UNIVERSITY OF CALIFORNIA SAN DIEGO

Three Investigations into Bayesian Measurement Modeling in Political Science

A dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Political Science

by

Bertrand L. Wilden

Committee in charge:

Professor James H. Fowler, Chair
Professor David Fortunato
Professor LaGina Gause
Professor Margaret E. Roberts

2024

The dissertation of Bertrand L. Wilden is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

EPIGRAPH

"It may be that Bayesian inference is the best general purpose method of inference known. However, Bayesian inference is much less powerful than we'd like it to be. There is no approach to inference that provides universal guarantees. No branch of applied mathematics has unfettered access to reality, because math is not discovered, like the proton. Instead it is invented, like the shovel."

Richard McElreath

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my advisor Professor James Fowler for providing exceptional support and guidance throughout each stage of the dissertation. You had a special way of instilling in me renewed optimism for my research after every one of our meetings. Thank you to my committee members, Professor Molly Roberts, Professor LaGina Gause, and Professor David Fortunato for challenging my ideas and helping me improve as a scholar. And thank you to Professor Philip Roeder for advising me on my undergraduate thesis and inspiring me to begin this academic endeavor.

I would also like to acknowledge my classmates, Austin Beacham, Alison Boehmer, Yeilim Cheong, Gabriel De Roche, Nhat-Dang Do, Kevin Flannagan, Stephanie Peng, Rachel Skillman, Malika Talgatova, Laura Uribe, and Sam Williams. I am grateful I was able to witness you grow and flourish during these past five years. I cannot wait to see what each of you accomplish in the next five. Your friendship means everything to me, and I would have never made it this far without it.

My deepest gratitude is reserved for my mother. Thank you for always believing in me.

VITA

2018 Bachelor of Arts, University of California San Diego
2021 Master of Arts, University of California San Diego
2024 Doctor of Philosophy, University of California San Diego

FIELDS OF STUDY

Major Field: Political Science

Professor James H. Fowler

ABSTRACT OF THE DISSERTATION

Three Investigations into Bayesian Measurement Modeling in Political Science

By

Bertrand L. Wilden

Doctor of Philosophy in Political Science

University of California San Diego, 2024

Professor James H. Fowler, Chair

Measurement is foundational to political science research. Theories are only testable to the extent that their abstract concepts can be connected to empirical reality. Political science is a field where many important questions deal with concepts whose measurement is not immediately obvious. Does democracy reduce corruption in a country? Does ideological extremism impact the electoral success of politicians? To answer these questions, valid measurement of the key variables is an essential first step.

In this dissertation I propose three projects which help improve measurement in political science. The first is a new method for measuring race/ethnicity when this data is missing. My model uses Bayes' theorem to predict the posterior probability that an individual identifies with a particular race or ethnicity, given other known attributes. I validate these predictions against voter registration data, and I show that my model is far more accurate compared to previous methods. I also develop an R package to provide easy implementation of my method.

The second project is a new model for measuring the political ideology of actors under extreme missing data conditions. Models of political ideology usually use observed actions, such as taking positions on legislation, to infer an actor's latent ideology. But I show that in contexts where most actors fail to take explicit positions on most pieces of legislation, the measurements from traditional models can quickly degrade. The model I develop directly accounts for these missing signals, thereby generating more accurate measurements of political ideology. I apply the model to data on federal interest group lobbying.

The final project is a method for incorporating measurement model uncertainty into an empirical theory testing model. Estimates of ideology from the model in project #2, as well as from any other statistical measurement model, produce more than a single value of the latent variable—they also produce some measure of the error/uncertainty for the true value. I show that failing to account for this measurement error in the theory testing stage can lead to misleading or biased conclusions. The method I propose in this project fixes this source of bias.

Chapter 1

Improved Bayesian Ethnorace Prediction

1.1 Introduction

Quantitative research on race and ethnic politics in the United States can be constrained by the lack of available individual-level data with these variables. Without individual-level data, researchers are sometimes forced to used aggregate racial data at census unit levels to measure disparate effects of policies such as voter ID laws on voter turnout (Kuk, Hajnal, and Lajevardi 2020) or of incarceration on political participation (Burch 2013). To overcome this challenge, I develop a method for imputing individual-level race and ethnicity values directly. This method uses Bayes' rule to make predictions by combining information from nationwide distributions of six ethnorace categories over other characteristics, such as names, geographies, political party identification, and others. Intuitively, an individual with a name that is highly unique to a particular race or ethnicity, and who lives in an area with many other people of the same ethnorace, will be given a high posterior probability of belonging to that group.

My implementation builds off existing methods which use a similar prediction algorithm (Elliott et al. 2009; Imai and Khanna 2016; Voicu 2018; Clark, Curiel, and Steelman 2021) The method described by Imai and Khanna (2016) in particular, and the associated R package **wru**, has become popular in recent studies on race and ethnicity. Its application has been used widely to help advance research on racial protests and voting patterns (Enos, Kaufman, and Sands 2019); disparities in campaign financing (Grumbach, Sahn, and Staszak 2020; Grumbach and Sahn 2020), evictions (Hepburn, Louis, and Desmond 2020) and voter turnout (Fraga 2018); the impact of electoral institutions on local representation (Abott and Magazinnik 2020); and public health issues such as suicide rates (Studdert et al. 2020).

My ethnorace prediction method improves upon Imai and Khanna (2016) in several ways. Whereas their method only takes as inputs distributions over surnames, geolocation, political party, age, and gender, my implementation adds information from a nationwide list of first names (Tzioumis 2018) as well as from whether an individual is living in a single-family home. Additionally, I incorporate insights from the machine learning literature on classification algorithms to further improve predictive performance. The result of these modifications is a substantial increase in predictive power compared to Imai and Khanna (2016) in validation tests. The predictive gains are most striking for non-White groups. My method is available in an easy-to-use R package **bper**.[1]

In the next section I provide some background and specifics regarding the inputs and outputs of my ethnorace prediction method. Then I explain the methodology and compare my implementation with previous methods. Next, I demonstrate the predictive performance of my method when validated against the combined North Carolina and Florida voter file (n = 23,754,749). The inclusion of self-reported ethnorace identifiers in these data allows me to compare my predictions against a ground truth. In order to better inform researchers whose target populations may be

[1] **bper**: Bayesian Prediction for Ethnicity and Race. https://github.com/bwilden/bper

different than those represented by the North Carolina and Florida voter files, I also run tests on sub-sets of demographics characteristics of these states (urban/rural, income level, percentage foreign born, and education level). And lastly, I replicate Grumbach and Sahn's (2020) findings in their study on racial disparities in campaign finance. I compare the empirical results using my prediction method against those from Imai and Khanna (2016). Grumbach and Sahn's original findings, that non-White groups are under-represented among political donors and that the presence of co-ethnic candidates acts to counteract this pattern, are strengthened when using the improved ethnorace predictions from my method.

1.2 Background

1.2.1 Outputs

What does it mean to "predict race and ethnicity"? These are categories which, although relatively immutable compared to other identities, do not have universally accepted delineations and meanings (Omi and Winant 2014; Sen and Wasow 2016; Davenport 2020). I follow the convention from previous ethnorace prediction methods by using the US Census Bureau categorizations (Elliott et al. 2009; Imai and Khanna 2016; Voicu 2018; Clark, Curiel, and Steelman 2021). In this framework, individuals can be classified as non-Hispanic White, non-Hispanic Black or African American, non-Hispanic Asian and Pacific Islander, non-Hispanic American Indian and Alaska Native, Hispanic or Latino alone, and non-Hispanic Other Race.[2] Because Hispanic identity is defined by the Census, and understood commonly, as an ethnicity, rather than as a race, I use the term "ethnorace" in this paper to refer to any of these six groups eligible for prediction.

Relying on Census categorizations comes with drawbacks. One downside is that it obscures

[2]Non-Hispanic Other Race includes individuals who identify as belonging to two or more races or ethnicities, as well as those who may not identify with the other Census categories.

substantial heterogeneity that may exist within each group. Among Asian and Hispanic Americans, for example, there is considerable variation in terms of national ancestry. A growing body of literature shows that neglecting to account for differences within these groups obscures important variation in the political behavior of these individuals (Sanchez 2006). Furthermore, the unfortunate necessity of a catch-all "Other Race" category, which also includes those who identify as multi-racial, ensures that other important sources of diversity are lost.

In addition to the problem of lumping a diverse set of individuals into only a few broad categories, ethnorace prediction methods like the one described here risk reifing an existing racial hierarchy. Ethnoracial categorizations are ultimately socially constructed and exist outside of Census definitions. Researchers should avoid treating ethnorace predictions as something that maps onto essential characteristics for individuals. Rather, these predictions should be understood as approximating ethnoracial self-identification within the constrains of the Census classification system.

There are, however, benefits to using Census ethnorace categorizations. These definitions roughly capture a common understanding of race and ethnicity in the US, and correspond to groups studied frequently in social science research. The data sources of these groups' distributions that serve as inputs to the prediction formula also rely on the Census categorization. There is currently no feasible alternative data source that offers the same level of detail as the Census.

1.2.2 Inputs

In order to generate predicted probabilities that an individual self-identifies as a particular ethnorace, the method uses their observed characteristics as inputs. Earlier methods such as Elliott et al. (2009) use only surnames and geolocations for inputs, whereas Imai and Khanna (2016) add political party as well as age and gender. In addition to these input characteristics, my method

uses first name data and whether an individual lives in single-family housing or not.

First Names. The first names list I use comes from Tzioumis (2018). It is drawn from mortgage applications between the years of 2007 and 2010, and contains ethnorace counts in each of the six groups across 4,250 first names. Relative to Census data, this list of first names may be unrepresentative of the larger US population. Mortgage applicants are wealthier than the average American, and are more likely to be employed. To the extent that first name distributions differ by ethnorace given these unobserved characteristics, this may be a concern. But the predictive benefits from using first name data likely overcome these worries.

Last Names. For my last names data, I draw from the 2000 and 2010 Census Surnames Lists.[3] These lists come from the decennial Census and contain over 160,000 common US last names (those occurring 100 or more times in the population). Like the first names list, these data include counts of individuals in each of the six ethnorace categories across each last name. The **bper** software packages accesses these data, and all other Census data, using wrapper functions for the R package **censusapi** which calls the Census API directly.[4]

Geolocations. The source of my ethnorace distributions by geolocations also comes from the Census. In decreasing order of average population, these geographies include State, County, Census place, ZIP code, Census tract, and Census block. Predictions tend to improve with more precise levels of geography. With this in mind, my implementation automatically matches each individual to the most fine-grain level of geography available in the input data.

Other Inputs. While ethnorace distributions over names and locations are responsible for most of the heavy lifting regarding the algorithm's predictive performance, other input data can be added. The first of these is political party affiliation. This data come from the American National

[3]https://www.census.gov/topics/population/genealogy/data/2010_surnames.html

[4]The Census API loads Census block data prohibitively slowly so I access these data from an external repository. https://github.com/bwilden/bper_data

Election Study[5] aggregated across decades. The three categories of political party I include are Republican, Democrat, and Other (including Independents and "don't knows").

In addition to political party, age and gender can be added as inputs in the algorithm. Like geolocations data, ethnorace distributions over age and gender are accessed via the Census API. These variables do not contain much predictive power in terms of ethnoracial classification (gender distributions over ethnorace are essentially flat), nevertheless, I find that their inclusion in the algorithm helps slightly in some contexts.

Lastly, **bper** allows researchers to input the residency type (single-unit or multi-unit) into the prediction algorithm. Individuals are matched to these probabilities if their address contains "Apt", "Unit", "#", or other such identifier.

1.2.2.1 Data Structure

The raw data sources I describe above all contain counts of individuals with a particular attribute (i.e. the first name JOHN, or the ZIP Code "92092") per ethnorace category. Taking proportions by cell across a given attribute tells us Pr(*Ethnorace*|*Attribute*), and taking proportions by cell across a given ethnorace group tells us Pr(*Attribute*|*Ethnorace*). These two conditional probabilities form the building blocks of the classification algorithm described later.

If any cell in the input data is empty (i.e. if there are no individuals of a particular ethnorace with some attribute), then the conditional probabilities Pr(*Ethnorace*|*Attribute*) and Pr(*Attribute*|*Ethnorace*) will be zero. As explained further in the Methodology section, if either of those two probabilities for an individual equal zero for a given ethnorace, the algorithm will predict a zero probability that the individual belongs to that ethnorace. This will occur even if some other attributes about that individual predict a high probability of belonging to the ethnorace.

[5]American National Election Studies. 2021. ANES 2020 Time Series Study Full Release [dataset and documentation]. July 19, 2021 version. www.electionstudies.org

For example, an individual could have first and last names that are highly predictive of being Hispanic, but reside in a Census block which had zero Hispanic occupants during a decennial Census. Blocks typically contain only around 400 individuals—so this is a real possibility. For this hypothetical person the input data claims that $\Pr(\textit{Hispanic}|\textit{Block} = x) = 0$, which yields $\Pr(\textit{Hispanic}) = 0$ due to the structure of the prediction algorithm. To resolve this issue, I apply a technique from the machine learning literature known as Laplace smoothing to my input data. This works by adding some constant, or pseudo-count, to number of individuals in every cell in the input data, then calculating the aforementioned conditional probabilities $\Pr(\textit{Ethnorace}|\textit{Attribute})$ and $\Pr(\textit{Attribute}|\textit{Ethnorace})$.[6] Laplace smoothing is commonly used in Naive Bayes algorithms in computer science to help improve predictions, but has not been used before in Bayesian ethnorace prediction methods.

1.3 Methodology

The method I develop here computes posterior probabilities for each of the six ethnorace categories for each individual. These tell us, given some set of first name, last name, geolocation, party ID, age, gender, and address type, what is the probability that an individual identifies as a particular ethnorace. Bayes' rule, Equation 1.1, provides a template for how to answer this sort of conditional probability problem.

$$\Pr(R|X) = \frac{\Pr(X|R)\Pr(R)}{\Pr(X)} \tag{1.1}$$

Where R is one of six possible ethnorace categories (White, Black, Asian, Native American, Hispanic, or Other race), and X is the joint probability of an individual having a particular

[6] Missing counts are only a problem in the first names, last names, and geolocation data.

profile of attributes (first name, last name, geolocation, party ID, age, gender, and address type). Unfortunately, the joint probability for X in Equation 1.1 is intractable due to both data constraints and the astronomically large number of combinations of possible attribute profiles. There simply is not data on ethnorace distributions by unique combinations of surnames, first names, and geolocations.

If however, we assume conditional independence of ethnorace among each attribute in X, we can rewrite Equation 1.1 in terms of less complex conditional probabilities as Equation 1.2:

$$\Pr(R|X) = \frac{\Pr(R|x') \prod_{j=1}^{6} \Pr(x_j|R)}{\sum_{i=1}^{6} \Pr(R_i|x') \prod_{j=1}^{6} \Pr(x_j|R_i)} \tag{1.2}$$

Where x is the vector of individual attributes indexed by j. The particular attribute x' comes from using the chain rule to decompose the joint probability $\Pr(R, X)$. The choice of which attribute to use for x' is atheoretical, but all previous prediction methods have used last names (Elliott et al. 2009; Imai and Khanna 2016; Voicu 2018). During my validation exercises, I found that the choice of x' has potentially large consequences for predictive performance. This is because the expression $\Pr(R|x')$ is typically much greater than any $\Pr(x_j|R)$, and therefore contributes more weight to the final posterior probability.[7] For example, using last names for x' appears to help predictions of Whites—but to the detriment of non-Whites. In light of these trade-offs, my method cycles through first name, last name, and geolocation as the choice of x' and computes $\Pr(R|X)$ for each. These posterior probabilities are then averaged within each ethnoracial category to generate final predicted probabilities that an individual belongs to a particular ethnorace. The result of this smoothing is more balanced predictions across each ethnorace. Equation 1.3 shows the final equation that is calculated by the **bper** package is:

[7]To see why, imagine x = Smith and R = Black. According to the 2010 Census Surnames list, the probability that an individual is Black given they have the last name Smith, $\Pr(R|x)$, is 0.23. But the probability that an individual is named Smith given that they are Black, $\Pr(x|R)$, is only 0.0000029.

$$\Pr(R|X) = \frac{1}{3}\sum_{i=1}^{3}\left\{\frac{\Pr(R|x')\prod_{j=1}^{6}\Pr(x_j|R)}{\sum_{i=1}^{6}\Pr(R_i|x')\prod_{j=1}^{6}\Pr(x_j|R_i)}\right\} \tag{1.3}$$

The conditional independence assumption necessary for transforming Equation 1.1 to Equation 1.2 says that knowing both a particular attribute of an individual, and that individual's ethnorace, should give us no extra knowledge of any other attribute for that individual. Stated formally, $\Pr(x_j|R, X_{-j}) = \Pr(x_j|R)$ for all x_j. This assumption is almost certainly violated in the present context. One example that has been demonstrated empirically is that last name distributions by race vary across regions in the US (Crabtree and Chykina 2018). Violations of the conditional independence assumption are commonplace in most applications of similar Naive Bayes classification algorithms. Nevertheless, these prediction methods perform well in many contexts (Lewis 1998; Domingos and Pazzani 1997; Rish 2001). This is likely because of the decision rule governing the final classifications—the posterior probabilities of the true class do not necessarily have to be well calibrated, they only need to be higher than those of every other class to be accurately classified.

1.4 Validation

To test the performance of the model, I apply the predictions to the combined North Carolina and Florida State voter file.[8] These files contain snapshots of the registered voters in their respective states and provide individual-level data for first names, last names, address, political party, age, gender, and—crucially—self-identified ethnorace. After combining the two voter files, I then geocode each unique address in the sample. This allows me to match individual observations to Census places, tracts, and blocks. Then I apply the prediction algorithm described above using the **bper**

[8]North Carolina version dated January 1, 2019 and Florida version dated February 12, 2019

package in R and calculate each individuals' predicted ethnorace. In order to compare my method against an existing benchmark, I also use the R package **wru** (Imai and Khanna 2016) to calculate ethnorace predictions for the same individuals.

Unlike typical machine learning techniques, the methods implemented by both **bper** and **wru** do not fit a model on some sub-sample, or training set, of the data and then compare predictions against a held-out test set. Instead, the conditional probabilities for each attribute and ethnorace described in the Inputs section are merged into voter file from the input data sources. This allows the entire voter file to be used for validation. Together, these two states represent 23,754,749 individuals. Compared to nationwide percentages, Florida has a higher proportion of Hispanics and North Carolina has a higher proportion of African Americans. When combined they form a reasonably ethnoracially diverse population—2% Asian, 16% Black, 11% Hispanic, 6% Other Race, 65% White.

