International Review of Law and Economics 68 (2021) 106020

Contents lists available at ScienceDirect

International Review of Law and Economics

The efficacy of measuring judicial ideal points: The mis-analogy of IRTs

Joshua Y. Lerner[a,*], Mathew D. McCubbins[b], Kristen M. Renberg[c]

[a] NORC at the University of Chicago, United States
[b] Duke University, Department of Political Science and School of Law, United States
[c] Duke University School of Law, United States

ARTICLE INFO

Article history:
Received 18 August 2020
Received in revised form 24 August 2021
Accepted 25 August 2021
Available online 28 August 2021

Keywords:
Ideal points
Measurement
Judicial politics
Item response theory
Replication

ABSTRACT

IRT models are among the most commonly used latent trait models in all of political science, particularly in the estimation of ideal points of political actors in institutions. While widely used, IRT models are often misapplied, and a key element of their estimation, the item parameters, are almost always ignored and discarded. In this paper, we look into the application of IRT models to the estimation of judicial ideology scores by Martin and Quinn (2002). Building off of a replication and extension of Martin and Quinn (2002), we demonstrate that the often-ignored item parameters are, in fact, inconsistent with the assumptions of IRTs. Then, using a post-estimation fix that is designed to ameliorate the problem, we run the model again, generating new scores. We then compare our new ideal points to the existing ideal points and discuss the implications for both ideal point modeling generally and in judicial politics specifically. We conclude by replicating a prominent study in judicial politics that demonstrates how inconsistencies in the estimation of IRT models can be consequential and bring up concerns with the implications for what this could mean for the usefulness of scores estimated via IRT models.

1. Introduction

Latent concepts and measurements are ubiquitous in political science, law, and economics. Because the core assumption in any paradigm of thought must be about concepts that we fundamentally cannot observe (Lakatos, 1976), such as preferences, ideology, or aptitude — there are extensive literatures devoted to how precisely we can measure their shadow on the real world, such as by observing people's choices. Since the highly influential work of Poole and Rosenthal (1991, 2000), numerous scholars have used spatial voting models to estimate ideological preferences from roll-call votes and other choice data; these measures and models are often vital to significant swaths of the literature in American and comparative politics (Imai et al., 2016) and law.

A widely used method to measure these latent traits is the item response theory model (IRT). These models, borrowing from literature developed in education testing and psychometrics, are a relatively fast way of fitting a latent trait model on diverse sets of choice data that arise from different choice environments. IRT methods have been applied to study: public opinion formation (Treier and Sunshine Hillygus, 2009; Tausanovitch and Warshaw, 2014), the ideology of actors in political institutions (Clinton et al., 2004; Martin and Quinn, 2002; Bonica, 2014), aggregation of expert rating models (Clinton and Lapinski, 2006; Treier and Jackman, 2008; Linzer and Staton, 2012) among many others.

Many of these approaches, however, disregard an important output of the IRT model that is vital to understanding the model's fit. Because of this disregard, IRT models are often used in a way that violates the fundamental assumptions of the IRT models and the estimates of latent traits. In this paper, we argue the main limitation to existing IRT models used in service of ideal point estimation in judicial politics is a misspecification of item parameters.[1] We will show that the choice of test questions to measure unobserved aptitude, the purpose of which IRT was created, may not be a good analogy to voting and ideology. In particular, we will show that we need to pay better attention to the 'item' part of the IRT models if we are to limit (or eliminate) bias in our measures.

* Corresponding author.
E-mail address: joshlerner1@gmail.com (J.Y. Lerner).

[1] We also highlight two other issues with existing applications of IRT models the first is the potential implicit multidimensionality of voting space and the second is the overall limits in our ability to turn nominal data (votes) into cardinal or ratio-level ideal point estimates.

https://doi.org/10.1016/j.irle.2021.106020
0144-8188/

In the sections that follow, we present and discuss the exact formulation of IRT models, how they are used in political science and legal scholarship, and the development of the Martin and Quinn's (2002) Dynamic Ideal Point Model, the most widely used ideal point model in the study of judicial politics and law. We then demonstrate the bias in their estimates and correct these estimates for the problems we identify. We then follow this by replicating two recent studies of judicial behavior using our new judicial ideology scores. We then show that the specification and construction of judicial ideology scores do indeed impact empirical models of judicial behavior. We follow this with our reservations about assuming that judicial ideology is unidimensional and we suggest improvements that can be made within the political science latent scale modeling literature more broadly.

2. The mis-analogy of applying IRT models in politics and law

2.1. IRT by the numbers

Item response theory methods are a class of latent trait models developed in education testing to capture test taker aptitude along with an assessment of a questionnaire (i.e., test or exam) (van der Linden and Hambleton, 2013). Item response models provide the basis for most modern standardized tests and represent improvements made to the education testing literature from classical test theory over the last seven decades. The broad adoption of IRT models in psychometrics and education testing is primarily due to two important features of this class of models: they are highly flexible when it comes to specifications, and they model the difficulty and utility of any given test question, which allows educators and employers to choose test questions that capture the latent aptitude they want to measure. This means that, unlike in classical test theory, tests constructed using an IRT framework have an adaptive approach to question weighting and force all questions to be evaluated on their ability to discriminate between test-takers based on their latent aptitude or skill (De Boeck and Wilson, 2004; Fox, 2010; van der Linden and Hambleton, 2013).

In education testing and psychometrics, IRT models are utilized primarily for the ease with which they fit what is thought of as the "skill" parameter for test-takers. This framework creates some problems given how it is used in social science and legal research because the models are not designed to scale infinitely (see Tahk, 2018 for a nonparametric solution to this problem). While studying the consistency of ideal point models, Londregan (1999) linked the psychometrics testing literature with the spatial theory of legislative voting in order to derive important statistical insights. In particular, Londregan (1999) demonstrates that when the preferential choices are nominal (as in voting 'yes', 'no', 'partially yes' or 'partially no') consistency of ideal points in its usual statistical sense is not possible. With nominal choices, maximum likelihood estimators that attempt simultaneously to recover legislators' ideal points and roll call or item parameters must inherit the granularity of the choice data and so cannot capture the (supposed) underlying continuous parameter space.[2] Londregan (1999) solves this identification problem in IRT models by using constraints on underlying ideal points of certain legislators, which induces additional variability to all spatial IRT models. This added constraint provides additional subjectivity to the analysis of IRT models in judicial politics. Namely, the estimates generated by even the best possible model will ultimately be contingent on certain arbitrary elements that make specific point estimations inconsistent.

Indeed, Ho and Quinn (2010) go as far as to argue that the ordinality of these ideal point measures, by virtue of being consistent, should be the main takeaway from these scores rather than the starting point. The specific location along any ideological line is compromised by the specifics of the subjective aspects of any modeling choices made. Ho and Quinn (2010, pg. 847) go on to argue that "the cardinal scale of the latent dimension is not identified from the data. In the standardized testing analogy, we have no sense of whether the 100-point difference between a score of 2100 and 2000 should be the same as the difference between 2300 and 2400. While prior assumptions about cutpoints affect such cardinal scaling, they generally do not affect the relative ranks." What we take from this is even more basic: we cannot and should not believe that we can generate ratio-level output from a model with nominal-value inputs. You cannot make a silk purse from a sow's ear.

Moreover, with regards to applications of IRT models to the Supreme Court, the underlying IRT models were never intended to be estimated over such a small selection of 'test-takers'. Given the complexity of the traits often modeled in political science and law, such as the ideology of legislators, presidents, bureaucrats, judges, and justices — it is no wonder the discipline has had to stretch the psychometrics of IRTs.

