiley.

Jackman, Simon. 2014. *pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University.* Version 1.4.6. http://pscl.stanford.edu/

Johnson, Valen E., and James H, Albert. 2000. *Ordinal Data Modelling*, Springer: NY,NY.

Levendusky, Matt. S., Pope Jeremy. C., & Jackman Simon. 2008. "Measuring District-Level Partisanship with Implications for the Analysis of U.S. Elections," *Journal of Politics*. 70(3), 736-53.

Maranto, Robert, and Karen M. Hult. 2004. "Right Turn? Political Ideology in the Higher Civil Service, 1987-1994." *American Review of Public Administration* 34:199-222.

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* 10:134-153.

Martin, Andrew D., Kevin M. Quinn, and Jong Hee Park. 2011 "MCMC pack: Markov Chain Monte Carlo in R," *Journal of Statistical Software*. 42.

McCarty, Nolan, Keith T. Poole, and Howard Rosenthal. 2006. *Polarized America: The Dance of Ideology and Unequal Riches*. Boston, MA: MIT Press.

Pemstein, Dan, Stephen Meserve, and James Melton. 2010. "Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type." *Political Analysis* 18 (4), 426–449.

Peress , Michael, and Arthur Spirling. 2010. "Scaling the Critics: Uncovering the Latent Dimensions of Movie Criticism With an Item Response Approach," *Journal of the American Statistical Association*, 105(489): 71-83.

Poole, Keith, and Howard Rosenthal. 2007. *Ideology and Congress*. New Brunswick, NJ: Transaction Press.

Quinn, Kevin M. 2004. "Bayesian Factor Analysis for Mixed Ordinal and Continuous Responses," *Political Analysis* 12: 338:353.

Selin, Jennifer L. 2014. "What Makes an Agency Independent?" *American Journal of Political Science*.

Shor, Boris, and Nolan McCarty. 2001. "The Ideological Mapping of American Legislatures," *American Political Science Review* 105(3): 530-551.

Spearman, Charles. 1904 "General Intelligence,' Objectively Determined and Measured," American Journal of Psychology, 15:201-93.

Volden, Craig and Alan E. Wiseman. 2014. *Legislative Effectiveness in the United States Congress: The Lawmakers*. NY,NY: Cambridge University Press.

Wawro, Gregory J., and Eric Schickler. 2007. *Filibuster: Obstruction and Lawmaking in the U.S. Senate*. Princeton, NJ: Princeton University Press.