Researchers might be interested to know the extent to which prediction accuracy in North Carolina and Florida are transferable to other parts of the country. To this end I calculate separate prediction metrics for subsets of these two states based on demographic features. These features include the urban/rural Census classification, as well as income level, percentage foreign-born, and percentage with a Bachelor's degree or higher divided into terciles (Low, Middle, High). All these demographic features are calculated at the Census tract level. The goal with these sub-sample tests is to provide some information about how accurate **bper**'s ethnorace predictions are in places with more specific demographic profiles.

1.4.1 Predictive Performance

The predictive performance of **bper**'s classifications varies by the specific set of input data used as well as by the demographic characteristics of the target population. To assess its performance

based on different constraints, I ran separate tests on every combination of geography (county, place, tract, block), party/no party, and age+gender/no age+gender as inputs. All tests use first name and last name inputs. This yielded approximately 150 total tests.

For each of these tests I computed several metrics quantifying predictive performance. The first is the Accuracy score, or Overall Error Rate. This number is the proportion of correctly classified individuals in the sample. Figure 1.1 shows the Accuracy results sorted in ascending order across all tests. The tick marks in the bottom panel of Figure 1.1 show which input data or demographic sub-sets were included in a given test.

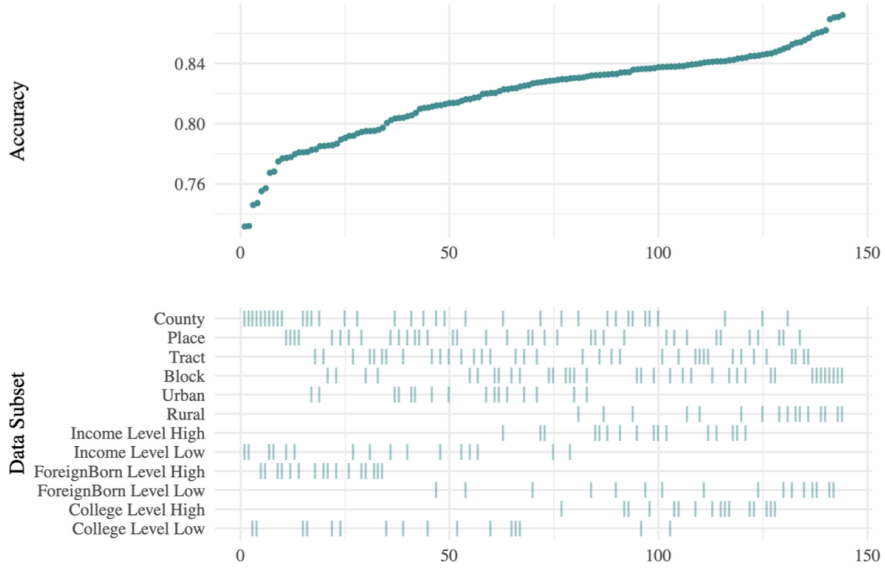

Figure 1.1: Predictive Performance: Accuracy

Accuracy scores range from 0.73 (using county-level geography and sub-setting to low income tracts) to 0.87 (using block-level geography and sub-setting to rural tracts). A few broad patterns emerge from the bottom panel in Figure 1.1. Accuracy generally improves with increased geographic precision (from counties to places, to tracks, to blocks). Rural areas have higher Accuracy than urban areas. High income areas have higher Accuracy than low income areas. Low foreign born percentage areas have higher Accuracy than high foreign born percentage areas. And high education

areas have higher Accuracy than low education areas.

Figure 1.2 compares the predictions from **bper** with **wru** (Imai and Khanna 2016). The baseline of zero in the upper panel corresponds to Accuracy scores from each model where the difference is zero in the test, and the height of the bars display the change in Accuracy using **bper**. Across all tests, **bper** wins 98.6% of comparisons. When using place-level geography in low education level areas, **bper**'s Accuracy is 9.5 percentage points higher than **wru**.

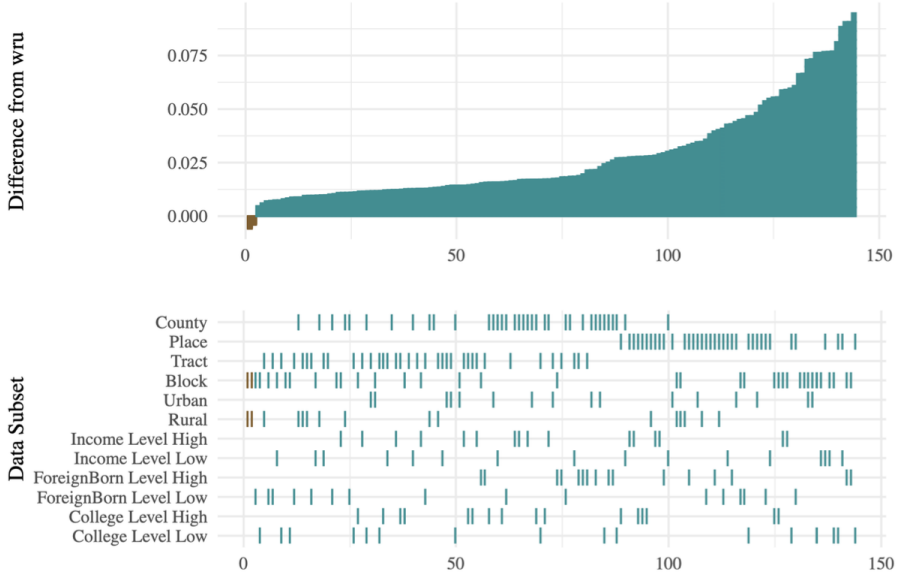

Figure 1.2: **bper** vs **wru** Predictive Performance: Accuracy

Accuracy scores, however, are an incomplete metric for assessing predictive performance. In contexts where the true distribution of classes is highly imbalanced, Accuracy can provide overly-optimistic results. For example, if we were to simply classify every individual as White in the North Carolina/Florida voter file, we would achieve 65% Accuracy automatically. We can evaluate the models in a more rigorous way by looking at each ethnorace category separately.

Two metrics useful for analyzing group-level predictions are Precision and Recall. Precision is the percentage of correctly classified individuals among all individuals predicted to belong to a specific ethnorace. It answers the question of how likely an individual's predicted ethnorace in

our sample matches their self-identified, ethnorace. Recall, also known as Sensitivity or the True Positive Rate, is the percentage of all individuals belonging to a given ethnorace group which the model correctly identifies.

Precision and Recall reflect substantively important concerns for real-world applications of the method, and the inherent trade-offs between optimizing for either metric provide a balanced assessment of the method's predictive performance. On the one hand, Precision rewards very conservative classification procedures. We could, for example, only classify individuals as White if their predicted probability of being White was greater than 99%. This would ensure a very high Precision score for Whites because we are only capturing the low-hanging fruit. A conservative classification procedure like this, however, would likely result in extremely low Recall for Whites. If we only capture the low-hanging fruit, a greater share of White individuals will be mis-classified as non-White. Likewise, optimizing the algorithm for perfect Recall for Whites is trivial. By classifying every individual as White, we ensure 100% of Whites are correctly classified. Of course this procedure would result in low Precision for Whites because every non-White individual would be classified as White as well. Achieving both high Precision and high Recall, therefore, is a difficult task.

Figure 1.3 through Figure 1.10 compare the Precision and Recall scores of **bper** and **wru** across the four predominant ethnoracial groups: Hispanic, White, Black, and Asian.

In 95% of comparisons, **bper** out-performs **wru** on Hispanic Precision. And in 81.9% of comparisons, **bper** out-performs **wru** on Hispanic Recall (Figure 1.3 and Figure 1.4). Using block-level geography, an individual predicted to be Hispanic by **bper** is potentially 30 percentage points more likely to self-identify as Hispanic compared to the same prediction made by **wru**.

Predictions made by **bper** for White individuals (Figure 1.5 and Figure 1.6 follow a similar pattern as those from the Hispanic tests. Across a significant majority of tests, **bper** out-performs

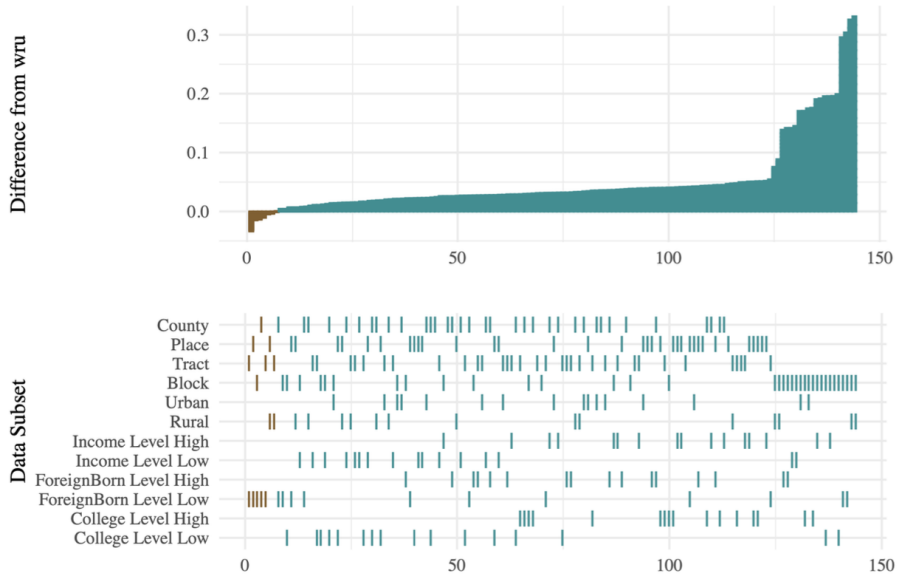

Figure 1.3: **bper** vs **wru** Predictive Performance: Hispanic Precision

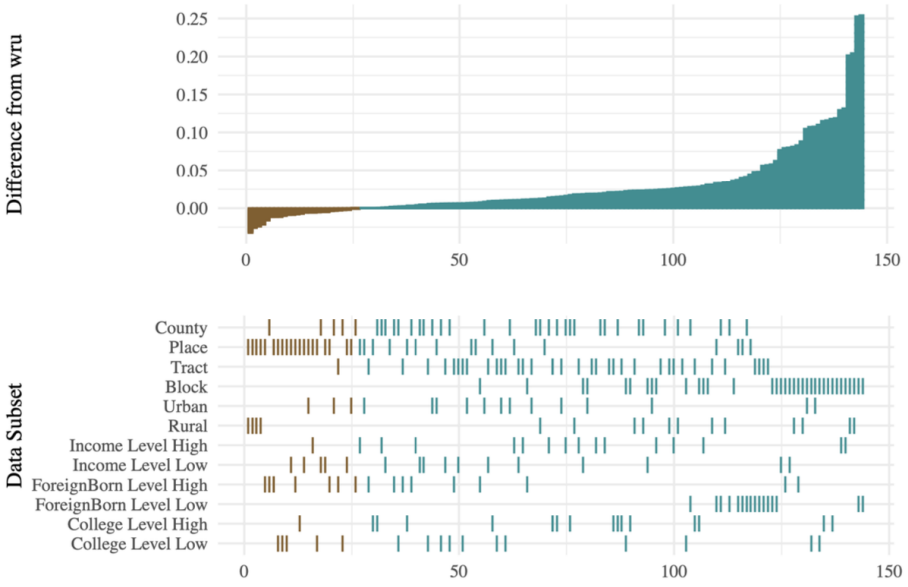

Figure 1.4: **bper** vs **wru** Predictive Performance: Hispanic Recall

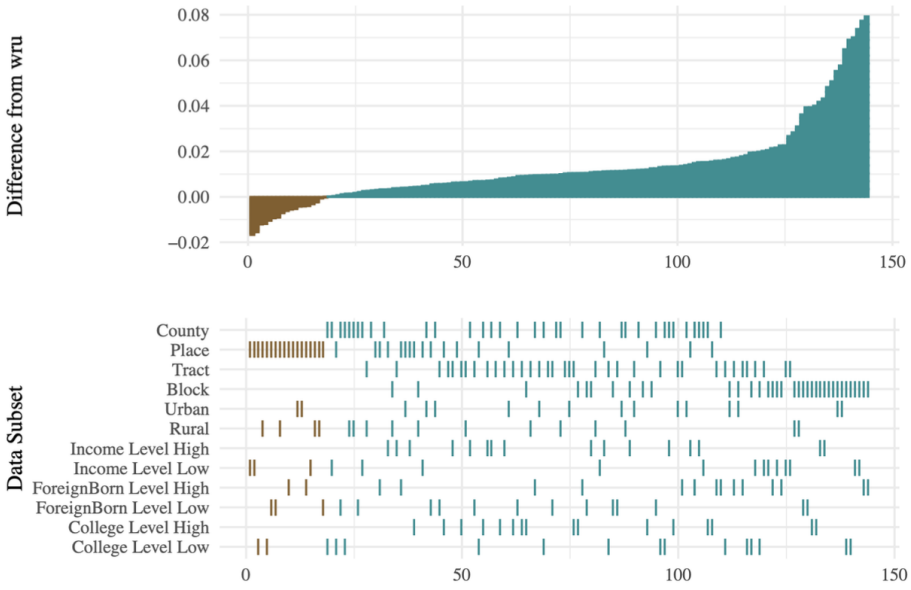

Figure 1.5: **bper** vs **wru** Predictive Performance: White Precision

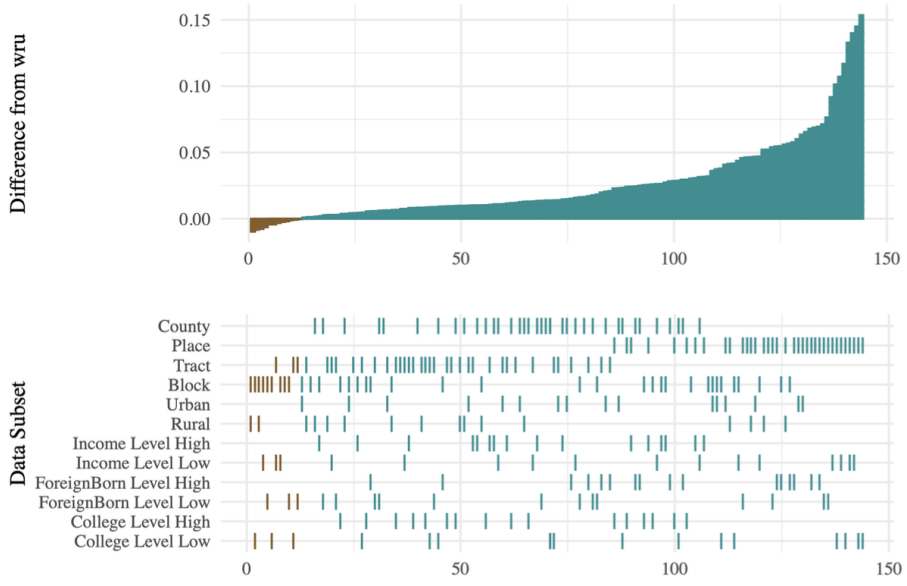

Figure 1.6: **bper** vs **wru** Predictive Performance: White Recall

wru in terms of Precision and Recall. The magnitudes for White predictions are lower than that of Hispanics' partly due to ceiling effects. Both models already do a much better job on White Precision and Recall relative to other ethnoraces.

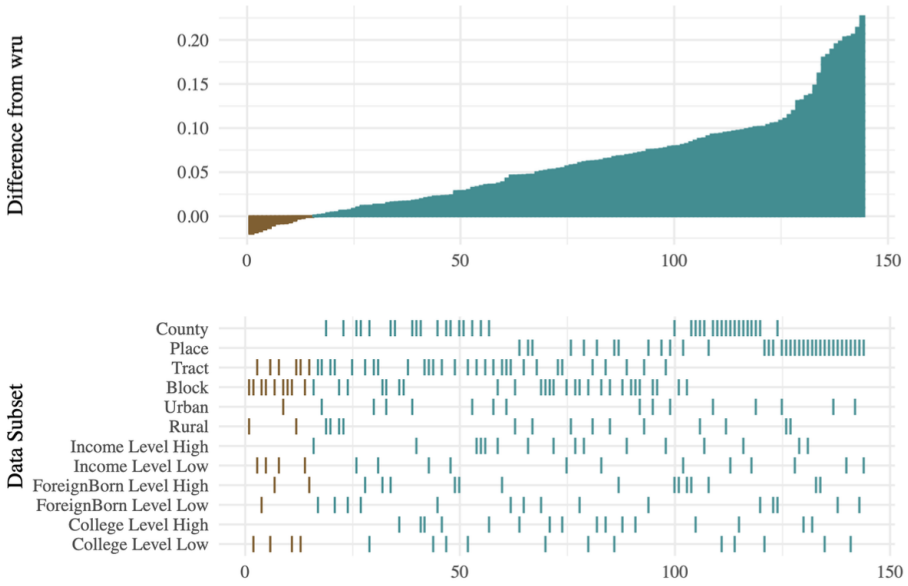

Figure 1.7: **bper** vs **wru** Predictive Performance: Black Precision

Figure 1.7 and Figure 1.8 show the predictive performance comparisons for African Americans. Similar to the Hispanic and White comparisons, **bper** makes significant gains in both Precision and Recall across a majority of tests. The largest exception to this pattern is for place-level geography, which appears to favor **wru** for Black Recall. However, place-level geography leads to the highest performance gains for Black Precision when using **bper**. Place-level predictions from **bper** increase the chance that someone predicted to be Black self-identifies as black by 10 to 20 percentage points.

The comparison tests for Asian Precision, shown in Figure 1.9, are roughly split between **bper** and **wru**. When using block-level geography and age/sex variables, however, **bper**'s Asian Precision jumps 20 percentage points higher than **wru**'s. Those inputs also lead to substantial gains in Asian Recall shown in Figure 1.10. Across all comparisons, Asian Recall is higher when using **bper**.