2.2. Item parameters and IRT models

Within an IRT model, an individual's latent ability and the characteristic function or curve of each item (which defines the relationship between individual's ability and the probability they answer the item correctly) are jointly modeled to predict the performance of individuals across a range of abilities (Baker and Kim, 2004, 2017; Baker, 2001; Lalor et al., 2016). Ideally, the items should be free-response items graded dichotomously as correct or not. The two-parameter IRT model takes into account the difficulty of an item and how well an item discriminates among individuals. The function of two-parameter (2-PL) IRT models is presented in a standard mathematical model known as the item characteristic curve (ICC). The ICC is the cumulative form of the logistic function and represents an individual's ability (sometimes referred to as the aptitude) to answer an item correctly (Baker and Kim, 2004, 2017; Baker, 2001). The equation for a two-parameter IRT model is presented below, with and $\boldsymbol{\beta}$ representing the two parameters:

where

$\boldsymbol{e}$ is the base of the natural logarithm that is a constant 2.781,

$\boldsymbol{\alpha}$ is the item discrimination parameter,

$\boldsymbol{\beta}$ is the item difficulty parameter, and

$\boldsymbol{\theta}$ is the latent ability.

The item difficulty parameter, denoted as $\boldsymbol{\beta}$ above, is defined as the point on the scale of ability (θ) where the probability of correct response to a given item is $P(\theta) = 0.5$.[3] More simply, the difficulty parameter ($\boldsymbol{\beta}$) indicates the level of ability (θ) a test-taker needs in

[2] The usual argument given by Poole (2005) regarding the roughly similar estimation of Nominate, is that we can count the 'yes' and 'no' votes and divide the number of 'yes' votes by the number of total votes. While we will defer an argument about whether this means that we can perform higher level mathematics on the counts (Poole would argue since we calculate the total number of 'yes' votes and the total number of 'no' votes we already have ratio-level input), or not, and whether this works on legislative votes (Poole's target), we will only point out that judicial votes are not just tallies of votes for and against the same item (such as a bill or amendment), but rather justices sign onto majority opinions, concurring opinions, dissents and so on. To assume these items are then all on the same dimension, measured with the same yardstick, as all the other is a much tougher argument. Judicial politics and legal scholars may want to consider assuming that concurrences (no matter how disagreeable) and partial concurrences (no matter what part is dissent and what is a concurrence) and other forms are the same as a vote for the majority opinion (Lerner et al., 2019).

[3] Theoretically, this parameter can range from $-\infty$ to ∞. However, this parameter typically ranges from -3 to 3 (Baker and Kim, 2004, pg. 18). Also, latent ability is set to have a mean of zero and a standard deviation equal to one.

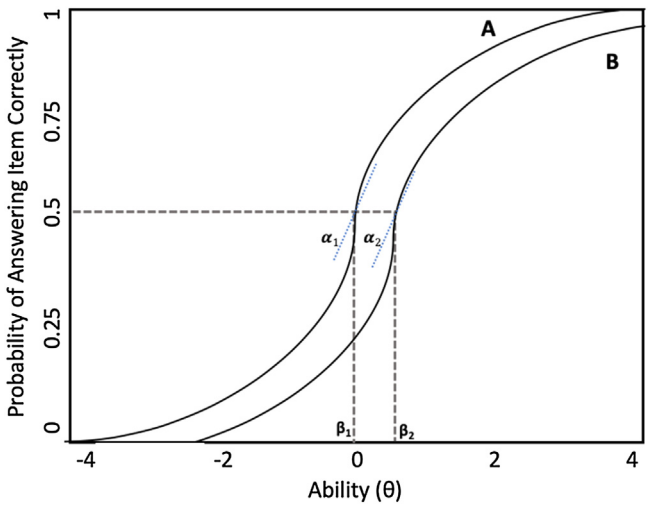

Fig. 1. Two Items with the Same Discrimination Parameters and Different Difficulty Parameters.

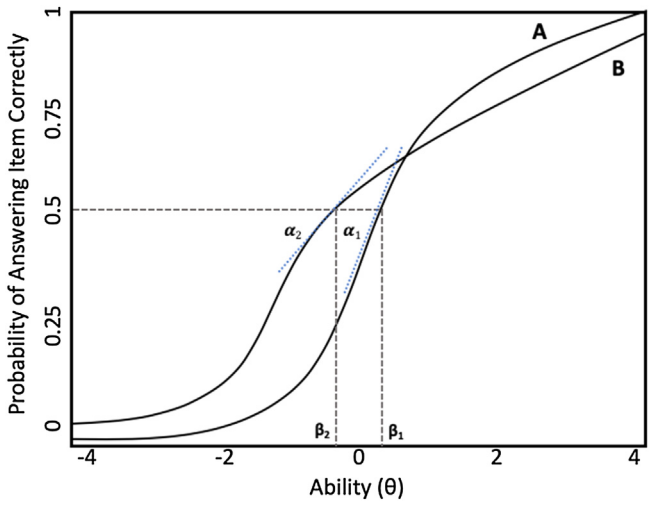

Fig. 2. Two Items with Different Discrimination Parameters and Different Difficulty Parameters.

order to have a 50 % chance of answering the question correctly. The discrimination parameter, $\boldsymbol{\alpha}$, is a scalar multiple of the slope of the curve (in the ICC) where $\theta = \boldsymbol{\beta}$ and $P(\theta) = 0.5$. Items with discrimination parameters that have high values are better at differentiating among individuals as those with lower ability are more likely to answer incorrectly than those with higher ability. We should also note the equation above is for a two-parameter IRT model, and there are other more nuanced IRT models that include constructs such as a 'guessing parameter'.

In general, scholars in political science and law have focused on using estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to calculate the latent trait (θ) — as it is this parameter that can provide researchers with an estimate of the latent ideology of a judge or justice. The rest of the IRTs' output, the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, are then unexplored and disregarded. In turn, the behavior of the item parameters in IRT models – specifically the discrimination parameter ($\boldsymbol{\alpha}$) is essentially overlooked in political science literature. When scholars ignore the item parameters, they are essentially throwing out much of the relevant information derived from an IRT model.[4] The following figures depict the relationship between ability (θ), ICC (items A and B), and the item parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.[5]

There are two items (test questions), A and B, whose ICCs are presented as S curves in Fig. 1. Both items A and B have the same slope, suggesting that the items have identical discrimination parameters ($\alpha_1 = \alpha_2$). However, we can see the two items have different difficulty parameters – as the point where an individual has a 0.5 probability of answering the item correctly has a lower ability score (θ) for item A than it does in item B. This implies that it takes more ability to answer item B correctly as compared to item A (where $\beta_1 < \beta_2$). For both items A and B, we should note that as an individual's ability increases, so does the probability they answer item A and item B correctly.

In Fig. 2, there are again two items, A and B, whose ICCs are presented. Items A and B have different discrimination and difficulty parameters in Fig. 2. Item A has a higher difficulty parameter than item B, as more ability (θ) is needed for an individual to have a 0.5 probability of answering the item correctly. Likewise, item A has a steeper slope at α_1, implying that it is able to discriminate better among individuals. Again, we note that the ICCs drawn in Fig. 2 indicate that as an individual's ability increases, so does the probability they answer item A and item B correctly.

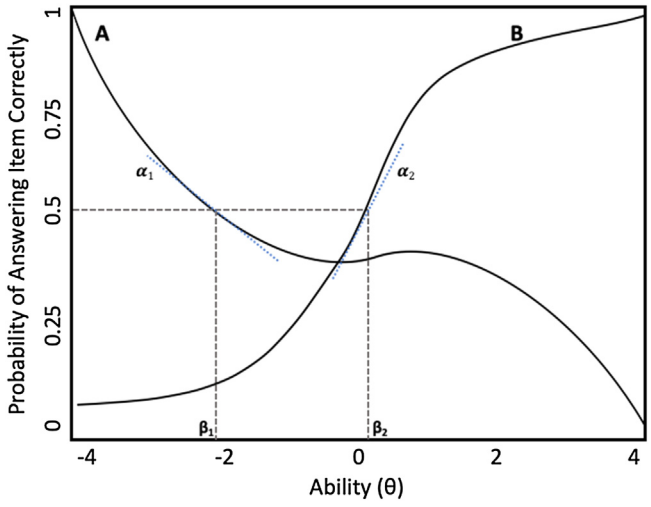

Fig. 3. Two Items with Different Discrimination Parameters and Different Difficulty Parameters, with $\boldsymbol{\alpha}_1 < 0$ and $\boldsymbol{\alpha}_2 > 0$.

In Fig. 3, we once again present two items, A and B, and their ICCs. Items A and B continue to have different discrimination and difficulty parameters. However, item A suffers from negative discrimination ($\alpha_1 < 0$) – where the probability of a correct answer actually decreases as ability (θ) increases.[6] Baker and Kim (2004, pg. 25) indicate that there is something wrong with any item with negative discrimination – as the item is either poorly written or there is some kind of misinformation among highly-able individuals. Item B in Fig. 3 has a positive slope and therefore has a positive discrimination parameter ($\alpha_2 > 0$), implying that as an individual's ability (θ) increases, the more likely they will answer item B correctly.

2.3. Item parameters and IRT models in judicial politics

Beyond the interpretability of the item parameters, we should note that scholars have also glossed over the set of assumptions

[4] De Boeck and Wilson (2004) provide examples of how item parameters are utilized in the psychometrics field.

[5] With θ ranging from -3 to 3, $\boldsymbol{\alpha} = 0.5$, and $\beta = 1$. The implied probability of a correct answer ranges from $P(-3) = 0.12$ to $P(3) = 0.73$.

[6] Ho and Quinn (2010) refer to this concept of negative discrimination as 'reverse discrimination'.

that underlay IRT models. First, this class of models assumes that individuals differ from each other on an unobserved latent trait dimension (generally referred to as ability) measurable for everyone on the same unidimensional metric and the probability of correctly answering an item is a function of an individual's latent trait (Lalor et al., 2016, pg. 2; Baker and Kim, 2004; Baker and Kim, 2017).[7]

Second, in an IRT model, the discrimination item parameter ($\boldsymbol{\alpha}$) is monotonic and has the same sign for all ICCs. Therefore, there can be no reversals and all $\boldsymbol{\alpha}$ values must be in the same direction (Jackman, 2001; van der Linden and Hambleton, 2013). Without a unidirectional set of $\boldsymbol{\alpha}$s, making comparisons between items is impossible in that the model cannot "discriminate" between the various skill levels. Fig. 3 above provides an example of the ICCs when this assumption is violated in an IRT model.

From this, we propose the following: if we are to take computation of θ's seriously, like most judicial politics studies that use judicial ideology scores by Martin and Quinn (2002) or any congressional study which computes ideology scores of legislators (e.g., Clinton et al., 2004), then the item parameters must correspond to something meaningful as well. More succinctly, we argue that if the item parameters in an IRT model do nothing (that is there are many items with exactly the same $\boldsymbol{\alpha}$), or behave in incorrect ways (such as when the $\boldsymbol{\alpha}$ of some items is positive and the $\boldsymbol{\alpha}$ of some items is negative) — then we cannot trust the estimates of θ.

Within the original estimation of Martin and Quinn (2002), most cases resolved by the Supreme Court (the items) in the IRT model are estimated to have negative $\boldsymbol{\alpha}$ values. As such, we believe the cases resolved by the Supreme Court which receive positive $\boldsymbol{\alpha}$ values in the original estimation by Martin and Quinn (2002) potentially capture variance from a different underlying latent trait as compared to the other cases, and not the unidimensional ideal point captured by the estimated θ. If this holds, then the reversals of the discrimination parameters is particularly troubling for studies built on spatial theories of voting, as these require a unidimensional space (e.g., Poole and Rosenthal, 2000).

Our claim here is simple: because Martin and Quinn's (2002) approach generates both positive and negative values for $\boldsymbol{\alpha}$, estimations made using these items violate an assumption of the IRT model and are thus flawed. Our solution here to overcome the problem of reversals is to exclude the cases (items) which received positive $\boldsymbol{\alpha}$ values in the initial estimation and then re-estimate the IRT model on the remaining cases. This procedure should lead to more consistent estimates of θ.[8]

Why our proposed solution works: the inherent logic of test design. The utility of any test question in an IRT model is how well it allows us to distinguish among the test takers. Within the original psychometric setting of IRT models, if a test question is performing irregularly or getting results that are inconsistent with known behaviors of good test questions, the designers of the test will remove the questions after the fact (van der Linden and Hambleton, 2013; Baker and Kim, 2004). There are two main reasons elimination may work as a post-estimation technique. First, to a certain extent, we already engage in this behavior by omitting cases that add no information to the model (i.e., unanimously decided cases) in order to make the model more identifiable. So, in a sense, we already engage in an elimination procedure at the design stage of the IRT model.

The second reason to engage in a post-estimation reanalysis is that it allows the data to determine which parts of the process are problematic and which require additional attention. Whatever approach we take, we are altering the basic relationship between the θ values and the items. That said, we argue that our post-estimation reanalysis approach is less problematic than allowing biased results to be treated uncritically. Again, if we were to keep the metaphor of IRT models as assessments of tests and test-takers, then the use of iterative trimming of irregularly behaved test questions is the standard. If we believe that the cases in the IRT model are doing a poor job of measuring the construct we care the most about, then removing them is the best way to keep the integrity of the test intact.

3. Martin and Quinn (2002)

The primary work we address is one of the most essential IRT-based ideal point models for political science and law: the Dynamic Ideal Point Model by Martin and Quinn (2002). For this paper, we focus on the Dynamic Ideal Point Model for two reasons: first, it is widely cited and has provided the basis for much of the modern study of judicial behavior. The second reason is that it is a very believable model that makes many reasonable assumptions. Thus, while we focus our attention on the ideal points of Supreme Court justices generated by Martin and Quinn (2002) and some of the most notable pieces of scholarship which use these scores to examine judicial behavior, we think of this paper as more of a general critique of the spatial model of ideology as applied to justices on the Supreme Court rather than just to Martin and Quinn (2002) specifically.

The Dynamic Ideal Point Model by Martin and Quinn (2002) is a modification of a classic two-parameter IRT model, with three significant differences. First, the model is set up in an entirely Bayesian framework, allowing for distributional flexibility in the estimation of the parameters (subject to prior knowledge). This means that they allow for MCMC simulations to generate the posterior estimates of the parameters of interest, which means that the error distribution is descriptive rather than prosaic (unlike in classic frequentist models), the specifications of the distributions are flexible, and prior knowledge is allowed to be used in estimating model fit.[9]

Second, unlike most models of latent traits, the θ is allowed to vary over time (modeled as a random walk between terms), which means that the trait is not fixed to a given rate over all the items seen, but instead changes every term on the Supreme Court, and is therefore allowed to drift from one year to another. Linking between terms is done by setting priors for each term to the value of θ estimated in the previous year. Third, the underlying link function is not a standard logistic regression, but a dynamic linear model that allows for general smoothing over time.

The Dynamic Ideal Point Model results by Martin and Quinn (2002) provide scholars ideology scores of Supreme Court justices ranging from the 1953 term to the present term. In turn, scholars have used these estimated ideal points in several forecasting studies (Ruger et al., 2004), studies on the influence of interest group activities before the Supreme Court (Collins 2008), research which seeks to identify the role of precedents in judicial decision-making (Clark and Lauderdale, 2010; Bartels, 2009). Altogether, Martin and

[7] Lalor et al. (2016) refer to the assumption that responses to different items are independent of each other for a given ability level of the individual and responses as the 'local independence assumption'.

[8] Indeed, on this point, Ho and Quinn (2010), in a discussion of how to use the judicial ideal generated by Martin and Quinn (2002), go as far as to discuss what happens at the extremes of $\boldsymbol{\alpha}$, where it gets too far into the negatives, it provides no useful information, and when it gets close to zero it is an extremely useful case. We would also like to eliminate items with extreme $\boldsymbol{\beta}$s and items that seem to have repeated $\boldsymbol{\alpha}$s and $\boldsymbol{\beta}$s – however this would lead to too few observations to estimate θ.

[9] Notably, prior knowledge of general Supreme Court ideological positioning is required for the model to run consistently, as are constraints on the θs that can be generated. This is consistent with limitations observed by Londregan (1999) and Jackman (2001).