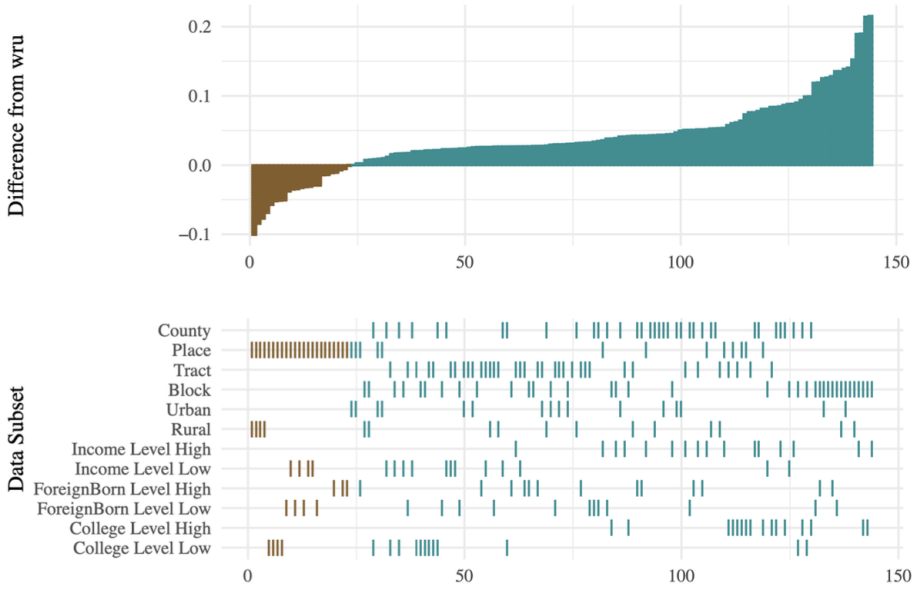

Figure 1.8: **bper** vs **wru** Predictive Performance: Black Recall

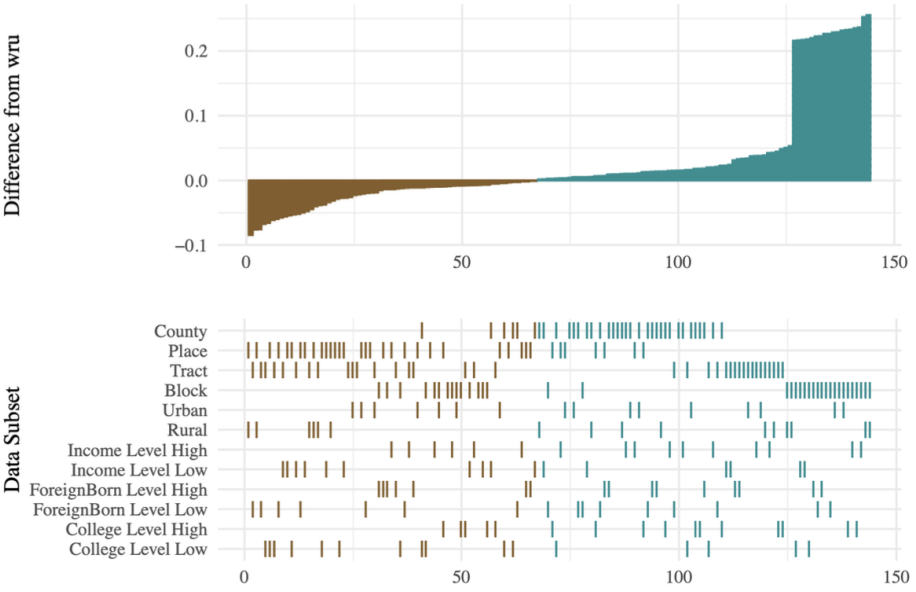

Figure 1.9: **bper** vs **wru** Predictive Performance: Asian Precision

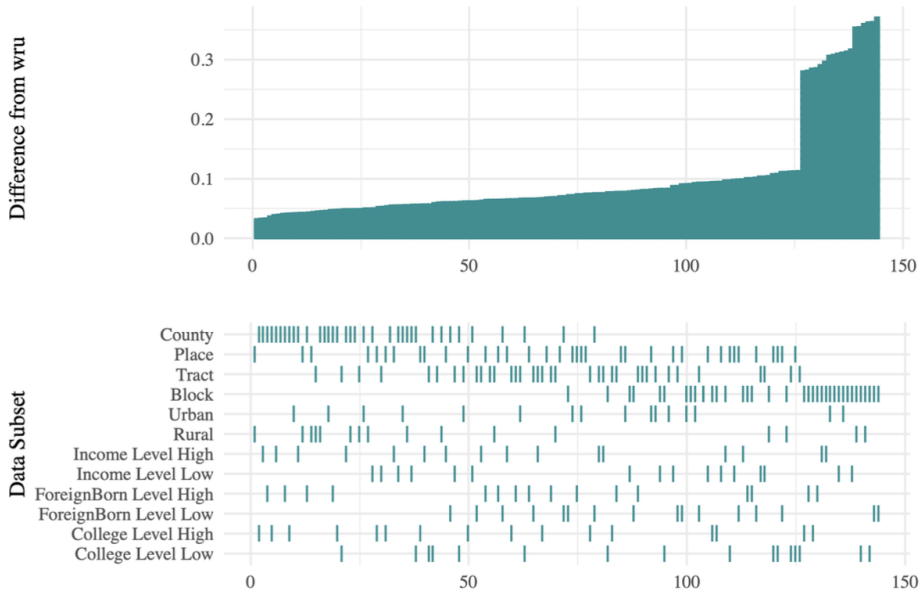

Figure 1.10: **bper** vs **wru** Predictive Performance: Asian Recall

1.4.2 wru Update 1.0.0

The analysis in the previous section was completed in February 2022 and built upon the method developed by Imai and Khanna (2016) and version 0.1-12 of the **wru** package. In May 2022 the authors released a new version of **wru** (1.0.0) which addressed some of the improvements I had incorporated into **bper**. For a comprehensive explanation of the updates and reanalysis of **wru** see Imai, Olivella, and Rosenman (2022). After re-running the comparison tests above, the two packages are now roughly comparable in overall predictive performance.

The latest version of **wru** adds two main changes. First, to address the issue of zero-counts I discussed above in the methodology section (Census geolocations with zero members of a particular ethnorace leading to zero probability that anyone living there belongs to that ethnorace), **wru** 1.0.0 uses a full Bayesian measurement model as opposed to the original Naive Bayes classifier algorithm. Ethnorace counts from geolocations are now built from a multinomial likelihood with a uniform Dirichlet prior, thereby ensuring that each ethnorace has a non-zero probability of belonging to any given geolocation. This method is more statistically principled than the Laplace smoothing

solution used in **bper**, which simply adds a constant pseudo-count value for each ethnorace in each geolocation. But overall, the two approaches towards dealing with zero-counts likely do not differ much in their impact on predictive performance. There were not large differences in **bper**'s predictive metrics with or without Laplace smoothing. Laplace smoothing, as opposed to full Bayesian measurement modeling, however, is much less computationally expensive.

The second major change in **wru** 1.0.0 is the addition of first and middle name data. These inputs add valuable information to the model and are likely the main source of the gains in predictive performance. Unlike **bper**, which uses first name data from mortgage applications (Tzioumis 2018), **wru** gets its name data from the state voter files of Alabama, Florida, Georgia, Louisiana, North Carolina, and South Carolina. While every state is required by federal law to compile voter lists, only these six Southern states provide information on voters' self-reported race and ethnicity. I do not use name data from these voter files in **bper** for two reasons. First, the data from some states cost as much as $5,000 to access. So collecting it is prohibitively expensive. Second, I use the two states with free voter files, North Carolina and Florida, as my validation data. Using names from these states with which to train the classification model will therefore over-fit to the sample, and may lead to predictions which generalize poorly to other states.[9]

1.5 Replication Study

In "Race and Representation in Campaign Fiance" (2020), Jacob Grumbach and Alexander Sahn investigate racial inequality in campaign contributions. Descriptively, they find that Latinos and African Americans are underrepresented among individuals who donate to US House campaigns. They also use differences-in-differences and regression discontinuity analyses to show that when

[9]Imai, Olivella, and Rosenman (2022) handle this over-fitting issue by running their predictive checks on each of the six states individually, using only the name data from the remaining five states.

candidates of color run, the contributor class becomes more ethnoracially representative because these candidates garner more co-ethnic donations.

Grumbach and Sahn rely primarily on the Dataset on Ideology and Money in Elections (DIME) from Bonica (2013). These data contain information—such as first names, last names, genders, addresses (for contributors), and political parties (for recipients)—for both the contributors and recipients of campaign contributions from 1980 to 2014. Crucially, however, the DIME data do not contain individuals' self-reported race or ethnicity. Grumbach and Sahn therefore use **wru** to predict the ethnorace of both contributors and recipients. By replacing the ethnorace predictions used in their study with those from **bper** I am able to demonstrate the substantive implications of switching to more accurate predictions.

The DIME data contain Census tract data for contributors, so I use that level of geography along with first name, last name, and gender as input data in **bper** for predicting contributor ethnoraces. The predictions from **wru** use these same inputs excepting first names. Candidate data is limited to state-level geography, which I use along with first names, last names, and party ID for **bper** predictions. The **wru** package does not have an option for state-level geography, so I only use last name and party ID inputs for its candidate ethnorace predictions.

Table 1.1 shows the prediction metrics tested in the Validation section for the set of variables used as inputs for contributor ethnorace predictions. Because individuals who donate to political campaigns are more affluent than the average voter, I used the validation test which sub-setted to only high income areas. Although it is impossible to validate the ethnorace predictions on the DIME data set to the same degree as the North Carolina and Florida voter files, the metrics in Table 1.1 should give a rough sense of how trustworthy the ethnorace predictions used by Grumbach and Sahn (2020) are. Across all metrics, **bper** out-performs **wru** when using tract-level geography and gender as inputs. Individuals are more likely to self-identify as their predicted ethnorace

Table 1.1: Prediction Metrics of DIME Contributor Ethnoraces

Metric	bper	wru	bper Difference
Accuracy	0.834	0.819	0.015
Precision			
Aapi	0.582	0.544	0.038
Black	0.313	0.244	0.069
Hispanic	0.758	0.721	0.037
White	0.970	0.964	0.006
Recall			
Aapi	0.574	0.498	0.076
Black	0.719	0.668	0.051
Hispanic	0.735	0.709	0.026
White	0.856	0.844	0.012

(Precision) and more individuals from each group are correctly predicted (Recall). Despite these gains in predictive performance, however, Black and Asian predictions appear to be unreliable. There is less than a 1 in 3 chance for Black predictions and a 1 in 2 chance for Asian predictions to accurately reflect the individual's self-identified race. And close to half of all Asian contributors may be incorrectly predicted.

Figure 1.11 (Figure 2 in Grumbach and Sahn 2020) shows the proportion of total campaign contributions by ethnorace by year.[10] The left-hand panel shows these results using **wru**'s predictions and matches closely the figure from the published article. Compared to **wru**, the results from **bper** show lower shares of contributions from individuals of color. Additionally, there appears to be a dramatic decline over time in contributions from African Americans using **bper** predictions. African Americans accounted for 5.95% of the contributions in 1980 compared to only 1.9% in 2014. This pattern exists, but is less dramatic in the **wru** panel (7.59% Black contributions in 1980 compared to 2.9% in 2014). Switching to **bper** predictions thereby reveals a contributor class which is less diverse than the one described by Grumbach and Sahn.

[10] As in the original figure, contributions from Whites are omitted to display Asian, Black, and Hispanic contributions better.

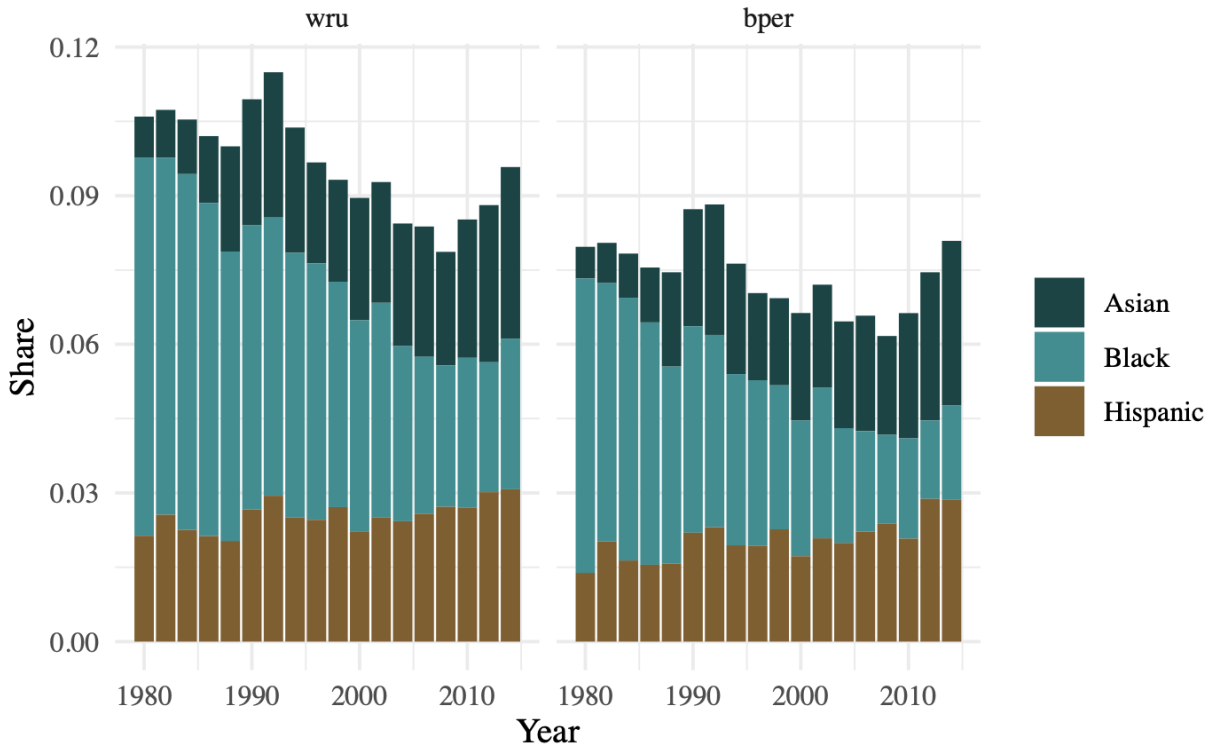

Figure 1.11: Ethnoracial Composition of the Contributor Class (1980-2014)

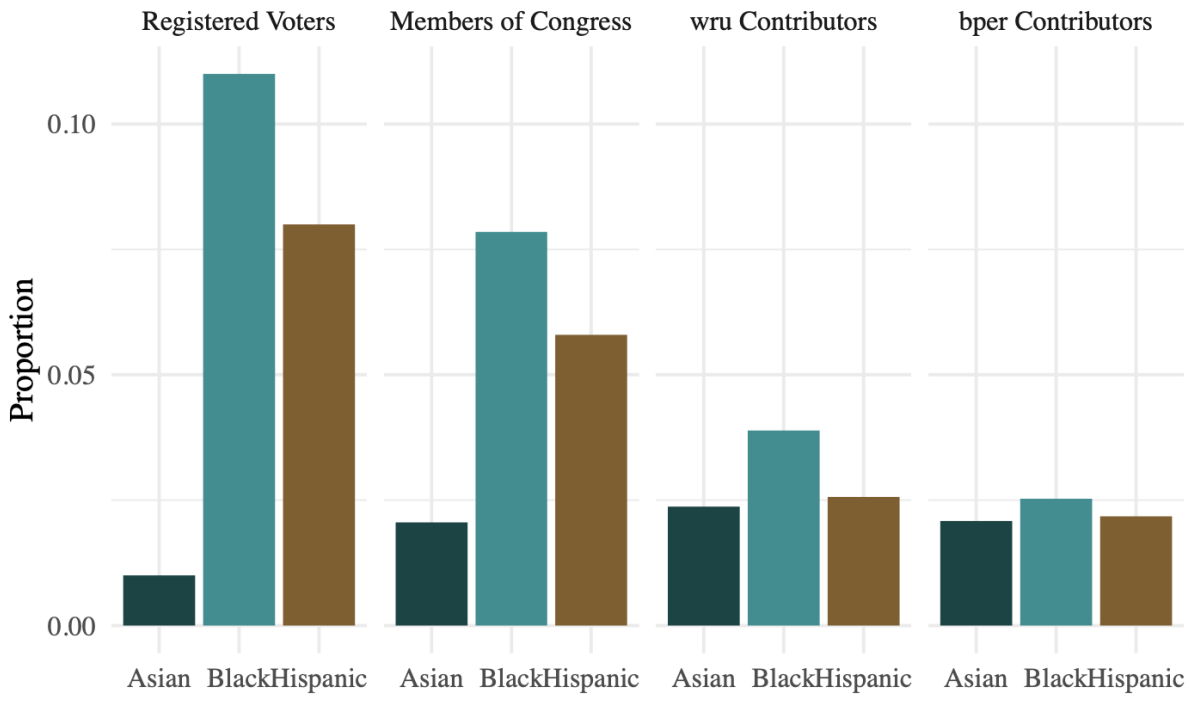

Figure 1.12: Ethnoracial Composition of the Contributor Class Versus Electorate

Whereas Figure 1.11 showed the proportions of contributions by ethnorace, Figure 1.12 (Figure 3 in original) displays the proportions of contributors by ethnorace relative to other forms of political participation. Both **wru** and **bper** predictions show that Black and Hispanic Americans are underrepresented in the contributor class. This finding, along with the increased representation of Asian Americans relative to their proportion among registered voters, replicates Grumbach and Sahn's descriptive analysis from the published article. However, **bper** predictions show a substantially lower share of Black contributors compared to the predictions from **wru**.