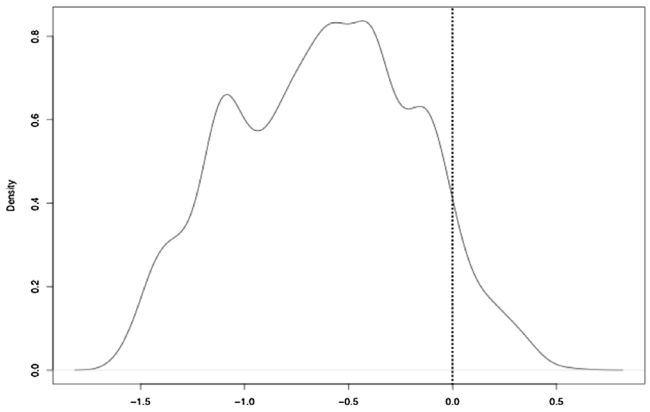

Fig. 4. Distribution of Alpha Parameter Estimated for Supreme Court Cases in Replication of Martin and Quinn (2002).

Quinn (2002) has garnered over 1430 citations from other academic publications.[10]

3.1. Replication of Martin and Quinn (2002) and our proposed solution for IRT models

Thanks to the excellent documentation of replication materials by Martin and Quinn (2002), along with the widely cited Supreme Court Database (SCDB) (Spaeth et al., 2014), we were able to successfully replicate the Dynamic Ideal Point Model and observe the ideal points for Supreme Court justices.

As noted above, the original estimation by Martin and Quinn (2002) generates both positive and negative values for the α parameter, which leads us to think the estimated ideal points for individual justices are likely amiss. This has important implications for scholars who rely on these scores to identify which justice's ideology is best captured in the content of the majority's opinion (Carrubba et al., 2012; Bonneau et al., 2007) and scholars who have claimed the justices' ideologies have shifted over time (Epstein et al., 2007a, 2007b). Fig. 4 presents the distribution of α parameters from the replication of Martin and Quinn (2002) with the value of the α noted on the x-axis. Most cases have a negative α parameter. However, a handful of cases have α parameters greater than 0.5. Altogether, 335 cases (of 4705 cases resolved between the 1946 term through the 2015 term) receive positive α parameters ($\sim$ 7%), with the remaining cases receiving negative α parameters.

As indicated above, our proposed solution to the problem of having cases with positive and negative α parameters in an IRT model is actually a rather simple one. If we think that this inconsistency is being driven by something inseparable from the data-generating process itself—if we think that the existence of a meaningful second dimension found in these cases drives this result—then the only thing we can do is prune the data so that we only study cases that are explained by the first dimension. On the other hand, if we believe this to be an anomaly, the best solution is to disallow this result in the model write-up stage and continue by truncating the distribution of the cases.[11] As such, the procedure we propose is rather simple: take an existing IRT approach—in this scenario, the IRT at the heart of the Dynamic Ideal Point Model by Martin and Quinn (2002)—and run it. If one finds that there are items that have α values that violate the distributional assumptions of the IRT model, then rerun the method without those items.

Some may view our post-estimation proposal as introducing more problems than it solves. We acknowledge that by altering the consistency of the α parameter, we are potentially impacting how the model performs. But we also acknowledge that our proposed solution is justifiable as inconsistencies and non-monotonicity in item parameters would lead to greater difficulty in generating precise ideology scores in IRT models. Identifiability is already a problem with Bayesian IRT models (see Martin and Quinn, 2002 and Clinton et al., 2004 for approaches for making models more identifiable); this is simply an additional concern that can lead to unstable results. It is important to note that in nine of the seventy terms (1946–2015) our re-estimation procedure identifies a different median justice than the original estimation by Martin and Quinn (2002).[12] Given that many scholars have relied on the median justice as a proxy for the preferences of the Supreme Court (e.g., Martin et al., 2005 along with Hall, 2014), we find this variation in the identity of the median highly important.

Fig. 5 compares the estimated ideology scores for Supreme Court justices via our replication of Martin and Quinn (2002) (in dotted blue) and the estimated ideology scores for the justices with our proposed post-estimation pruning approach (in solid red). In general, our approach estimates the justices to have more moderate ideological estimates than the primary results generated by Martin and Quinn (2002). The ideology scores estimated by Martin and Quinn (2002) are hereafter referred to as MQ scores.

Fig. 5 demonstrates a relatively consistent pattern that suggests that our post-estimation procedure leads to smaller year-to-year ideological changes for the justices. Overall, Fig. 5 also indicates the relative ideological spread of the bench is smaller under our proposed post-estimation procedure. These patterns are significant in light of the notable research on ideological drift of justices and polarization of the Supreme Court which relied on MQ scores reflexively.[13]

We demonstrate in the remainder of this section that the cases which have a positive α parameter, a reversed discrimination parameter, are distinct from cases that have negative α parameter in the original estimation of MQ scores.[14] Therefore, we have reasons to suspect there are systematic and meaningful characteristics of cases that have led to some cases having positive α parameters and most having negative α parameters. We rely on the Supreme Court Database (SCDB) (Spaeth et al., 2014) to highlight characteristics of the cases and observe how the distribution of α parameters varies across the type of legal issue (*issueArea* in SCDB), the natural court (*naturalCourt*), the size of the majority coalition (*majVotes*), and the ideological direction of the Supreme Court's decision (*decisionDirection*). We also implement the binary *saliency* measure by Epstein and Segal (2000) to observe if a case's saliency (or lack of) is related to the direction of its discrimination parameter (α). We show that there are important landmark cases that received a positive discrimination parameter (α) in the original estimation of Martin and Quinn (2002) and other important cases which received a negative discrimination parameter (α).

[10] Reported by GoogleScholar on August 15, 2021.

[11] Indeed, with regards to our proposed solution, Ho and Quinn (2010), in a discussion of how to use the estimated ideal points of Martin and Quinn (2002), imply a case will likely provide no useful information when α has an extremely negative value. This is similar to our point that cases (items) where α and β are identical (or nearly so) provide no additional information, but serve to merely make us believe the standard errors are smaller than we think.

[12] This aspect of our re-estimation of MQ scores is highlighted and discussed at length in the following section concerning applications of MQ scores in judicial politics research.

[13] In light of the re-estimated MQ scores, this literature seems primed for a re-assessment (in particular Martin et al., 2005 along with Westerland et al., 2010).

[14] We also suggest in the conclusion section that the presence of positive and negative α parameters *might* be driven by the presence of an additional latent dimension If this is true, and we care about the meaningfulness of the unidimensional model, forcing these cases to look like non-deviant cases only hides this problem, it does not solve it.

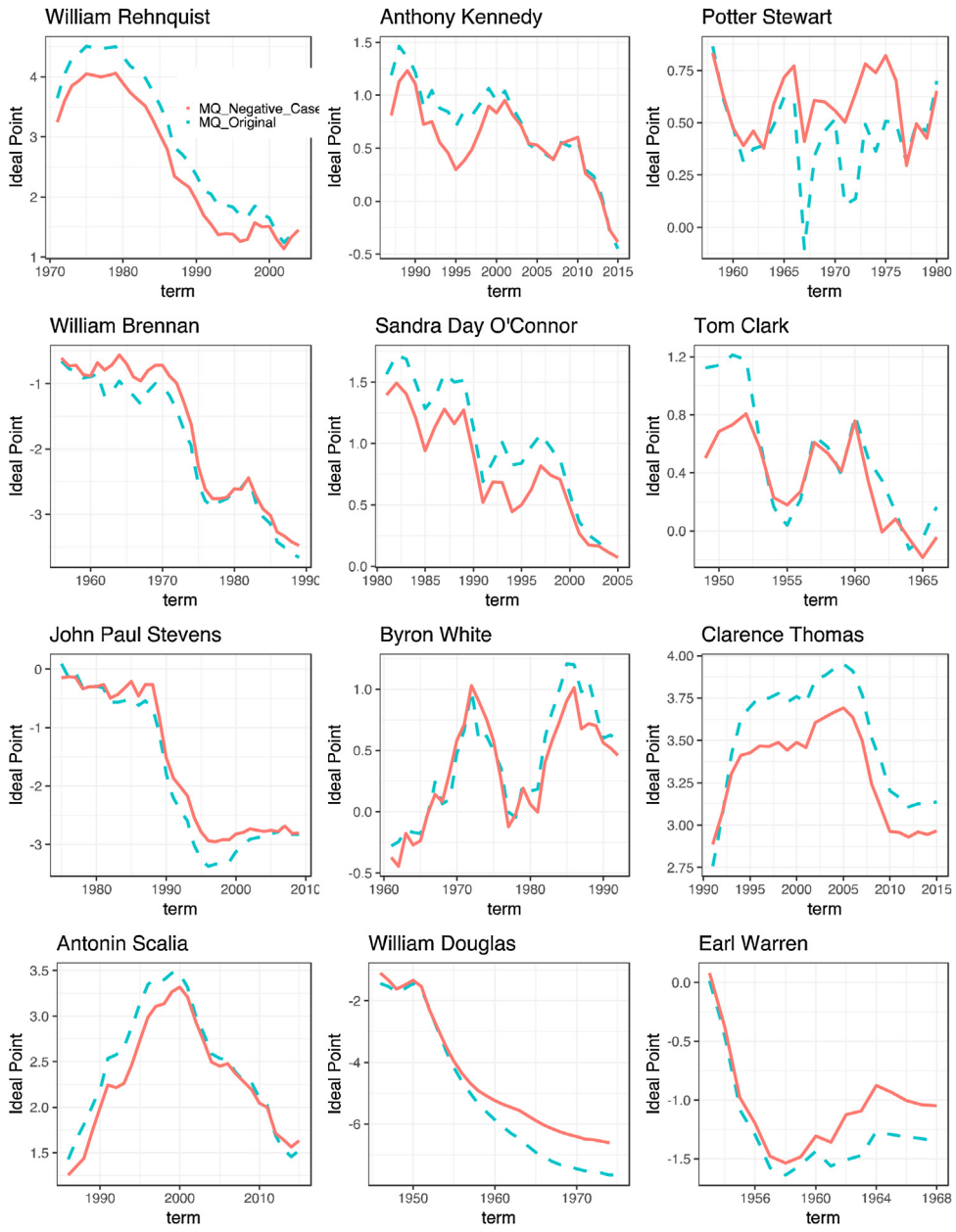

Fig. 5. Comparison of Ideal Points Over Time for Select Supreme Court Justices.
Note: the solid line represents our re-estimated MQ scores and the dotted line represents the original MQ scores.

Fig. 6 displays the distribution of cases with negative and positive α parameters across issue areas of law. Almost every area of law has some cases with reverse discrimination, a positive α. The results in Fig. 6 imply that cases with a positive α parameter are often observed in criminal procedure cases, civil rights, and First Amendment cases.

Next, Fig. 7 suggests there are temporal trends in the frequency of positive α parameters. Cases with reversed discrimination parameters frequently occurred towards the end of the Vinson Court and at the beginning of the Burger Court. In the last half of the Rehnquist Court, where no membership changes occurred, and in the Roberts Court, where few membership changes have occurred, we observe little-to-none of the resolved cases receive positive α parameters. This distribution is insightful for many reasons. The patterns may indicate that the justices' ideology is becoming more impactful on the Supreme Court over time and the decision-making processes of the justices may similarly rely more often on a singular dimension at these times.[15]

Fig. 8 also provides a great deal of insight into which cases receive a positive α parameter in the replication of Martin and Quinn (2002). When we distribute and identify the type of α parameter, either positive or negative, across the size of the majority coalition behind each case, we find that the narrower the majority coalition, the more likely the Dynamic Ideal Point Model pro-

[15] A suggestion has been made to run the model with only the positive α valued items instead of just the negative α items as a robustness check. The idea is in principal a good one, though in practice there is a severe limitation: there are not enough positive α items to have any robust estimation of these ideal points. Since the dynamic ideal point model requires enough votes per term to converge, the absence of any positive α items in 19 out of 47 Supreme Court Terms (and less than 10 in an additional 17 terms) makes estimating any similarly dynamic model impossible.

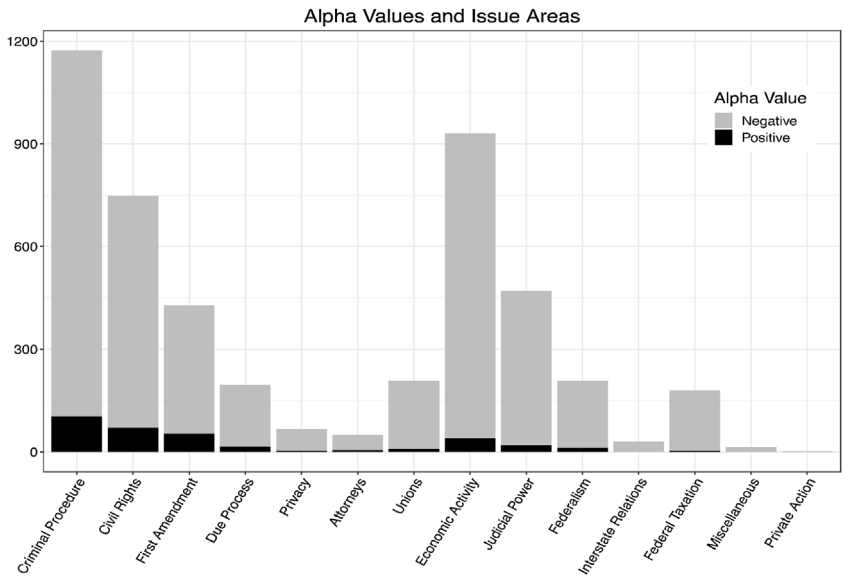

Fig. 6. Distribution of Discrimination (α) Parameter Across Issue Areas.

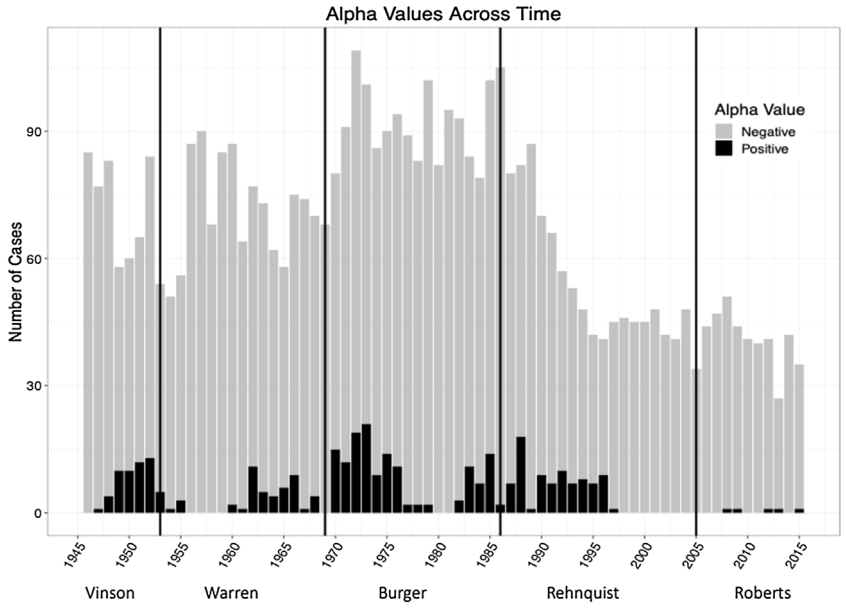

Fig. 7. Distribution of Discrimination Parameter (α) Across Natural Courts.

duces a positive α parameter. Specifically, 315 of the 1606 cases resolved by a five-to-four majority have a positive α parameter. Furthermore, none of the cases resolved through the formation of seven-to-two or eight-to-one majorities received a positive α parameter.