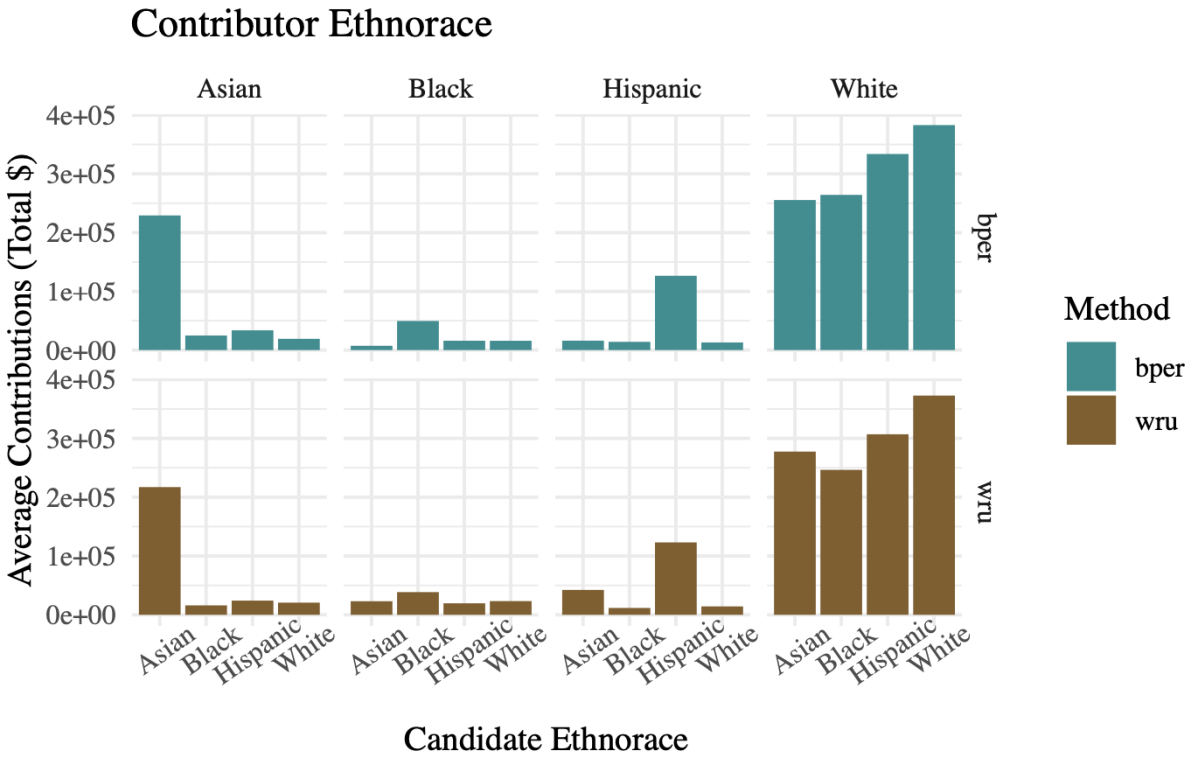

Figure 1.13: Average Contributions by Ethnorace

Figure 1.13 (Figure 4 in original) uses both the ethnorace of contributors (columns) and the ethnorace of candidates (x-axis). It displays the average total contributions candidates receive from contributors of a given ethnorace. Similar to the original study, I find strong evidence of co-ethnic contributing using **bper**. Among each ethnorace's contributors, either a plurality or a majority of

contributions go to candidates who share the same ethnorace. This finding, however, is supported more strongly by the **bper** predictions than the **wru** predictions. Contributions from Black donors appear to be more heavily concentrated among Black candidates when using **bper** compared to **wru**. The same pattern of increased co-ethnic concentration appears for Hispanic contributors and candidates.

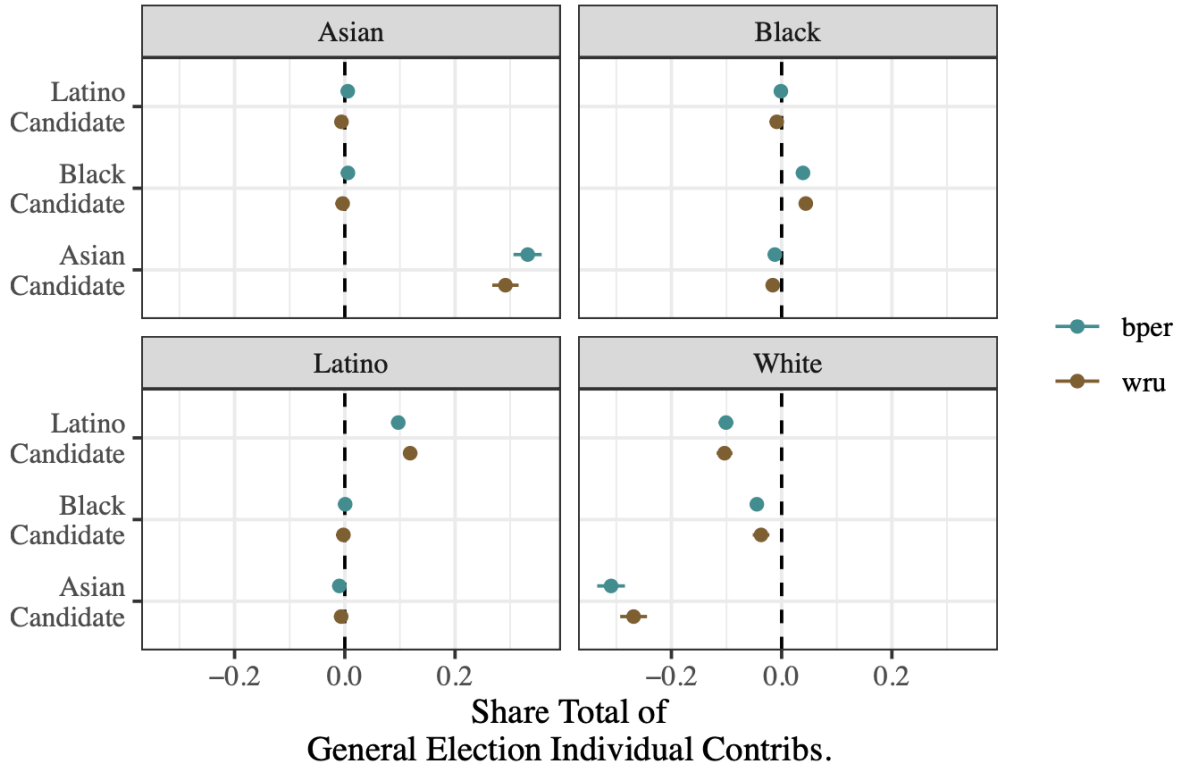

Figure 1.14: Effect of Candidate Ethnorace on Share of Contributions by Ethnorace

While Figure 1.13 shows strong descriptive evidence of co-ethnic contributions, Grumbach and Sahn attempt to further isolate this causal effect by using a differences-in-differences analysis. Figure 1.14 show the main results of their two-way fixed effects models they use to estimate this effect. If their causal assumptions hold, the coefficients in each panel of Figure 1.14 show the effect of a candidate belonging to a co-ethnic group (relative to a White candidate) on the share of total general election individual contributions from donors of the same ethnorace. Models using

predictions made by **bper** generally match those by **wru**. The presence of a co-ethnic candidate increases the amount of contributing by co-ethnic donors.

1.6 Conclusion

This paper describes a powerful method for predicting individuals' race or ethnicity based on other known attributes. Through validation tests, I have shown that my method produces better predictions than older methods used in recent social science research. The primary gains in predictive performance come from using individuals' first names as inputs. These improvements are not trivial. Depending on the input data available, my method can nearly double the probability that someone predicted to be Black or Asian self-identifies as that race relative to predictions generated from other methods (Imai and Khanna 2016). These improved predictions reveal new empirical findings as demonstrated by my replication of Grumbach and Sahn (2020).

1.7 References

Abott, Carolyn, and Asya Magazinnik. 2020. "At-Large Elections and Minority Representation in Local Government." *American Journal of Political Science* 64 (3): 717–33. https://doi.org/10.1111/ajps.12512.

Bonica, Adam. 2013. "Ideology and Interests in the Political Marketplace: *Ideology and Interests in the Political Marketplace*." *American Journal of Political Science* 57 (2): 294–311. https://doi.org/10.1111/ajps.12014.

Burch, Traci. 2013. *Trading Democracy for Justice: Criminal Convictions and the Decline of Neighborhood Political Participation*. The University of Chicago Press.

Clark, Jesse T., John A. Curiel, and Tyler S. Steelman. 2021. "Minmaxing of Bayesian Improved Surname Geocoding and Geography Level Ups in Predicting Race." *Political Analysis*, November, 1–7. https://doi.org/10.1017/pan.2021.31.

Crabtree, Charles, and Volha Chykina. 2018. "Last Name Selection in Audit Studies." *Sociological Science* 5: 21–28. https://doi.org/10.15195/v5.a2.

Davenport, Lauren. 2020. "The Fluidity of Racial Classifications." *Annual Review of Political Science* 23 (1): 221–40. https://doi.org/10.1146/annurev-polisci-060418-042801.

Domingos, Pedro, and Michael Pazzani. 1997. "Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier." *Machine Learning* 29: 103–30.

Elliott, Marc N., Peter A. Morrison, Allen Fremont, Daniel F. McCaffrey, Philip Pantoja, and Nicole Lurie. 2009. "Using the Census Bureau's Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities." *Health Services and Outcomes Research Methodology* 9 (2): 69–83. https://doi.org/10.1007/s10742-009-0047-1.

Enos, Ryan D., Aaron R. Kaufman, and Melissa L. Sands. 2019. "Can Violent Protest Change Local Policy Support? Evidence from the Aftermath of the 1992 Los Angeles Riot." *American Political Science Review* 113 (4): 1012–28. https://doi.org/10.1017/S0003055419000340.

Fraga, Bernard L. 2018. *The Turnout Gap: Race, Ethnicity, and Political Inequality in a Diversifying America*. https://doi.org/10.1017/9781108566483.

Grumbach, Jacob M., and Alexander Sahn. 2020. "Race and Representation in Campaign Finance." *American Political Science Review* 114 (1): 206–21. https://doi.org/10.1017/S0003055419000637.

Grumbach, Jacob M., Alexander Sahn, and Sarah Staszak. 2020. "Gender, Race, and Intersectionality in Campaign Finance." *Political Behavior*, June. https://doi.org/10.1007/s11109-020-09619-0.

Hepburn, Peter, Renee Louis, and Matthew Desmond. 2020. "Racial and Gender Disparities Among Evicted Americans." *Sociological Science* 7: 649–62. https://doi.org/10.15195/v7.a27.

Imai, Kosuke, and Kabir Khanna. 2016. "Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records." *Political Analysis* 24 (2): 263–72. https://doi.org/10.1093/pan/mpw001.

Imai, Kosuke, Santiago Olivella, and Evan T. R. Rosenman. 2022. "Addressing Census Data Problems in Race Imputation via Fully Bayesian Improved Surname Geocoding and Name Supplements." *Science Advances* 8 (49): eadc9824. https://doi.org/10.1126/sciadv.adc9824.

Kuk, John, Zoltan Hajnal, and Nazita Lajevardi. 2020. "A Disproportionate Burden: Strict Voter Identification Laws and Minority Turnout." *Politics, Groups, and Identities*, June, 1–9.

https://doi.org/10.1080/21565503.2020.1773280.

Lewis, David D. 1998. "Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval." In, edited by Claire Nédellec and Céline Rouveirol, 1398:4–15. Berlin, Heidelberg: Springer Berlin Heidelberg. http://link.springer.com/10.1007/BFb0026666.

Omi, Michael, and Howard Winant. 2014. *Racial Formation in the United States*. Routledge.

Rish, Irina. 2001. "An Empirical Study of the Naive Bayes Classifier." *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence* 3 (22): 41–46.

Sanchez, Gabriel R. 2006. "The Role of Group Consciousness in Latino Public Opinion." *Political Research Quarterly* 59 (3): 12.

Sen, Maya, and Omar Wasow. 2016. "Race as a Bundle of Sticks: Designs That Estimate Effects of Seemingly Immutable Characteristics." *Annual Review of Political Science* 19 (1): 499–522. https://doi.org/10.1146/annurev-polisci-032015-010015.

Studdert, David M., Yifan Zhang, Sonja A. Swanson, Lea Prince, Jonathan A. Rodden, Erin E. Holsinger, Matthew J. Spittal, Garen J. Wintemute, and Matthew Miller. 2020. "Handgun Ownership and Suicide in California." *New England Journal of Medicine* 382 (23): 2220–29. https://doi.org/10.1056/NEJMsa1916744.

Tzioumis, Konstantinos. 2018. "Demographic Aspects of First Names." *Scientific Data* 5 (1): 180025. https://doi.org/10.1038/sdata.2018.25.

Voicu, Ioan. 2018. "Using First Name Information to Improve Race and Ethnicity Classification." *Statistics and Public Policy* 5 (1): 1–13. https://doi.org/10.1080/2330443X.2018.1427012.

Chapter 2

Ideal Point Estimation with 99% Missing Data

2.1 Introduction

Many variables of interest in the social sciences defy easy measurement. Some latent characteristics, such as ideology, cannot be observed directly and must instead be inferred from the actions taken by political actors. One popular statistical method for estimating latent characteristics from observed actions is known as Item Response Theory (IRT). These models originated in psychometric research in an effort to measure the ability of individual test takers (Rasch 1980), but have also been adapted to measure political ideology (Clinton, Jackman, and Rivers 2004). In general, IRT models can be used to measure latent traits when a set of actors signal their observed preference among two or more choices. A common example of is legislators voting yea or nay on bills in Congress. IRT has been used to measure the political ideology of the US public (Treier and Hillygus 2009; Caughey and Warshaw 2015), of Supreme Court justices (Martin and Quinn 2002), of Twitter users (Barberá 2015), and the alignment of countries in the UN General Assembly (Bailey, Strezhnev,

and Voeten 2017).

While these ideal point models do not strictly require that the entire set of actors in question take visible positions on every policy choice presented before them, in practice this is almost always the case. Most members of the United States Congress vote on most bills, most Supreme Court justices take positions on cases presented before them, and most survey respondents answer all questions on their survey. There are other contexts, however, where abstaining from taking a position is more common. Failing to factor abstentions into models of actors' ideology is valid only if these missing positions are ignorable—in other words, if the actors are abstaining at random rather than due to their ideological characteristics. Rosas, Shomer, and Haptonstahl (2015) show that non-ignorable abstentions can lead to misleading estimates of legislator ideology in legislatures outside the US. In the Israeli Knesset, for example, abstentions are common and likely represent some aspect of legislators' ideologies.

Another context with high rates of missing data are interest groups' signals of support or opposition for particular pieces of legislation. Using Map-light data[1] on interest group positions, Crosson, Furnas, and Lorenz (2020) measure the political ideology of interest groups lobbying in the US Congress. Although the Map-light data has information on 8,494 groups taking one or more positions across 16,436 bills, the probability that any particular group sends a signal on a particular bill is extremely low. *Most* groups fail to signal a position on *most* bills. If interest groups in the US were a legislative assembly, they would have an abstention rate of over 99.99%.

In this paper I build on the IRT model developed by Rosas, Shomer, and Haptonstahl (2015) to account for the massive amounts of missing data among interest group positions. This model allows each group to have an independent "indifference" parameter which controls the probability

[1]Lorenz, Geoffrey M., Alexander C. Furnas, and Jesse M. Crosson. "Large-N Bill Positions Data from Map-Light.Org: What Can We Learn from Interest Groups' Publicly Observable Legislative Positions?" *Interest Groups & Advocacy* 9, no. 3 (September 2020): 342–60.

that they will signal any position on a particular bill or choose to abstain. The indifference parameter could reflect a number of interest group characteristics. Some interest groups, such as the business-focused Chamber of Commerce, could have issue-areas which touch a wider range of legislation than more narrowly focused groups, such as the National Rifle Association. "Indifference" could also reflect the amount of resources an interest group has—sending signals on legislation requires time and effort.

My primary goal is extending the analysis done by Rosas, Shomer, and Haptonstahl (2015), and showing how common IRT techniques used in political science dramatically fail in contexts where a vast majority of signals are missing. I also contribute to overall use of IRT models in political science by incorporating the latest advances in the field of Bayesian computational methods into the analysis. Finally, I replicate the ideal point estimation from Crosson, Furnas, and Lorenz (2020) and show how my model yields a diverging view of political polarization among interest group lobbying.

2.2 Ideal Point Models

Ideal point models in political science are types of measurement models. These methods use observed actions to inform us about the latent traits of various political actors. Using interest groups as an example, let's imagine that each group, i has an political ideal point, θ_i which lies on a single left-right, or liberal-conservative, scale. The probability that a given group signals its support or opposition to a particular piece of legislation, j is determined by the distance between the group's ideal point and the ideological content of the bill, γ_j. We also assume that each bill has a component, ξ_j which controls the probability, independent of political ideology, that it receives support from an interest group. Putting these parameters together, along with an assumption that interest groups have some Normal(0, 1) idiosyncratic error in their signalling behavior (which we

encode using the standard normal cumulative density function, Φ), gives us the following model which can be estimated from observed data:

$$\Pr(y_{ij} = \text{Support}) = \Phi(\gamma_j \theta_i + \xi_j) \tag{2.1}$$

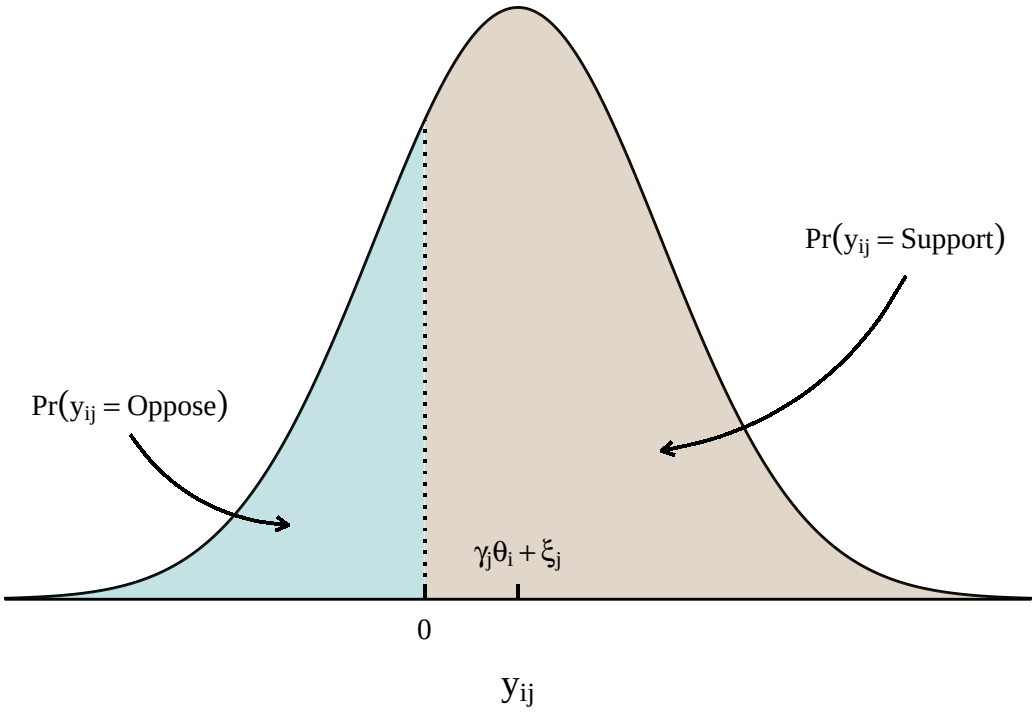

Figure 2.1: Binary IRT Model

The binary IRT model from Equation 2.1 is shown graphically in Figure 2.1. The shaded density regions are proportional to the relative probability of a particular interest group either supporting or opposing a particular bill, y_{ij}. Larger values of $\gamma_j \theta_i + \xi_j$ lead to higher probabilities that $y_{ij} = \text{Support}$, and lower values lead to higher probabilities that $y_{ij} = \text{Oppose}$. If $\gamma_j \theta_i + \xi_j = 0$, the interest group is 50-50 on whether to support or oppose the bill. This follows from the mathematical evaluation of the Normal(0, 1) CDF at zero: $\Phi(0) = 0.5$.

Equation 2.1 is equivalent to the 2-parameter IRT model used in psychometric research (Fox

2010). As it stands, however, Equation 2.1 is not identified. This means that, for given data, y_{ij} there is not a unique set of parameter values for θ_i, γ_j, and ξ_j. There for three reasons for this. First, there is no unique location for the latent scale because adding a constant to the $\gamma_j\theta_i$ term can be offset by subtracting the same constant from β_j. In other words, while it may be natural to think of zero as the center of a political ideology scale, nothing in Equation 2.1 defines the center of the latent variable. Second, the scale can be arbitrarily stretched or compressed by multiplying and dividing the terms by the same constant. And third, the polarity of the scale is not uniquely identified. There is no information in Equation 2.1 to tell use whether larger values of θ_i correspond to more liberal ideal points or to more conservative ideal points. I discuss how each of these identification issues are handled in the following section.

While Equation 2.1 assumes that the response data are binary (support or oppose) and uses the probit link function, Φ to transform the nonlinear predictor into a probability, IRT models can support a wide range of other outcome types (Bürkner 2020). Public opinion surveys often ask respondents to rank their support or opposition to some policy on a five or seven point scale. In these situations either an ordinal or multinomial link function is more appropriate rather than collapsing responses to a binary support/oppose (Hill and Tausanovitch 2015). Likelihoods such as the poisson distribution can also be used if the observed outcomes are a count of positions taken (Slapin and Proksch 2008). As I demonstrate in the next section, the choice of response function can have a dramatic effect on inferences that come out of IRT ideal point models.

2.3 Abstention Ideal Point Model

Recall that the Map-Light data reports interest group positions (support/oppose) on bills brought before the US Congress. While it may appear that these observed positions are binary, interest groups have a third option available to them when a piece of legislation is introduced: to

abstain from taking any position at all. In fact, this is by far the most common response interest groups choose. For any given bill, over 99% of interest groups will not send a signal of either support or opposition. Most interest groups focus on narrow policy areas, and sending a signal of support an opposition is costly, so this high abstention rate should be expected. Ideal point models, however, assume that *all* pieces of legislation contain some amount of ideological content along the uni-dimensional left-right scale. Therefore we should consider interest group abstentions as something like a middle-ground position that lies between signals of outright support and outright opposition.

Rosas, Shomer, and Haptonstahl (2015) develop an IRT ideal point model which explicitly accounts for abstentions in legislatures. Every actor is given an additional parameter τ_i which controls its level of indifference. The larger the absolute value of τ_i, the less likely the actor will signal support or opposition on a given bill. Equation 2.2 and Figure 2.2 show the expanded model:

$$\begin{aligned}
\Pr(y_{ij} = \text{Support}) &= \Phi(\gamma_j\theta_i + \xi_j - \tau_i) \\
\Pr(y_{ij} = \text{Abstain}) &= \Phi(\tau_i - (\gamma_j\theta_i + \xi_j)) - \Phi(-\tau_i - (\gamma_j\theta_i + \beta_j)) \\
\Pr(y_{ij} = \text{Oppose}) &= 1 - \Phi(\gamma_j\theta_i + \xi_j + \tau_i)
\end{aligned} \tag{2.2}$$

It is important to note that the model in Equation 2.2, while almost always superior to the simple binary IRT model in Equation 2.1, may not be appropriate if the data generating process governing missing data differs from the one described above. I am assuming that all interest groups have the opportunity to send signals on all federal bills. Their signal is missing if they choose to abstain, which is a choice controlled independently by their value of τ. Another plausible way to explain missingness in interest group positions, however, is that they are dependent on the content or topic of the bill. The interest group Planned Parenthood will never send a signal of support or opposition to a bill on steel tarriffs because steel tarriffs have nothing to do with the issue-area of Planned Parenthood.

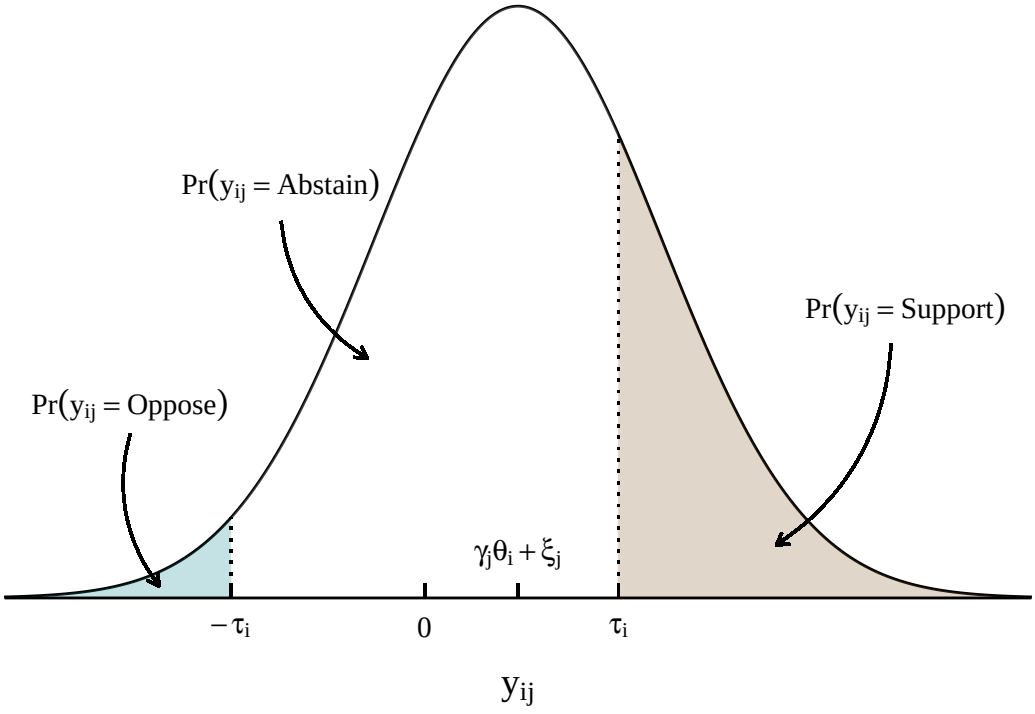

Figure 2.2: Abstention IRT Model

This general pattern of selective interest group behavior could be modeled in a couple of ways. First, rather than being a static characteristic of the interest group, τ could be allowed to vary with an additional parameter related to each bill. This interaction would have the effect of growing or shrinking the probability of a group's abstention based on the content of the bill. A second modeling option would be to add a hurdle component. In a hurdle model, the binary IRT model would be used to determine whether a group signals support or opposition to a bill, and a separate process would control whether the bill is sufficiently salient for the group to send any signal in the first place. Statistically, this would look like a nested probit model. The first probit deals with the question of abstention versus signalling, and the second probit deals with the question of support versus opposition conditional on having chosen to signal.

These alternative modeling strategies offer a more nuanced view of interest group behavior compared to the one encoded in Equation 2.2. Ideally, a new bespoke ideal point model should

be built for each set of political actors in question—using the particularities of each unique data generating process to guide model construction. But, because researchers tend to prefer off-the-shelf models in practice, I provide the model in Equation 2.2 as an alternative to the classic binary ideal point model. As I will demonstrate later in this project, Equation 2.2 offers a straightforward improvement to existing ideal point methods in contexts with large amounts of missing data.

We can estimate the group and bill parameters in Equation 2.2 using a Bayesian ordered probit model with group-specific thresholds. A fully-specified Bayesian model requires that we place priors on all parameters. This has the added benefit in the context of IRT models of resolving the location and scale identification issues described earlier. Priors work to constrain the plausible space of parameter values, which means we can use them to enforce a center on the ideal point distribution as well as set a reasonable scale to the values. In my model I assign the following prior distributions to the main parameters:

$$\begin{aligned} \tau_i &\sim \text{Normal}(0, 4) \\ \theta_i &\sim \text{Normal}(0, 3) \\ \gamma_j &\sim \text{Normal}(0, 3) \\ \beta_j &\sim \text{Normal}(0, 2) \end{aligned} \tag{2.3}$$

In order to identify the polarity of the ideal point scale I include covariates to help predict the θ_i and γ_j parameters.[2] A binary indicator for whether the interest group is classified as representing business interests, along with a binary indicator for whether the bill was authored by a Republican member of Congress, are added to the model. I constrain each of these coefficients to be positive by using a LogNormal(-1, 1) prior distribution. The center of this distribution is $exp(-1) = 0.368$ which is reasonably conservative on the probit scale. But the right-skewed nature of the LogNormal distribution allows these coefficients to be much larger if required by the data. Because θ_i and γ_j

[2]The parameter ξ does need any covariates because it represents the non-ideological component of the bill, and therefore is not relevant to the polarity identification issue.

now modeled with a linear predictor, they are given intercepts α_θ and α_γ.

$$\begin{aligned}
\theta_i &\sim \text{Normal}(\alpha_\theta + \delta_1 \text{Business}_i, 3) \\
\gamma_j &\sim \text{Normal}(\alpha_\gamma + \delta_2 \text{Republican}_j, 3) \\
\xi_j &\sim \text{Normal}(0, 2) \\
\delta_1, \delta_2 &\sim \text{LogNormal}(-1, 1) \\
\alpha_\theta, \alpha_\gamma &\sim \text{Normal}(0, 2)
\end{aligned} \tag{2.4}$$

The method of including informative covariates into ideal point models is not widely practiced in political science. Instead, in order to identify the polarity of the scale researchers typically fix the ideal points of two actors to constants (Bafumi et al. 2005). An example from the US Congress would be to fix liberal Bernie Sanders's ideal point to -2 and conservative Ted Cruz's to 2. This method has a few drawbacks. First, the results become sensitive to the modeler's *a priori* expectations for where certain actors are located on the latent scale. This might not be such a problem in contexts which have been heavily studied like US Congress members, but there are fewer theoretical expectations regarding where the ideal points of specific interest groups are located. Second, it is not necessarily possible to fix actors to specific locations when using hierarchical modeling techniques.

Hierarchical Bayesian modeling is another under-utilized technique among IRT ideal point models in political science. Rather than estimating each actor and bill parameter in a vacuum, hierarchical modeling (also referred to as multilevel modeling) allows the groups to partially-pool information from the population of actors and bills. This is done by modeling each parameter's prior as a function of the group mean, which itself is given a prior. Extending out the model with hierarchical priors we get:

$$
\begin{aligned}
\tau_i &\sim \text{Normal}(0, 4) \\
\theta_i &\sim \text{Normal}(\theta_{obs}, \sigma_\theta) \\
\theta_{obs} &= \alpha_\theta + \delta_1 \text{Business}_i \\
\gamma_j &\sim \text{Normal}(\bar{\gamma}, \sigma_\gamma) \\
\bar{\gamma} &= \alpha_\gamma + \delta_2 \text{Republican}_j \\
\xi_j &\sim \text{Normal}(\bar{\xi}, \sigma_\xi) \\
\delta_1, \delta_2 &\sim \text{LogNormal}(-1, 1) \\
\sigma_\theta &= 1 \\
\sigma_\gamma, \sigma_\xi &\sim \text{HalfCauchy}(0, 2)
\end{aligned}
\tag{2.5}
$$

The σ_θ parameter is fixed at 1 in order to help identify the scale of the ideal points (Bürkner 2020; Rosas, Shomer, and Haptonstahl 2015), but the rest of the variance parameters are allowed to vary. Hierarchical modeling is very powerful in contexts like the Map-Light data. Many interest groups only signal support or opposition a few times during the time period covered, so it is important that we use the population distribution of interest group ideal points to help inform us about these rare cases. Hierarchical models are also much better at predicting out of sample compared to ordinary "memory-less models" (Gelman and Hill 2007; McElreath 2020). This should give us more confidence that the ideal points produced by the model are a more accurate reflection of the interest groups' true latent ideology.

$$
\begin{aligned}
y_{ij} &\sim \text{Ordered.Categorical}(\mathbf{p}) \\
p_2 &= \Phi(\gamma_j\theta_i + \beta_j - \tau_i) \\
p_1 &= \Phi(\tau_i - (\gamma_j\theta_i + \beta_j)) - \Phi(\tau_i - (\gamma_j\theta_i + \beta_j)) \\
\tau_i &\sim \text{Normal}(0, 4) \\
\theta_i &\sim \text{Normal}(\theta_{obs}, \sigma_\theta) \\
\theta_{obs} &= \alpha_\theta + \delta_1 \text{Business}_i \\
\gamma_j &\sim \text{Normal}(\bar{\gamma}, \sigma_\gamma) \\
\bar{\gamma} &= \alpha_\gamma + \delta_2 \text{Republican}_j \\
\xi_j &\sim \text{Normal}(\bar{\xi}, \sigma_\xi) \\
\delta_1, \delta_2 &\sim \text{LogNormal}(-1, 1) \\
\sigma_\theta &= 1 \\
\sigma_\gamma, \sigma_\xi &\sim \text{HalfCauchy}(0, 2)
\end{aligned}
\tag{2.6}
$$

Putting all the pieces of the model together gives us Equation 2.6. This is an ordered probit model with three outcomes: support, abstain, and oppose. The cut-point p_1 reflects the probability of an interest group supporting a particular bill, and the cut-point p_2 reflects the probability of abstaining. The probability of opposing a particular bill is implicit in the remaining probability density, or $1 - \Phi(\gamma_j\theta_i + \beta_j + \tau_i)$. I fit this model using the probabilistic programming language Stan (Carpenter et al. 2017). Stan's Hamiltonian Monte Carlo (HMC) sampler does a better job of exploring the high dimensional parameter space in a model like Equation 2.6 than older Markov Chain Monte Carlo algorithms (Betancourt 2018). It also provides many diagnostic tools which alert users of possible problems with their models. This is invaluable for IRT models for which proper identification is a major concern.

The ordered probit model which accounts for interest group abstentions in Equation 2.6 (hereafter referred to as the "Abstention model") stands in contrast with the ideal point model used by Crosson, Furnas, and Lorenz (2020). The authors use the binary ideal point model developed by Clinton, Jackman, and Rivers (2004) via the R package **pscl**. The default **pscl** ideal point model

uses Equation 2.1 with the following priors:

$$\begin{aligned} \theta_i &\sim \text{Normal}(0, 1) \\ \gamma_j &\sim \text{Normal}(0, 25) \\ \beta_j &\sim \text{Normal}(0, 25) \end{aligned} \tag{2.7}$$

This method does not model abstentions explicitly, nor does it take advantage of interest group and bill covariates or use hierarchical priors. These drawbacks are not always apparent, and the **pscl** model does an excellent job recovering accurate ideal points in many contexts in which it has been used. But as I show in the following section, it is ill-suited for estimating ideal points when the data generating process produces massively abstention-skewed observations. And the MCMC algorithm used by **pscl** is not equipped with the same diagnostic tools as Stan's HMC sampler to alert users that these problems may be occurring.

2.4 Simulation Study

To test the accuracy of the Abstention model from Equation 2.6 and the **pscl** model, I simulate five data sets with 100 interest groups and 300 bills. Each interest group is given an ideal point drawn from a Normal(0, 1) distribution and an indifference parameter drawn from a $\text{Normal}_+(\mu, 0.5)$ distribution with a mean, μ varying across the five simulated data sets. Each interest group either supports, abstains, or opposes a given bill following the formula in Equation 2.2. Increasing μ from 0 to 4 by increments of 1 produced the following abstention rates in five data sets: 18.9%, 46.3%, 70%, 90.2%, 96.2%. The Abstention model is fit on each of these full simulated data sets, whereas the **pscl** model is fit on a version of each data set with the abstention observations dropped.[3]

Figure 2.3 through Figure 2.7 display the results of these simulation tests. In each plot the

[3]Simulated groups which never signal support or opposition are removed from both versions of the data so that each model is working with the same set of groups.