Finally, Table 1 demonstrates that the presence of a positive α parameter is not related to the ideological direction of the given case's outcome. In Table 1, we observe cases whose outcomes are in a conservative ideological direction and whose outcomes are in a liberal ideological direction receiving a positive α parameter in the original estimation of the Dynamic Ideal Point Model by Martin and Quinn (2002). The ideological direction of each case was provided by the SCDB (Spaeth et al., 2014). We find a pattern consistent with a weak relationship to case ideology, but not clear enough; negative α cases represent a mix of both liberal and conservative opinions, while positive α are more conservative than liberal, though not overwhelmingly.

In sum, our post-estimation procedure for correcting IRT models demonstrates how this post-estimation procedure generates more moderate and more stable ideology scores for Supreme Court justices as compared to the original MQ scores. Moreover, we also offered a considerable amount of evidence that the cases resolved by the Supreme Court that have a positive (reversed) α parameter are distinct from the cases that have negative α parameter.

Another alternative we considered would be to focus on the positive α cases and re-estimate ideal points on just those cases as well, essentially pairing this measure with the negative only cases measure discussed above. There are a few problems with this strategy: first, given that we want to maintain continuity with Martin and Quinn's approach as much as possible, running their dynamic IRT with the same priors they did would simply not work with the number of observations that are left if we subset to only the positive α. This is especially true when we take the dynamic part of the dynamic IRT seriously: as we see in Fig. 7, there would be sev-

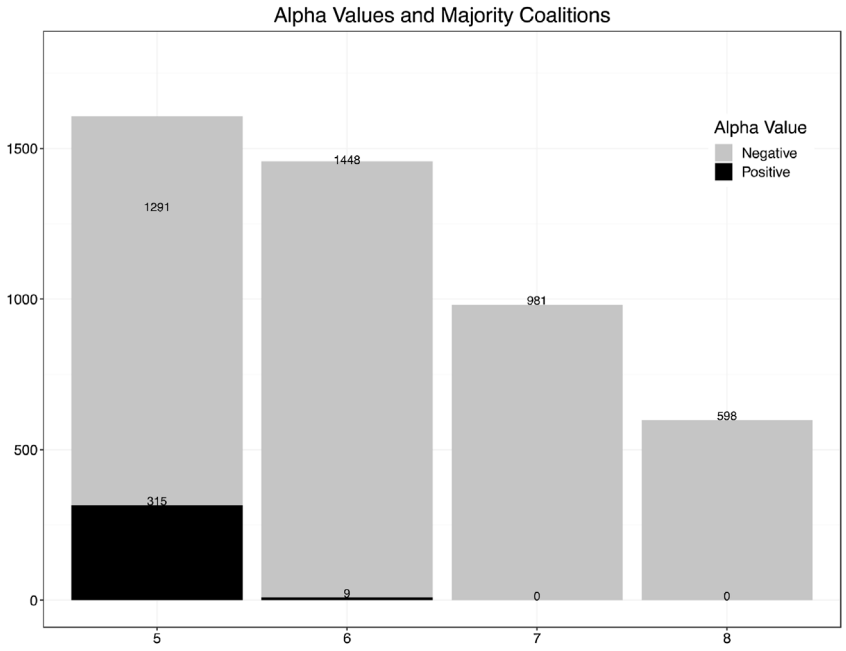

Fig. 8. Distribution of Discrimination Parameter (α) Across Majority Coalition Sizes.

Table 1
Discrimination Parameter (α) Values Across the Ideological Direction of Case Decisions.

		Number of Cases with Positive α Parameter	Number of Cases with Negative α Parameter
Ideological Direction	**Conservative**	273	2212
	Liberal	62	2114
	Unspecified	0	47

eral different terms in which there would be no cases preserved at all. Second, it seems that, even using a different IRT model altogether, we would have sample size problems that might make the new underlying model unstable. Beyond that, the positive α cases could be reverse-ordered and then reincluded in the model, consistent with some discussion of the psychometrics literature. But, this practice assumes that the positive α cases would be the same as the negative α cases save for the direction of the α, and we have some doubt as to the veracity, as we showed in the preceding section, there does seem to be distinct antecedents and predictions of positive α cases that make them distinct. A potential extension (for an entirely separate paper) would be to explore a multidimensional IRT approach that might solve this problem.[16]

4. Applications to judicial politics research

Suppose the estimates of any particular justice's ideology are incorrect. In this scenario, the results of empirical studies within judicial politics (particularly studies that use these scores as cardinal-level unidimensional independent variables in their models) might be misestimated. We first highlight differences between the median justices identified by the original MQ scores and our recoded MQ Scores. Next, we use our re-estimated MQ scores to replicate a well-known paper that relies on MQ scores, Carrubba et al. (2012). Here, we first replicate the author's primary models and we then replace the original MQ scores with our recoded MQ scores. We then follow up with an estimation approach that takes into account the uncertainty generated by the scores themselves. This process allows us to illustrate how sensitive important results in judicial politics are to measures of ideology.

4.1. *Martin et al. (2005)* and new medians on the court

The following list identifies the terms in which the median justice identified in our procedure differed from the original Martin and Quinn (2002) estimation: 1950 (we identify Justice Tom C. Clark as the median where the original estimation identifies Justice Harold Burton); 1951 (we identify Justice Tom C. Clark as the median where the original estimation identifies Justice Harold Burton); 1956 (we identify Justice Sherman Minton as the median where the original estimation identifies Justice Felix Frankfurter); 1968 (we identify Justice Thurgood Marshall as the median where the original estimation identifies Justice William J. Brennan); 1970 (we identify Justice Potter Stewart as the median where the original estimation identifies Justice Byron White); 1975 (we identify Justice Potter Stewart as the median where the original estimation identifies Justice Byron White); 1976 (we identify Justice Harry Blackmun as the median where the original estimation identifies Justice Potter Stewart); 1991 (we identify Justice David Souter as the median where the original estimation identifies Justice Sandra Day O'Connor); 1994 (we identify Justice Anthony Kennedy as the median where the original estimation identifies Justice Sandra Day O'Connor). Fig. 9 illustrates how the location of the median member shifts between our scoring method and the original MQ scores from the Dynamic Ideal Point Model (based on a figure from Martin et al., 2005).

[16] Though this approach would require dramatically altering Martin and Quinn (2002) methodology. Thus, moving the instant project far afield from the replication exercise we sought to bring to light here. Potentially a future extension of this would be fruitful, though we are also skeptical that it will be a singular second dimension that explains this. Rather, given the coalitions we saw amongst these cases, we expect the dimensions that are being tapped into in these cases to be far more complex than just one additional dimension.

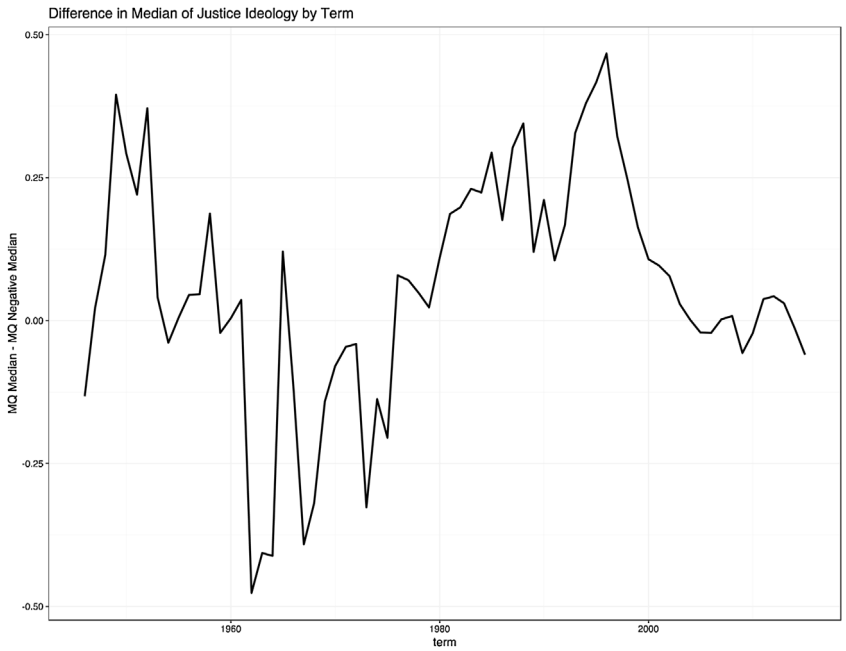

Fig. 9. Differences in the Median Justice's Ideal Point (1946-2015).