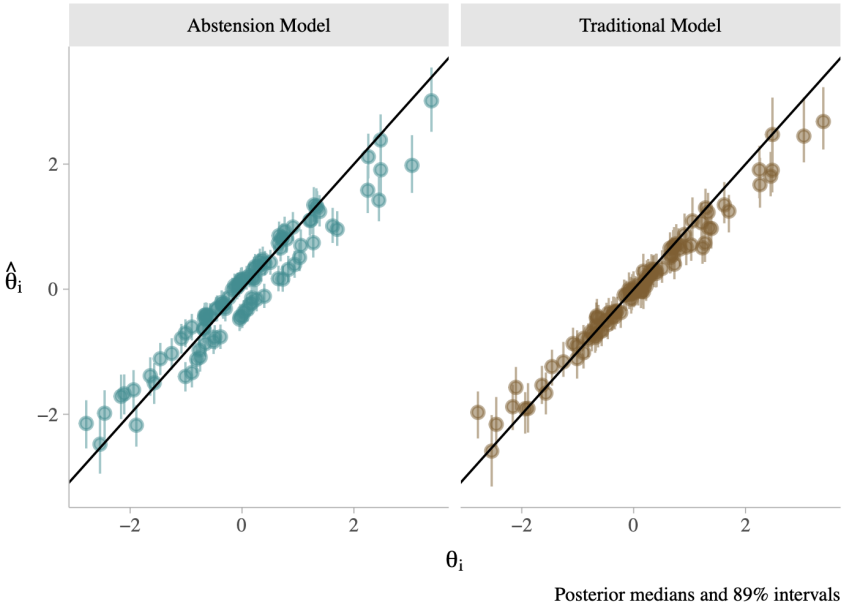

Figure 2.3: Simulation Results with 18.9% Missing Data

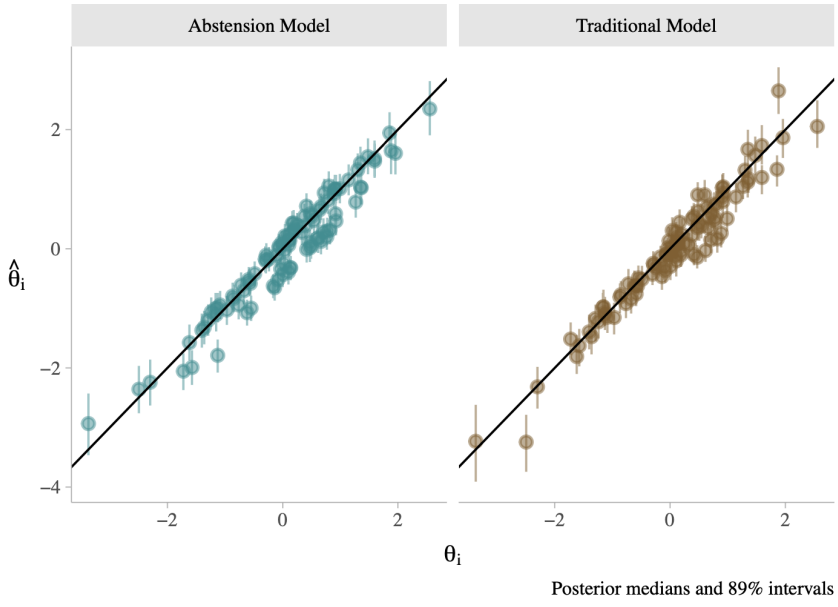

Figure 2.4: Simulation Results with 46.3% Missing Data

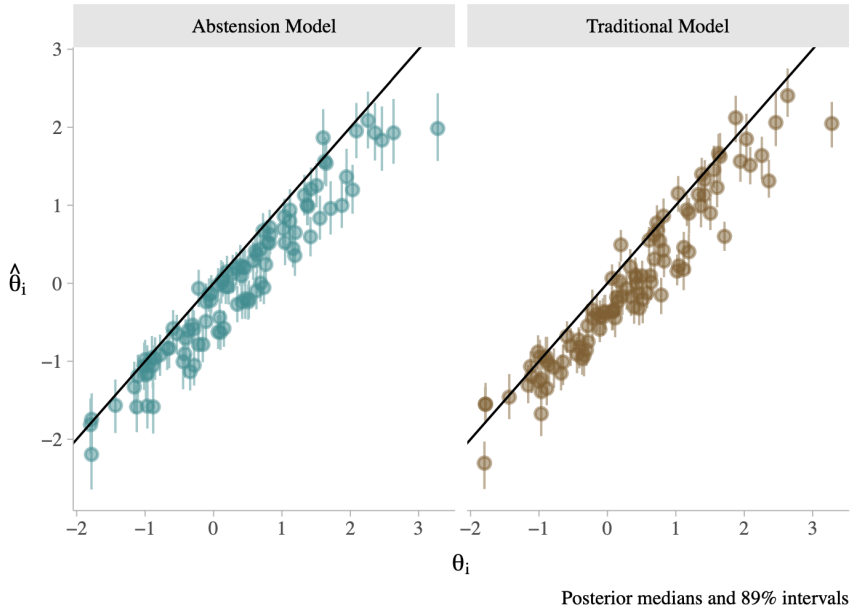

Figure 2.5: Simulation Results with 70.0% Missing Data

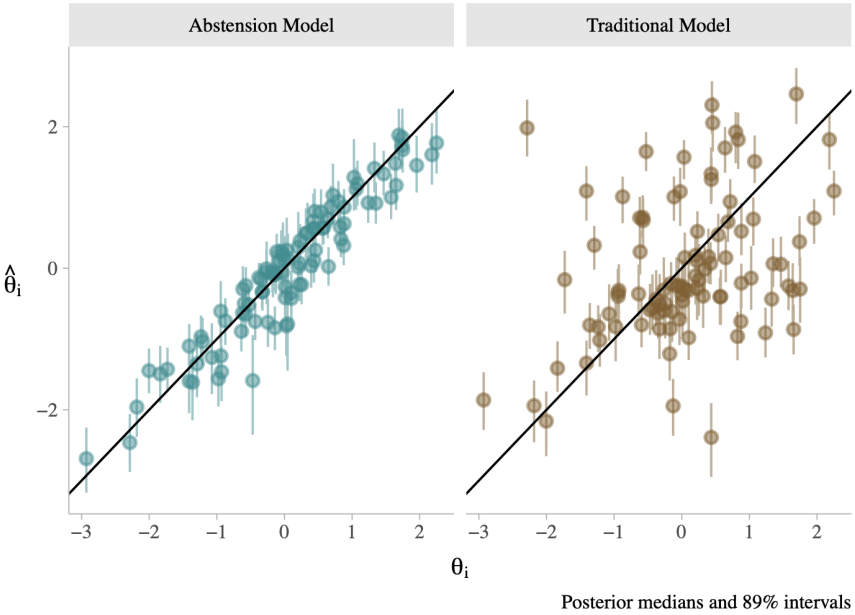

Figure 2.6: Simulation Results with 90.2% Missing Data

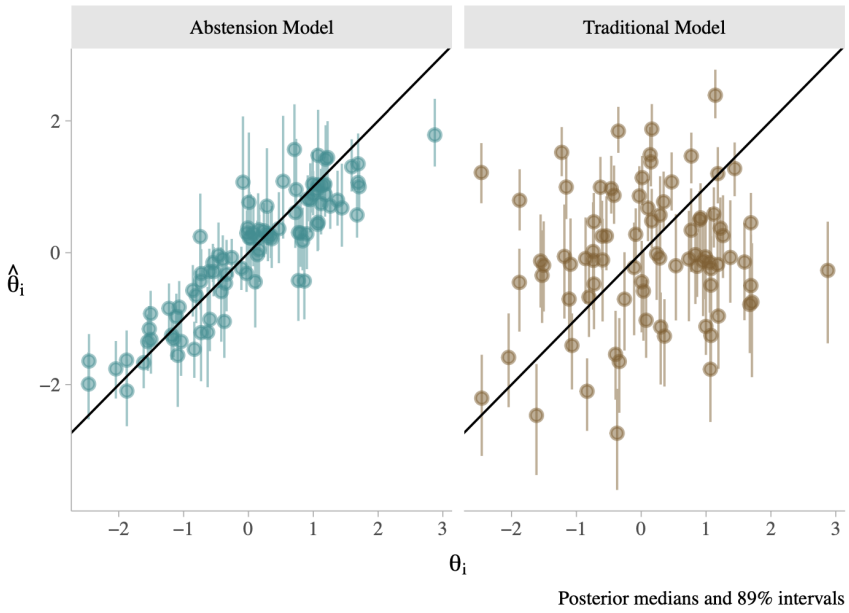

Figure 2.7: Simulation Results with 96.2% Missing Data

true interest group ideal points, θ_i are displayed along the x-axis and the ideal points estimated from each model, $\hat{\theta}_i$ are displayed along the y-axis. Each model's performance can be judged by how closely the estimated ideal points track with the true ideal points along the solid diagonal identity line. The 89% posterior intervals around the median ideal points provide a sense for how precise the estimates are.

Simulations with abstention rates below 75% (Figure 2.3, Figure 2.4, and Figure 2.5) show both models performing well. But once the abstention rate starts to exceed 90% in Figure 2.6, the **pscl** model begins to deteriorate. We see a decoupling of the estimates from their true values throughout the range of ideal points. Things really fall apart for the **pscl** model in Figure 2.7 with 96.2% abstentions. The estimated ideal points are all over the place and fail to discernibly correlate with the true ideal points. As we increase the percentage of abstentions, the Abstention model produces more uncertain estimates as well. Despite this, the Abstention model still generally recovers the true ideal point values in all five simulations.

Although the simulation results show some evidence that the Abstention model is better

equipped to handle missing data compared to the **pscl** model, the former has one major drawback. The HMC implementation of the Abstention model takes an extremely long time to sample compared to the **pscl** model. The combination of modeling all group and bill parameters hierarchically and giving each group its own independent thresholds on the ordered probit scale increase computational demands drastically. Each of the simulations in Figure 2.3 through Figure 2.7 above took roughly two hours to run on a 2019 MacBook Pro. Whereas fitting the **pscl** models took less than five seconds each. So for now the Abstention model is only feasible for small-scale data sets.

2.5 Replication of *Polarized Pluralism: Organizational Preferences and Biases in the American Pressure System*

In *Polarized Pluralism: Organizational Preferences and Biases in the American Pressure System* (2020) Crosson, Furnas, and Lorenz use the Map-Light data on interest groups' lobbying behavior in the US Congress to construct ideal points for interest groups. Rather than exhibiting a strong conservative bias as was previously theorized (Schattschneider 1960), the authors find that the ideological distribution of interest groups is distinctly bi-modal and roughly resembles the ideological make up of members of Congress. As mentioned previously, the authors use the **pscl** model from Clinton, Jackman, and Rivers (2004) to estimate these interest group ideal points (Equation 2.7).

Because of the expensive computation required by the Abstention model, I am only able to use a small portion of their data to replicate their results. Looking only at the 113th Congress (2013 to 2015), I select the 200 interest groups with the highest amount of support/opposition signals. I then select the 600 bills which received the highest amount of support/opposition signals and expanded the data set such that if a group did not signal support or opposition on a particular

bill its observed action was explicitly coded as an abstention.[4] The overall abstention rate in this reduced data is 91.7%

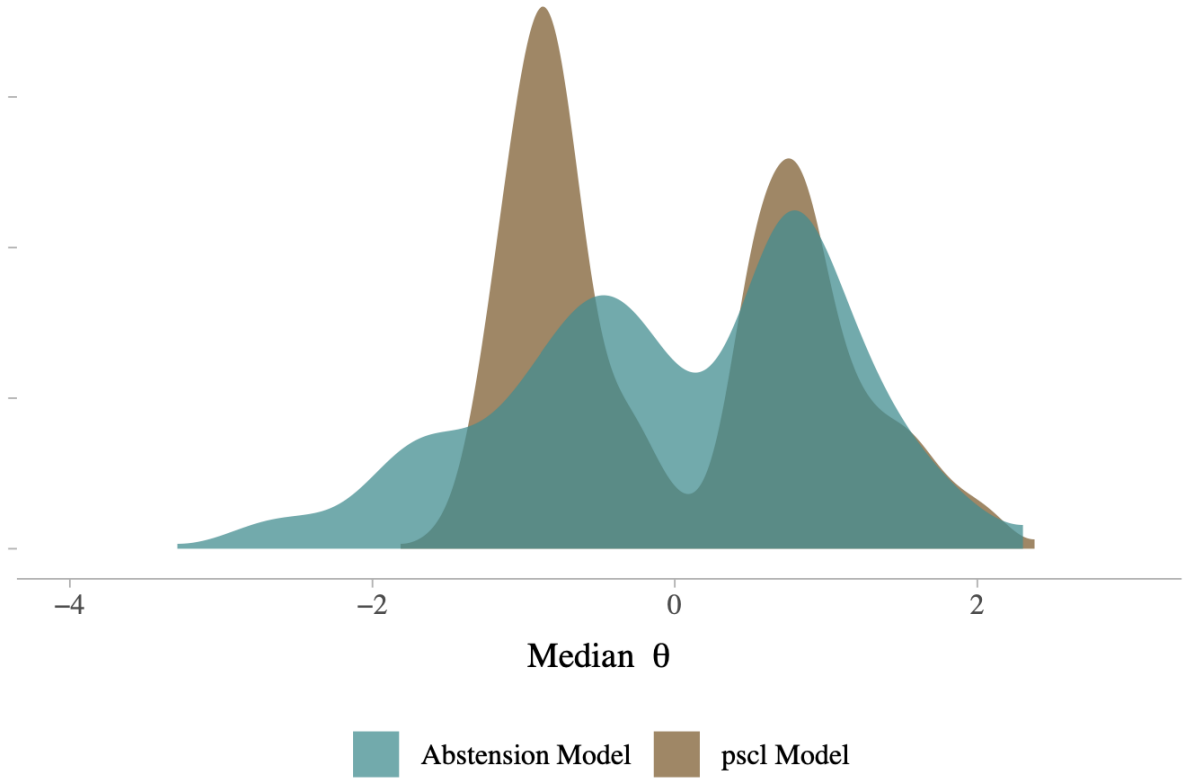

Figure 2.8: Federal Interest Group Ideal Point Distributions

Figure 2.8 shows the distribution of median θ_i estimates across the two models. The **pscl** model produces a distinctly bi-modal distribution of interest group ideology, which, despite the restricted sample, replicates the main findings from Crosson, Furnas, and Lorenz (2020). The Abstention model, on the other hand, shows many more interest groups falling in the middle of the ideological spectrum. There is still some evidence of bi-modality, but the peaks are far less pronounced. As debates about the causes and extent of US political polarization continue, these

[4]A handful of the 200 groups never signaled on any of the 600 bills and were thus dropped from the data. Also bills that received less than five support/oppose signals were dropped from the data. Unlike Crosson, Furnas, and Lorenz, I did not remove bills that received unanimous signals of support or opposition. While such bills might provide no evidence of a group's ideal point using the **pscl** model, the fact that a group choose to signal rather than to abstain provides valuable information in the Abstention model.

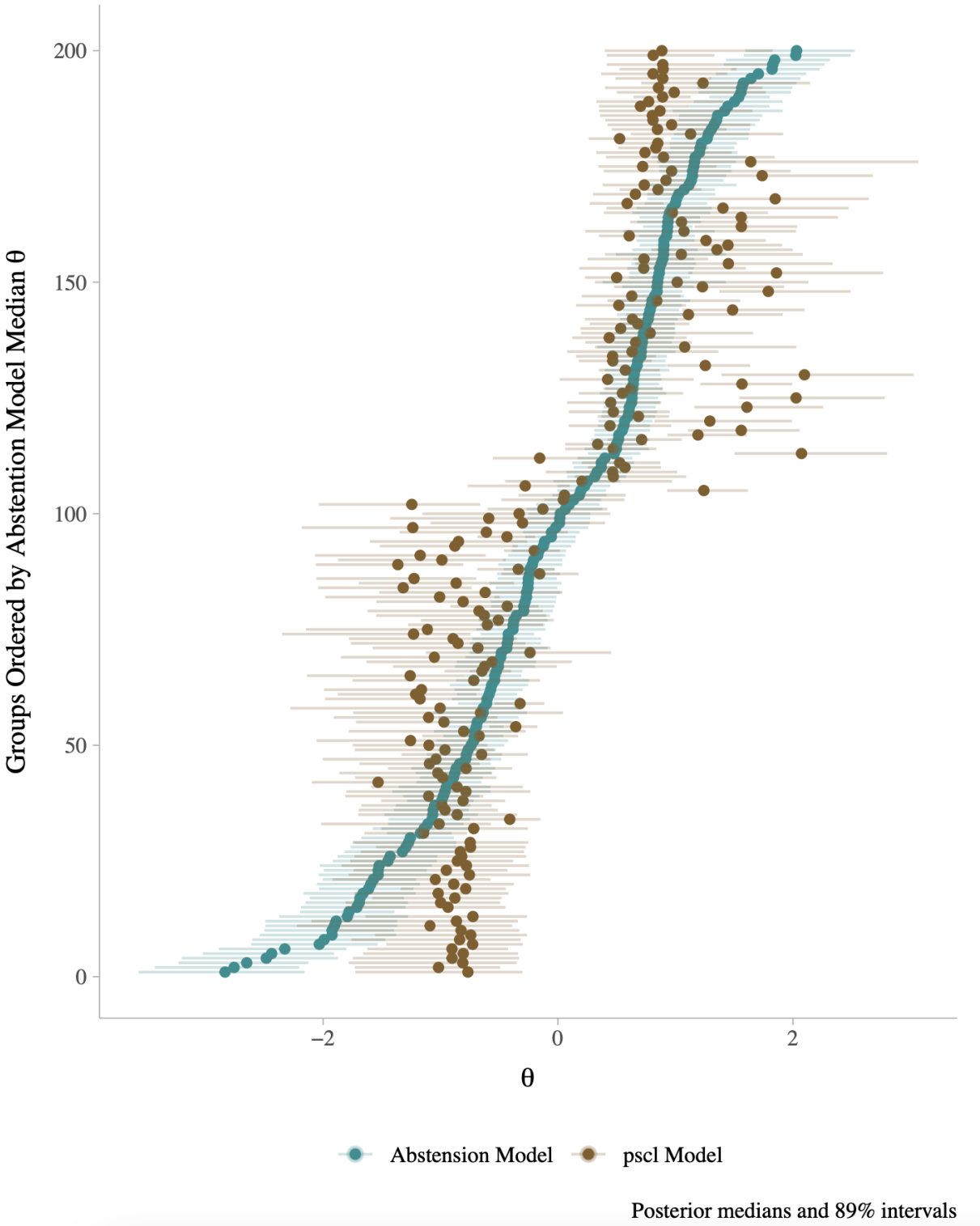

Figure 2.9: Federal Interest Group Ideal Point Comparison

results show how important good measurement is. The Abstention model, which better reflects the underlying data generating process, provides us with a new understanding of the ideological landscape of federal interest groups.

Figure 2.9 displays another comparison of interest group ideology model estimates. Each individual interest group's median θ_i from both models is shown along with the 89% posterior intervals. As in Figure 2.8, we can see the clumping of **pscl** estimates in a liberal cluster and a conservative cluster. We also see that the Abstention model's ideologically centrist results are not an artifact of hierarchical partial pooling towards a central average. Rather than all θ_i's being pulled to the center, the Abstention model produces estimates which are sometimes further liberal or further conservative than the **pscl** estimates.

2.6 Conclusion

This projects demonstrates some of the problems with using traditional ideal point models in contexts with large amounts of missing data. If actors abstain from taking positions on most pieces of legislation, my simulation results show that their ideology will be more accurately measured using the Abstention model developed in this project. When applied to real-world data, the Abstention model and the tradition ideal point model produce diverging views of the ideological landscape of federal interest groups.

2.7 References

Bafumi, Joseph, Andrew Gelman, David K. Park, and Noah Kaplan. 2005. "Practical Issues in Implementing and Understanding Bayesian Ideal Point Estimation." *Political Analysis* 13 (2): 171–87. https://doi.org/10.1093/pan/mpi010.

Bailey, Michael A., Anton Strezhnev, and Erik Voeten. 2017. "Estimating Dynamic State Preferences from United Nations Voting Data." *Journal of Conflict Resolution* 61 (2): 430–56.

https://doi.org/10.1177/0022002715595700.

Barberá, Pablo. 2015. "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23 (1): 76–91. https://doi.org/10.1093/pan/mpu011.

Betancourt, Michael. 2018. "A Conceptual Introduction to Hamiltonian Monte Carlo." *arXiv*, 60.

Bürkner, Paul-Christian. 2020. "Bayesian Item Response Modeling in R with Brms and Stan." *arXiv:1905.09501 [Stat]*, February. http://arxiv.org/abs/1905.09501.

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "*Stan* : A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1). https://doi.org/10.18637/jss.v076.i01.

Caughey, Devin, and Christopher Warshaw. 2015. "Dynamic Estimation of Latent Opinion Using a Hierarchical Group-Level IRT Model." *Political Analysis* 23 (2): 197–211. https://doi.org/10.1093/pan/mpu021.

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98 (2): 355–70. https://doi.org/10.1017/S0003055404001194.

Crosson, Jesse M., Alexander C. Furnas, and Geoffrey M. Lorenz. 2020. "Polarized Pluralism: Organizational Preferences and Biases in the American Pressure System." *American Political Science Review* 114 (4): 1117–37. https://doi.org/10.1017/S0003055420000350.

Fox, Jean-Paul. 2010. *Bayesian Item Response Modeling*. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4419-0742-4.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge ; New York: Cambridge University Press.

Hill, Seth J., and Chris Tausanovitch. 2015. "A Disconnect in Representation? Comparison of Trends in Congressional and Public Polarization." *The Journal of Politics* 77 (4): 1058–75. https://doi.org/10.1086/682398.

Martin, Andrew D., and Kevin M. Quinn. 2002. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999." *Political Analysis* 10 (2): 134–53. https://doi.org/10.1093/pan/10.2.134.

McElreath, Richard. 2020. *Statistical Rethinking: A Bayesian Course with Examples in r and Stan.* 2nd ed. CRC Texts in Statistical Science. Boca Raton: Taylor; Francis, CRC Press.

Rasch, George. 1980. *Probabilistic Models for Some Intelligence and Attainment Tests.* Chicago: University of Chicago Press.

Rosas, Guillermo, Yael Shomer, and Stephen R. Haptonstahl. 2015. "No News Is News: Nonignorable Nonresponse in Roll-Call Data Analysis" *American Journal of Political Science* 59 (2): 511–28. https://doi.org/10.1111/ajps.12148.

Schattschneider, E. E. 1960. *The Semisovereign People: A Realist's View of Democracy in America.* Hinsdale, Ill: Dryden Press.

Slapin, Jonathan B., and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52 (3): 705–22. https://doi.org/10.1111/j.1540-5907.2008.00338.x.

Treier, Shawn, and D. Sunshine Hillygus. 2009. "The Nature of Political Ideology in the Contemporary Electorate." *Public Opinion Quarterly* 73 (4): 679–703. https://doi.org/10.1093/poq/nfp067.

Chapter 3

Mis-Measuring Measurement Model

Measurement Error

3.1 Introduction

Variables of interest in the social sciences are often things we cannot directly observe or measure. Examples include the level of democracy or corruption in a country, or the political ideology of an individual or group. Latent variables such as these must be inferred through indirect processes. One common method is to build statistical models which purport to estimate latent variables using observable input data. I will refer to these as *measurement models*. The outputs of measurement models are then used in subsequent inference procedures to test substantive theories in social science. I will refer to this set of models as *theory-testing models*.

In practice, information about the latent variable is often lost when researchers move from measurement to theory-testing. Measurement models do not simply output a single value for the underlying latent variable. Instead, by virtue of being statistical models, they produce *estimates of uncertainty* for each observation. This is particularly true for Bayesian measurement models, whose

output is the full posterior distribution of values according their relative plausibility—not a single point estimate and standard error as is the case for frequentist models. Failure to propagate this uncertainty from the measurement model into the theory-testing model, as I will show, can lead to mistaken conclusions regarding the underlying research question. And unlike so-called "classical" measurement error, whose attenuation bias is generally well known, the mistakes I investigate can lead to bias in unpredictable directions.

In this paper I demonstrate the problems associated with failing to include measurement model measurement error in theory-testing models, and I develop a method for overcoming these issues. By faithfully incorporating measurement uncertainty into the theory-testing stage of analysis, I show how both attenuation and confounding bias can be mitigated. While the logic of this method can be applied to any measurement model which produces estimates of uncertainty, I focus specifically on continuous-valued latent variables generated from Bayesian measurement models.

3.1.1 The Problem

Theory-testing research which uses estimates from Bayesian measurement models typically reduce the associated posterior distributions down to a single value. In the case of continuous variables, researchers select some statistic of central tendency from each posterior distribution to use in subsequent analyses—such as the mean, median, or mode. This practice necessarily discards information from the full distribution. Figure 3.1 show four hypothetical posterior distributions that may arise from a Bayesian measurement model. Despite all having the same mean of zero, higher order moments such as variance (top-right), skew (bottom-left), and kurtosis (bottom-right) can generate distributions which vary widely.

In Figure 3.1, an estimate from distribution B should be treated as more uncertain than one from distribution A when used to test a theory. Failing to do so can lead to attenuation bias—or

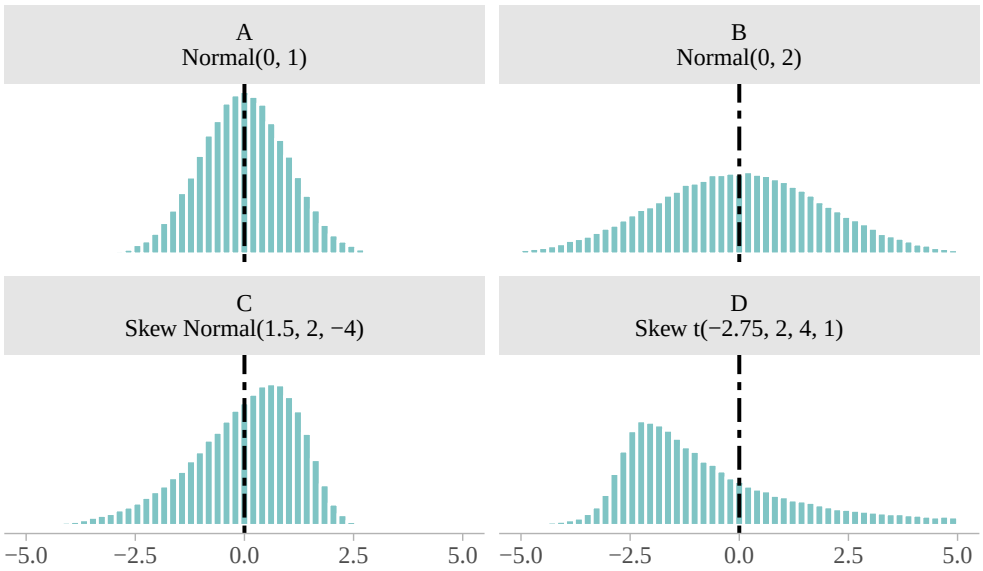

Figure 3.1: Ignoring Measurement Error in Measurement Models

the false conclusion that the latent variable has no association with an outcome when it in fact does. In other words, the method I propose can help increase the statistical power of theory tests by reducing the rate of false negatives. Panels C and D in Figure 3.1 show skewed distributions. Here the danger is that the skewness is caused by a third variable, which *also* causes the outcome of interest in the theory-testing model. This, as I will show, can lead to confounding bias if the skewness of the measurement output is not accounted for.

3.1.2 Method Overview

How can researchers avoid the issues highlighted above? In short, the measurement process and theory-testing procedure should happen simultaneously in a single model. This is handled straightforwardly using the Bayesian statistical framework, which, unlike the frequentist paradigm, does not draw such a sharp distinction between data and parameters (McElreath 2020). We start by specifying the full measurement model, whose posterior distributions for each observation's

value of the latent variable are then used as data in the theory-testing model. The stylized version of this joint model is shown in Equation 3.1, where $g(\cdot)$ is the measurement model which produces posterior estimates of the latent variable, θ_i for each observation based on some training data z. The posterior estimates for θ_i from the measurement model $g(\cdot)$ are then treated as data in the theory-testing model $f(\theta_i)$ using the outcome of interest y.

$$\begin{aligned} y_i &\sim f(\theta_i) \\ z_i &\sim g(\theta_i) \end{aligned} \tag{3.1}$$

There are two practical issues, however, with building a fully-specified joint measurement and theory-testing model. The first is computational. Bayesian statistical software uses notoriously expensive Markov Chain Monte Carlo (MCMC) sampling methods to derive its posterior distributions. Even on their own, measurement models which use MCMC can be extremely slow to sample given these types of models' high-dimensional nature. So attempting to sample from a model which also includes an arbitrarily complex theory-testing model, $f(\cdot)$, in addition to the measurement model may simply be unfeasible given the computing power that the average researcher has access to. The second challenge with the idealized joint model is that it requires researchers to write down a fully-specified measurement model, $g(\cdot)$. Compared to their theory-testing model, applied researchers likely have much less knowledge regarding the intricacies involved in estimating latent variables. Because latent variables have no objective scale, measurement models can often be challenging to fit in practice due to issues of model identification.

The method developed in this paper overcomes the two problems outlined above by simplifying the measurement model step, $z_i \sim g(\cdot)$ in the joint model. Rather than estimating the latent variable from scratch, I take the posterior distributions already provided from previously fitted measurement models and use those as approximations in the full joint model. The measurement

model $g(\cdot)$ becomes a probability distribution function with distributional parameters according to maximum likelihood estimates of the posterior. So if the posterior distribution of the measurement model appears normal, we would use $\theta_{obs[i]} \sim N(\theta_i, \sigma^2_{\theta[i]})$. The values $\theta_{obs[i]}$ and $\sigma^2_{\theta[i]}$ are estimated from the measurement model's posterior distribution, which allows the true, unobserved, value of the latent variable θ_i to be estimated for each observation. If the posterior distributions from the measurement model appear skewed, or have thicker tails than a normal distribution, the distributional parameters for these distributions can be used instead. These simplifications faithfully propagate the uncertainty in the outputs of the measurement model to the theory-testing model, while also being computationally tractable and straightforward to implement.

3.1.3 Motivating Example - Bayesian Models of Ideology

One of the most common measurement models in political science is the Bayesian Item-Response Theory (IRT) model used to measure the ideological leanings of political actors (Clinton, Jackman, and Rivers 2004; Bafumi et al. 2005). These models assume that political ideology is a latent characteristic that lies on a single left-right dimension. Observed actions, such as voting on legislation, are used as training data (z in Equation 3.1) to produce a posterior distribution of continuous values for each actor (e.g. member of Congress) on this left-right scale.

Let's say we want to estimate the effect of legislator ideology, θ on some outcome, y. Using the format developed in Equation 3.1, Equation 2.1 shows an example joint measurement and theory-testing model to answer this question using ideology estimates from an IRT model. The measurement model (bottom) predicts whether a legislator votes yes or no on a piece of legislation, z_j using the traditional 2-parameter IRT equation $\Phi(\gamma_j\theta_i + \xi_j)$. Estimates of the parameters θ_i from this model, are then treated like data in the linear regression theory-testing model (top) to estimate β_1 —the coefficient of substantive interest.

$$y_i \sim \text{Normal}(\beta_0 + \beta_1 \theta_i, \sigma^2)$$

$$z_{ij} \sim \text{Bernoulli}[\Phi(\gamma_j \theta_i + \xi_j)]$$

As mentioned previously, however, estimating Equation 2.1 would not generally be feasible due to computational constraints. Instead, I propose that the IRT ideology measurement model be fit beforehand. Then, for each legislator's posterior distribution of θ, the values $\theta_{obs[i]}$ and $\sigma^2_{\theta[i]}$ are calculated via maximum likelihood. These values are in turn used as data in the simplified measurement model in Equation 3.2 in order to estimate the latent θ_i for each observation.

$$\begin{aligned} y_i &\sim \text{Normal}(\beta_0 + \beta_1 \theta_i, \sigma^2) \\ \theta_{obs[i]} &\sim \text{Normal}(\theta_i, \sigma^2_{\theta[i]}) \end{aligned} \tag{3.2}$$

If the posterior estimates of θ from the IRT measurement model are truly normally distributed for each legislator, then Equation 2.1 and Equation 3.2 are essentially equivalent—thereby properly incorporating the measurement model measurement error in the theory-testing model. If, however, the IRT model produces posterior distributions that are not normal, then this simplification step could be throwing out important information. For this reason I extend the model to include a skewness parameter later in this project.

3.2 Measurement Error Models

In this section I provide additional motivation for why researchers should care about measurement model uncertainty when using latent variables in their theory-testing models. Usually, theory-testing models are used to answer some causal question: *what is the effect of X on Y*? The observed relationship between X and Y is often confounded by other variables in the system exerting causal influence. Theory-testing models, therefore, need to condition on these confounding variables

in order to get an unbiased estimate of the causal effect of interest. While this general method for theory testing is well-understood in the social sciences (Rubin 1974; Morgan and Winship 2007), it is less common to apply the same causal logic to measurement. Failing to do so, I argue, can lead to erroneous substantive conclusions.

The causal graph framework (Pearl 2000) is a useful way to demonstrate this argument. Causal graphs are heuristic tools which map out causal relationships between variables in a particular system. Each node represents a variable, and the directed edges between nodes represent hypothesized causal impacts of one variable on another. These directed-acyclic-graphs (DAGs) are useful because they allow us to determine the set of variables we need to condition on in order to get an unbiased estimate of the effect of our primary independent variable on the dependent variable. This set of confounders is defined by the variables which are needed to close every "backdoor" path between the primary independent and dependent variables.[1]

Using the logic of causal graphs, I will discuss two types of measurement model measurement error, and how the joint measurement theory-testing procedure laid out in Equation 3.1 helps fix them. First I consider random, or classical, measurement error. In this case the joint model will (in most cases) provide researchers with extra statistical power to test their theory by mitigating issues of attenuation bias and helping avoid false negatives. Then we will look at non-random measurement error scenarios, in which the measurement model error introduces confounding in the theory-testing model. I will show how the joint model from Equation 3.1 ameliorates this confounding bias. For both types of measurement error, I use simulation studies to demonstrate how effective each modeling approach is at recovering known parameter values from the theory-testing model.

[1] See Cinelli, Forney, and Pearl (2020) for a more complete introduction to deconfounding using DAGs.

3.2.1 Measurement Error Attenuation Bias

If the posterior distributions for the latent parameters θ_i from the measurement model follow a normal distribution, this is a form of classical measurement error. Here we do not assume that there is some relationship between the measurement error and the outcome of interest, rather, the errors are simply random "noise" in the measurement estimates. Classical measurement error leads to attenuation bias: a reduction of the main effect size in the theory-testing model towards zero. Thus, the wider the posterior distribution is for $\theta_{obs[i]}$, the more likely we are to make a false negative error in our theory-testing model.

Let's return to the Bayesian IRT measurement model used to estimate political ideology for legislators. Figure 3.2 shows the causal process which produces these ideal point estimates. The observed measurements of ideology, θ_{obs} come from an unobserved latent variable plus some measurement error. At least two variables can affect the measurement error, e^{θ} in Figure 3.2. First, the true ideal point of the group, θ influences the amount of measurement error for estimates of the group's ideology because, as we move further from the ideological center of the scale, uncertainty increases. Figure 3.3 shows a stylized example of what the distribution of posterior estimates from a Bayesian IRT model can look like. Groups further from the center have increasingly wider ideal point posterior distributions. The second variable affecting e^{θ} is participation, P —how much a particular legislator has taken positions on bills. Legislators that signal more positions on bills will have smaller levels of measurement error compared to those that signal fewer positions because we have more information on their true ideological preferences.

There is also likely unobserved measurement error in these types of models. IRT models assume that each legislator's decision to vote on bill is influenced solely by their inherent ideology, rather than by strategic concerns. A violation of this assumption, therefore, will produce biased estimates of θ_{obs}. There are also computational issues with fitting IRT models which can make

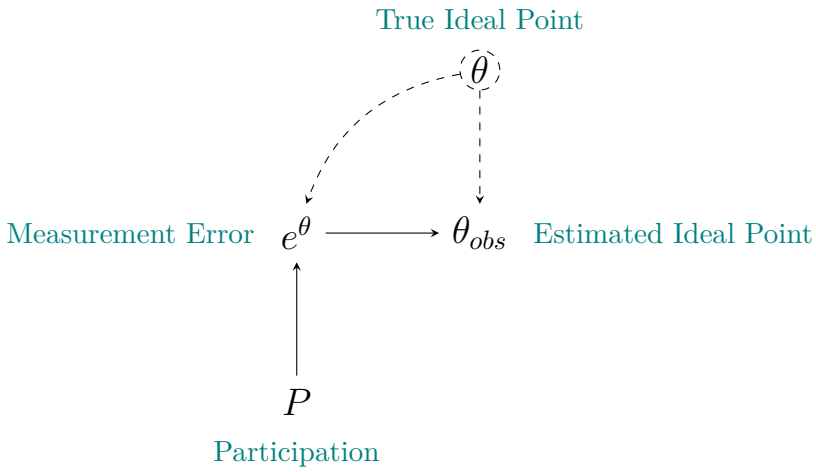

Figure 3.2: IRT Measurement Model

the posterior estimates untrustworthy. For the purposes of this illustration, however, I will assume that the posterior distribution, e^{θ} for θ_{obs} contains all relevant information about the measurement error for the true ideology θ.

Now let's expand Figure 3.2 into a theory-testing model with Figure 3.4. While there may exist some backdoor paths through e^{θ} and P in this hypothetical theory-testing model, I will assume that the outcome of interest, Y is unaffected by anything other than the direct causal effect $\theta_{obs} \longrightarrow Y$.[2] The purpose of this simplification is to highlight the consequences of random measurement error during parameter estimation of a theory-testing model.

3.2.1.1 Simulation Study: Attenuation Bias

Using the generative causal model in Figure 3.4, we can simulate data with a known parameter for the effect, β_1 of legislator ideology on the outcome Y. Then we fit two linear regression models to estimate this parameter. Equation 3.3 is the naive theory-testing model where $\theta_{obs[i]}$ corresponds to a legislator's mean ideal point estimate from the Bayesian IRT measurement model. This is in

[2]In principle, we could close any backdoor paths through P by conditioning on it directly because the level of group participation is directly observed.