4.2. Replication of Carrubba et al. (2012)

Carrubba et al. (2012) ask a highly relevant but highly disputed question in judicial politics: who controls the content of opinions by the Supreme Court? Within the paper, the authors identify two existing theories. The first theory argues that the preferences of the median justice should guide the content of the Supreme Court's opinion. This theory is known as the Median Justice Theorem (MJT Model) and is built on the median voter theorem. The second theory argues that the justice authoring the Supreme Court's opinion should have the most influence on the content of the opinion. This theory is known as the Opinion Writer Model (OW Model). In contrast to these existing theories, Carrubba et al. (2012) argue the median justice of the majority coalition in each case should exercise the most influence over the content of the Supreme Court's opinion, which they refer to as the Coalition Median Model (CM Model). Thus, the main contribution to judicial politics offered through this paper is a direct comparison of the three most contested theories on who controls the content, and therefore the law, in written opinions by the Supreme Court.

Methodologically, Carrubba et al. (2012) apply the original MQ scores to generate a set of 'treatments' which identify the ideological distance among justices. Their first measure is the *Distance to Court Median* variable, which estimates the distance between a given justice and the median justice on the bench. The next measure is the *Distance to Opinion Writer* variable, which estimates the distance between a given justice and the justice who is writing the majority's opinion. The final distance measure is the *Distance to Coalition Median* variable, which estimates the distance between a given justice and the median member of the majority coalition. The dependent variable relied upon by Carrubba et al. (2012) is binary and identifies whether a given justice authored a concurring opinion or not. Table 2 displays the replication results of Carrubba et al. (2012). The models estimated with our re-estimated MQ scores are also placed in Table 2 and noted as the *updated* results and the columns have a grey background.

We were able to exactly replicate the results by Carrubba et al. (2012) while using the original MQ scores.[17] Our replication using the re-estimated MQ scores are very similar to their original estimated results. However, the updated results for the CM Model are significantly smaller than the original results for the CM Model. The difference in the coefficients between the replicated and updated models suggests the use of the original MQ scores may have overstated the size of the effect. We also decided to run a set of regressions that take into account the measurement uncertainty of any IRT model. Columns 3, 6, and 9 all use an MCMC approach to account for different plausible measures of each ideal point, and what we report is the average effect and the average error terms. This is sometimes referred to as the "Method of Composition" discussed more heavily in Trier and Jackman (2008). Unsurprisingly, the main effect estimates are slightly different and the error terms are slightly larger. Even though the effects in all models are still positive and significant, the more precise measurement of the effect matters, and getting a better approximation of the effect of measurement error in this particular model is of particular note, since the only measures included are based on measures of ideal points. Although this is not a rejection of Carrubba et al.'s original finding—nor would we necessarily have expected this to be—we believe that this refinement is important.

Altogether, the replication of Carrubba et al. (2012) and comparison of median justices identified under our recoded MQ scores and original MQ scores, stressed how crucial it is to measure our central constructs carefully. Further, this section revealed that a misspecification of judicial ideology could bias outcomes when modeling judicial behavior.

[17] We also used a rare events logistic regression, since the choice to author a special concurrence only occurred 7.5% of the time. The results were unchanged from what is reported above.

Table 2
Replication and Extension of Carrubba et al. (2012).

	Coalition Median Model			MJT Model			Opinion Writer Model		
DV: Author Special Concurrence	(1) Original	(2) Updated scores	(3) Updated Scores & Uncertainty	(4) Original	(5) Updated scores	(6) Updated Scores & Uncertainty	(7) Original	(8) Updated scores	(9) Updated Scores & Uncertainty
Distance to Coalition Median+	0.324* (0.017)	0.285* (0.017)	0.258* (0.021)						
Distance to Court Median				0.145* (0.018)	0.143* (0.018)	0.130* (0.023)			
Distance to Opinion Writer							0.200* (0.014)	0.188* (0.014)	0.177* (0.017)
N	17,422	17,422	17,422	17,422	17,422	17,422	17,422	17,422	17,422

+ Median (Majority Case) or Opinion Writer (Plurality Case). The binary dependent variable indicates if a specific justice authored a special concurrence. Each model was estimated with a Logistic regression. The estimates are maximum likelihood, with asymptotic standard errors in parentheses. * indicates significance at $p < 0.05$. Columns 3,6, and 9 use the uncertainty in measuring the ideal points in the model itself. Estimates are based on 1000 MCMC runs of the model, following the suggestions of Treier and Jackman (2008).

5. Discussion and conclusion

As mentioned above, IRT models are a widely used methodology to measure the latent traits in political science and legal scholarship. These models were borrowed from the education testing and psychometrics literature. We believe that scholars have inappropriately ignored assumptions of IRT models in the process of applying them to judicial politics. Specifically, we focused on the violation of the uniformity of direction in the discrimination parameter (α). We cannot compare across items if item parameters α have different directions (Jackman, 2001; van der Linden et al., 2013). More explicit to applications of IRT models in judicial politics, if one case resolved by the Supreme Court is estimated to have a positive α and another case is estimated to have a negative α, then we cannot compare the two cases and understand how they provide insight into the ideology of the justices. Our solution to solve this specific application dilemma is to exclude the items that received positive α values in the initial estimation and then re-estimate the IRT model on the remaining cases.

In this paper we chose to solve only one application dilemma that generally occurs when IRT models are implemented on data from the Supreme Court. However, we would like to note the existence of other application dilemmas. Mainly these violations are related to the local independence of the estimation. As cases are resolved on the Supreme Court through majority coalitions, one justice's vote is not independent of the vote of another justice. Likewise, if we think courts and justices are constrained by precedent and follow the norm of *stare decisis,* then the cases resolved by the Supreme Court which serve as items in an IRT model are not independent of each other. Baker and Kim (2004), along with Wainer and Kiely (1987), are quite clear that when local independence is violated, the estimated latent scale will appear more precise (with standard errors biased downward) and the item parameters (difficulty and discrimination) will be larger than they should be.

Further, if the same coalition of Supreme Court justices, say five conservative justices, consistently find themselves in the majority coalition and four liberal justices consistently find themselves not in the majority coalition — then the cases which serve as items in an IRT model will be correlated and the scale will lack construct validity. Baker and Kim (2004, 2017) point out that in traditional educational testing we would not observe the same coalitions of individuals appearing across items and that we should delete the items that are highly correlated with other items in the model in order to prevent the correlated items from dominating the scale. These two application dilemmas of IRT models in judicial politics are beyond the scope of this paper. That said, future scholars may be interested in developing methodological techniques to overcome these issues.

It is also important to mention that the presence of both negative and positive α values for items in IRT models may indicate the presence of at least a second latent dimension. Fischman and Jacobi (2015) argue that there is likely a legally-focused dimension beyond a latent ideological dimension. Fischman and Jacobi (2015) also present evidence of a continuum between legalism and pragmatism that similarly divides the justices in ways that are hard to explain strictly in terms of the singular ideological dimension. This would put the ideology of judicial actors in line with the work of Poole and Rosenthal (1991, 2000), who estimate the ideology of members of Congress and notably find that two dimensions are required to explain Congressional voting behavior adequately.[18] We have no reason to believe that such a set of analyses done on Supreme Court justices would not lead to a similar conclusion, but that is beyond our purview here.

To conclude, we would like to highlight how this paper reflects one element of our greater research agenda, which explores the many limitations with the current ways we estimate ideology and ideal points in judicial politics. For example, future research will examine how concurrences are accounted for in the literature and argue that the current approaches misclassify them. Similarly, future research will explore the multidimensionality of ideology (in particular building off of the inconsistencies brought up in this paper by the deviations in the alpha parameters), as well as estimating measures of ideology that are based solely on non-voting-based approaches, in particular raw text and citations found within legal opinions. Overall, it may be that the future of modeling judicial ideology will be to move away from traditional IRT-based models and into something designed more explicitly for modeling law and the legal reasoning found within opinions by the Supreme Court. This paper, in that context, should be viewed as shedding a light on a critical limitation of traditional methods and providing an easy correction that improves the modeling of judicial ideology.

[18] Likewise, Aldrich et al. (2014) demosntrate that even the high levels of predictive power of the first dimension of DW-NOMINATE are not sufficient to prove the presence of only one dimension in Congressional ideology, and that it is likely that multiple dimensions are at play at a single time.

Author statement

Joshua Lerner: Conceptualization, Methodology, Validation, Formal Analysis, Writing-Original Draft, Writing - Review & Editing.

Mathew McCubbins: Conceptualization, Methodology, Supervision, Writing-Original Draft.

Kristen Renberg: Conceptualization, Validation, Formal Analysis, Writing-Original Draft, Writing - Review & Editing, Visualization.

References

Aldrich, John H., Montgomery, Jacob M., Sparks, David B., 2014. Polarization and ideology: Partisan sources of low dimensionality in scaled roll call analyses. Political Anal. 22 (4), 435–456.

Baker, Frank B., 2001. The Basics of Item Response Theory. For Full Text. http://ericae.net/irt/baker.

Baker, Frank, Kim, Seock-Ho, 2004. Item Response Theory: Parameter Estimation Techniques. CRC Press.

Baker, Frank B., Kim, Seock-Ho, 2017. The Basics of Item Response Theory Using R. Springer, New York, NY.

Bartels, Brandon L., 2009. The constraining capacity of legal doctrine on the US Supreme Court. Am. Polit. Sci. Rev. 103 (3), 474–495.

Bonica, Adam, 2014. Mapping the ideological marketplace. Am. J. Pol. Sci. 58 (2), 367–386.

Bonneau, Chris W., Hammond, Thomas H., Maltzman, Forrest, Wahlbeck, Paul J., 2007. Agenda control, the median justice, and the majority opinion on the US Supreme Court. Am. J. Pol. Sci. 51 (4), 890–905.

Carrubba, Cliff, Friedman, Barry, Martin, Andrew D., Vanberg, Georg, 2012. Who controls the content of Supreme Court opinions? Am. J. Pol. Sci. 56 (2), 400–412.

Clark, Tom S., Lauderdale, Benjamin, 2010. Locating Supreme Court opinions in doctrine space. Am. J. Pol. Sci. 54 (4), 871–890.

Clinton, Joshua D., Lapinski, John S., 2006. Measuring legislative accomplishment, 1877–1994. Am. J. Pol. Sci. 50 (1), 232–249.

Clinton, Joshua, Jackman, Simon, Rivers, Douglas, 2004. The statistical analysis of roll call data. Am. Polit. Sci. Rev. 98 (2), 355–370.

De Boeck, Paul, Wilson, Mark, 2004. A framework for item response models. In: Explanatory Item Response Models. Springer, New York, NY, pp. 3–41.

Epstein, Lee, Segal, Jeffrey A., 2000. Measuring issue salience. Am. J. Pol. Sci., 66–83.

Epstein, Lee, Martin, Andrew D., Segal, Jeffrey A., Westerland, Chad, 2007a. The judicial common space. J. Law Econ. Organ. 23 (2), 303–325.

Epstein, Lee, Martin, Andrew D., Quinn, Kevin M., Segal, Jeffrey A., 2007b. Ideological drift among supreme court justices: Who, when, and how important. Nw. UL Rev. 101, 1483.

Fischman, Joshua B., Jacobi, Tonja, 2015. The second dimension of the Supreme Court. Wm. & Mary L. Rev. 57, 1671.

Fox, Jean-Paul, 2010. Bayesian Item Response Modeling: Theory and Applications. Springer Science & Business Media.

Hall, Matthew E.K., 2014. The semiconstrained court: public opinion, the separation of powers, and the US Supreme Court's fear of nonimplementation. Am. J. Pol. Sci. 58 (2), 352–366.

Ho, Daniel E., Quinn, Kevin M., 2010. How not to lie with judicial votes: misconceptions, measurement, and models. Calif. Law Rev. 98 (3), 813–876.

Imai, Kosuke, Lo, James, Olmsted, Jonathan, 2016. Fast estimation of ideal points with massive data. Am. Polit. Sci. Rev. 110 (4), 631–656.

Jackman, Simon, 2001. Multidimensional analysis of roll call data via Bayesian simulation: identification, estimation, inference, and model checking. Political Anal. 9 (3), 227–241.

Lakatos, Imre, 1976. Falsification and the methodology of scientific research programmes. In: Can Theories Be Refuted? Springer, Dordrecht, pp. 205–259.

Lalor, John P., Wu, Hao, Yu, Hong, 2016. Building an evaluation scale using item response theory. NIH Public Access In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, Vol. 2016, p. 648.

Lerner, Joshua Y., McCubbins, Mathew D., Renberg, Kristen M., 2019. Diving the question: concurrences on the supreme court and ideal Point estimation. Working Paper. Presented at the Southern Political Science Association Annual Meeting.

Linzer, Drew, Staton, Jeffrey K., 2012. A Measurement Model for Synthetizing Multiple Comparative Indicators. Working Paper. Emory University, Atlanta, GA.

Londregan, John, 1999. Estimating legislators' preferred points. Political Anal. 8 (1), 35–56.

Martin, Andrew D., Quinn, Kevin M., 2002. Dynamic ideal point estimation via Markov chain Monte Carlo for the US Supreme Court, 1953–1999. Political Anal. 10 (2), 134–153.

Martin, Andrew D., Quinn, Kevin M., Epstein, Lee, 2005. The median justice on the United States Supreme Court. NCL rev. 83, 1275.

Poole, Keith T., 2005. Spatial Models of Parliamentary Voting. Cambridge University Press.

Poole, Keith T., Rosenthal, Howard, 1991. Patterns of congressional voting. Am. J. Pol. Sci., 228–278.

Poole, Keith, Rosenthal, Howard, 2000. Congress: a Political-Economic History of Roll Call Voting. Oxford University Press on Demand.

Ruger, Theodore W., Kim, Pauline T., Martin, Andrew D., Quinn, Kevin M., 2004. The supreme court forecasting project: legal and political science approaches to predicting Supreme Court decision making. Columbia Law Rev., 1150–1210.

Spaeth, Harold, Epstein, Lee, Ruger, Ted, Whittington, Keith, Segal, Jeffrey, Martin, Andrew D., 2014. Supreme Court Database Code Book.

Tahk, Alexander, 2018. Nonparametric ideal-point estimation and inference. Political Anal. 26 (2), 131–146.

Tausanovitch, Chris, Warshaw, Christopher, 2014. Representation in municipal government. Am. Polit. Sci. Rev. 108 (3), 605–641.

Treier, Shawn, Jackman, Simon, 2008. Democracy as a latent variable. Am. J. Pol. Sci. 52 (1), 201–217.

Treier, Shawn, Sunshine Hillygus, D., 2009. The nature of political ideology in the contemporary electorate. Public Opin. Q. 73 (4), 679–703.

van der Linden, Wim J., Hambleton, Ronald K., 2013. Handbook of Modern Item Response Theory. Springer Science & Business Media.

Wainer, Howard, Kiely, Gerard L., 1987. Item clusters and computerized adaptive testing: a case for testlets. J. Educ. Meas. 24 (3), 185–201.

Westerland, Chad, Segal, Jeffrey A., Epstein, Lee, Cameron, Charles M., Comparato, Scott, 2010. Strategic defiance and compliance in the US courts of appeals. Am. J. Pol. Sci. 54 (4), 891–905.