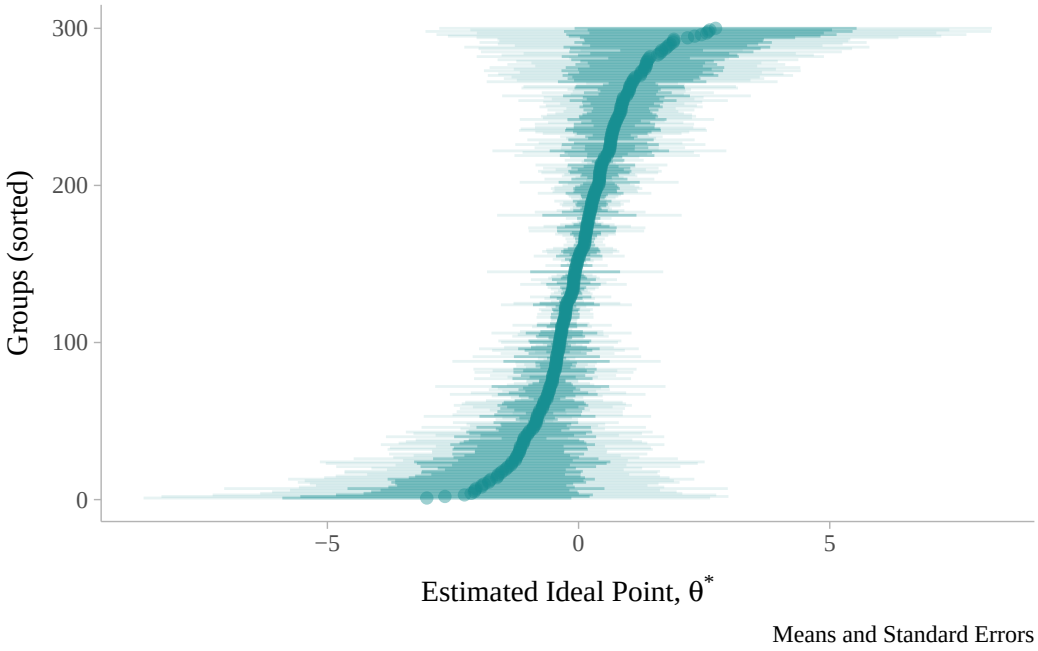

Figure 3.3: IRT Model Posterior Distributions

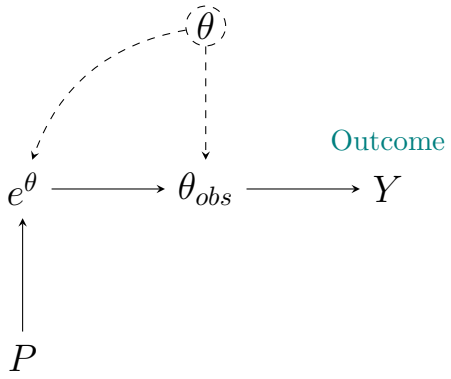

Figure 3.4: IRT Measurement Model in Hypothetical Theory-Testing Model

contrast to the joint measurement theory-testing model in Equation 3.4, which models $\theta_{obs[i]}$ as an outcome of the true ideology θ_i (an unobserved parameter for each observation) and $\sigma^2_{\theta[i]}$ which is the observed variance of the posterior distribution for each groups' ideal point. The parameters θ_i are also given hyperpriors π and τ for location and scale respectively. The estimates of θ_i from this simplified measurement model are then used in the linear model which predicts the outcome y. These, and all other models in this paper, are written in the probabilistic programming language Stan and fit using its Hamiltonian Monte Carlo sampler (Carpenter et al. 2017).

$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma^2) \\
\mu_i &= \beta_0 + \beta_1 \theta_{obs[i]} \\
\beta_0, \beta_1 &\sim \text{Normal}(0, 2) \\
\sigma &\sim \text{Half Student t}(3, 0, 2)
\end{aligned}
\tag{3.3}
$$

$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma^2) \\
\mu_i &= \beta_0 + \beta_1 \theta_i \\
\theta_{obs[i]} &\sim \text{Normal}(\theta_i, \sigma^2_{\theta[i]}) \\
\theta_i &\sim \text{Normal}(\pi, \tau) \\
\beta_0, \beta_1 &\sim \text{Normal}(0, 2) \\
\sigma &\sim \text{Half Student t}(3, 0, 2) \\
\tau &\sim \text{Half Student t}(3, 0, 2) \\
\pi &\sim \text{Normal}(0, 1)
\end{aligned}
\tag{3.4}
$$

Figure 3.5 shows how well each model recovers the true parameter value for β_1: the effect of the true ideal point on the simulated outcome. Each model was fit 40 times across a range of increasing values for $\sigma^2_{\theta[i]}$, thereby increasing random error in the independent variable (shown on the horizontal axis as the correlation between the simulated true ideal point and mean measurement error value approach zero). The mean, and 89% credible interval posterior estimates of each model's β_1 parameter are plotted with a loess fit. With little-to-no measurement error (left side of the

graph), both models reliably recover the true β_1 value of 1. But as the random measurement error increases, the β_1 estimates from the naive model from Equation 3.3 rapidly attenuate towards zero. This is in contrast to the estimates from the joint measurement theory-testing model in Equation 3.4 which remain much closer to the true β_1 value even after there is essentially zero correlation between the true ideal points and means from the ideal points with measurement error.

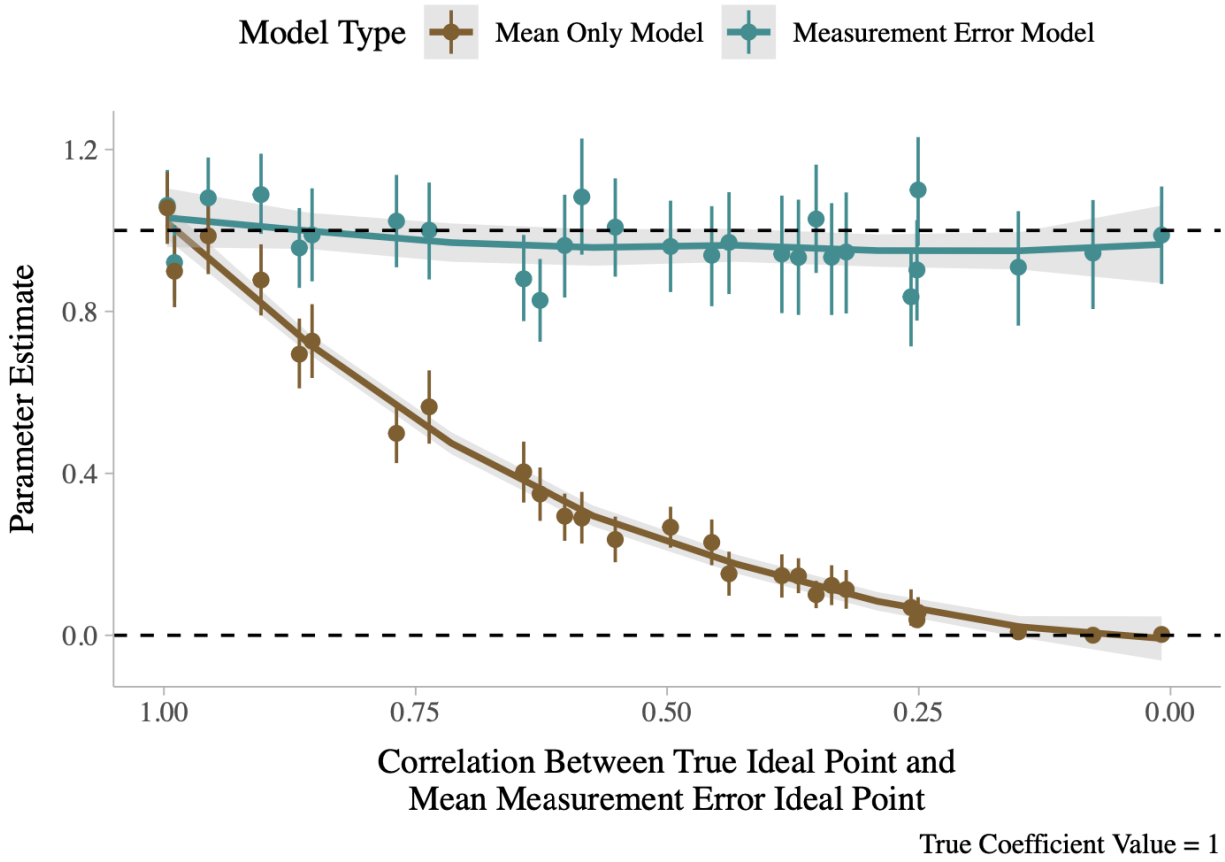

Figure 3.5: Parameter Recovery as Measurement Error Increases

In addition to showing how the joint measurement theory-testing method can help avoid attenuation bias, the results from Figure 3.5 show how this method more faithfully propagates measurement uncertainty into the theory-testing analysis. For each simulated model, the 89% credible intervals are wider for the measurement error model compared to the naive mean values model. These results demonstrate a dangerous combination of both increased bias, and increased

certainty, in the theory-testing model if researchers neglect to incorporate the measurement model measurement error.

3.2.2 Measurement Error Confounding Bias

The previous discussion highlighted how failing to account for random measurement error could lead to attenuation bias. The second problem I address in this project is measurement error-induced confounding bias. This general issue is also known as nonrandom, or unignorable, measurement error (Blalock 1970). In the political science methodology literature, methods such as multiple imputation (Blackwell, Honaker, and King 2017) and sensitivity analysis (Gallop and Weschle 2019; Imai and Yamamoto 2010) have been developed to deal with nonrandom measurement error. My method is another way of dealing with nonrandom measurement error, but in the context where the measurement error is known and comes from the output of some measurement model.

Building off from the DAG in Figure 3.4, let's now consider the hypothetical causal graph shown in Figure 3.6. As before, the main causal effect of interest is represented by the path $\theta_{obs} \longrightarrow Y$. In order to get an unbiased estimate of this causal effect we need to close all backdoor paths leading from θ_{obs} to Y, which in this case, flows through the unobserved variable U. This confounding variable represents anything that is a common cause of both the IRT model measurement error and the outcome of interest.

For the theory-testing model in Figure 3.6 it may be possible to directly condition on some variables in U in order to obtain an unbiased estimate of $\theta_{obs} \longrightarrow Y$. But with something as multi-faceted as political ideology, there is always some risk of residual confounding. My proposed method of building a joint measurement and theory-testing model fixes this issue by obviating the need to deal directly with U at all. This is because the measurement error, e^{θ} in Figure 3.6 is part of the backdoor path from θ_{obs} to Y. Therefore when we explicitly incorporate e^{θ} into a

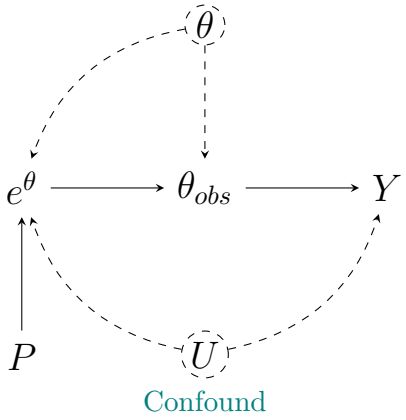

Figure 3.6: IRT Measurement Model in Hypothetical Theory-Testing Model with Confounding

model estimating $\theta_{obs} \longrightarrow Y$ we can obtain an unbiased (or at least less-biased, given unobserved measurement error) estimate of the causal effect of ideology on the theoretical outcome of interest.

3.2.2.1 Simulation Study: Confounding Bias

To demonstrate how the joint measurement theory-testing method I propose handles non-ignorable measurement error, I carry out a simulation study similar to that in the previous section. Given the causal graph Figure 3.6, I generate data such that Y is only a function of the unobserved confound U. The true effect of $\theta \longrightarrow Y$ is zero in the simulation. θ_{obs} is drawn from a Skew-Normal distribution whose location parameter, ξ equals the true θ value, but whose skew parameter, α is a function of the confound U. This corresponds to the $U \longrightarrow e^{\theta}$ path in Figure 3.6.

$$\frac{2}{\omega\sqrt{2\pi}} e^{-\frac{(x-\xi)^2}{2\omega^2}} \int_{-\infty}^{\alpha\left(\frac{x-\xi}{\omega}\right)} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt \tag{3.5}$$

Equation 3.5 is the probability density function for the Skew-Normal distribution. The distribution is a convolution of the Normal distribution and Half Normal (or folded Normal) distribution and has three distributional parameters: ξ for location, ω for scale, and α for skew. When $\alpha = 0$ the distribution collapses to the Normal distribution. Unfortunately there is not a

closed form solution for finding the maximum likelihood estimates of the distributional parameters so numerical methods need to be used. In practice this leads to estimation instability as α approaches zero (Azzalini and Capitanio 2014). Because of this, a choice must be made ahead of time about whether to use the Skew-Normal or regular Normal distribution for the measurement model.

After the data are generated, I fit two models: the ordinary linear regression using only θ_{obs} values as were used previously in the attenuation bias example (Equation 3.3), and a modified version of Equation 3.4 which substitutes the Normal distribution for the Skew Normal distribution (Equation 3.6). As before, the key parameter of theoretical interest is β_1 —which, according to the simulated data, should equal zero if unconfounded.

$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma^2) \\
\mu_i &= \beta_0 + \beta_1 \theta_i \\
\theta_{obs[i]} &\sim \text{Skew Normal}(\theta_i, \omega_{\theta[i]}, \alpha_{\theta[i]}) \\
\theta_i &\sim \text{Normal}(\pi, \tau) \\
\beta_0, \beta_1 &\sim \text{Normal}(0, 2) \\
\sigma &\sim \text{Half Student t}(3, 0, 2) \\
\tau &\sim \text{Half Student t}(3, 0, 2) \\
\pi &\sim \text{Normal}(0, 1)
\end{aligned}
\tag{3.6}
$$

Figure 3.7 shows how well each model estimates β_1. As expected, the bottom model from Equation 3.3 using only the θ_{obs} values produces a biased estimate of β_1. The open backdoor path $\theta_{obs} \longleftarrow e^{\theta} \longleftarrow U \longrightarrow Y$ from Figure 3.6 confounds the causal effect $\theta \longrightarrow Y$ if the latent variable was measured perfectly. In contrast, the top model from Equation 3.6 accurately reports a β_1 coefficient value of zero. This is because the measurement error e^{θ} is included in the model in the form of $\alpha_{\theta[i]}$ for each observation.[3] As in the attenuation bias example, the naive model's posterior for β_1 is also significantly narrower compared to the measurement error model. The measurement

[3]The scale parameter $\omega_{\theta[i]}$ is also included in Equation 3.6 to help avoid attenuation bias.

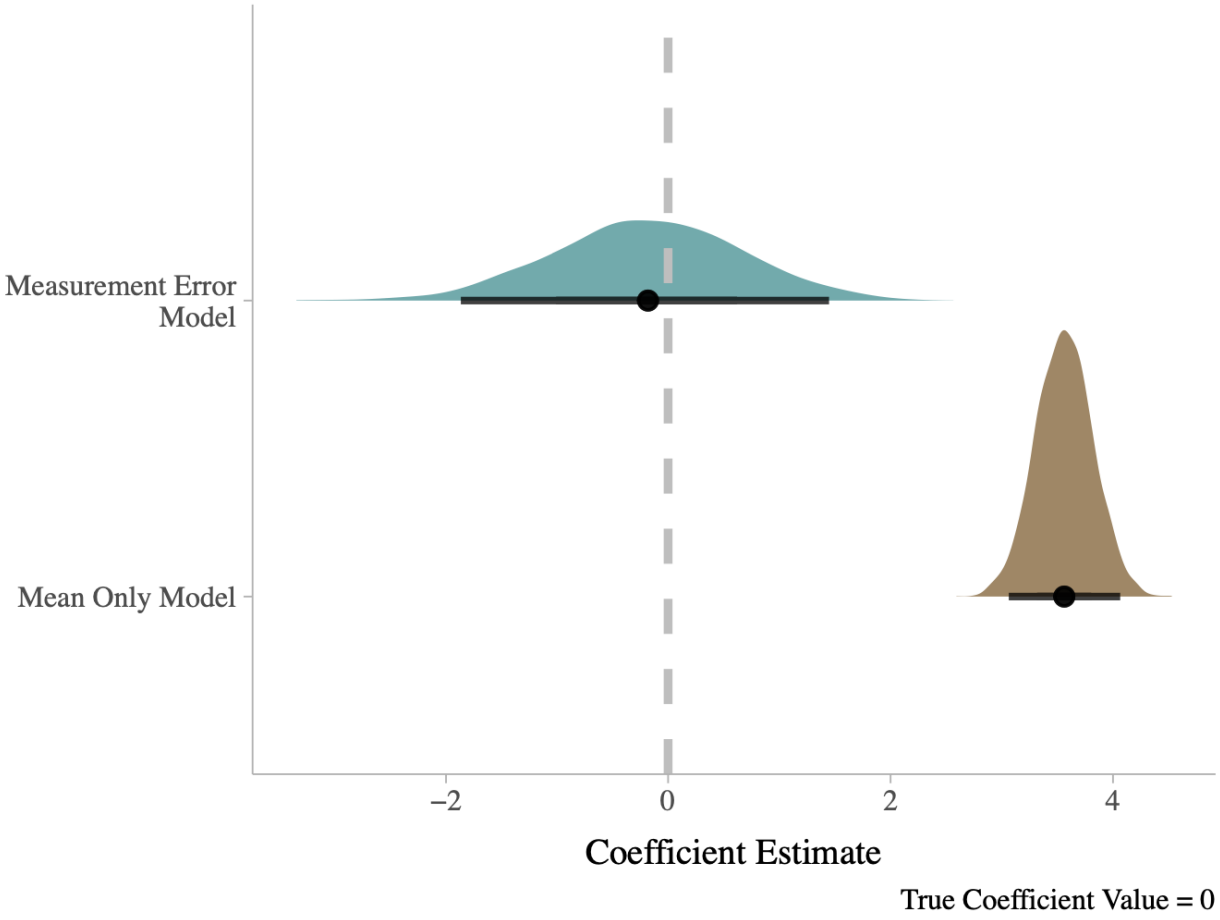

Figure 3.7: Parameter Recovery Under Confounding

error model is faithfully propagating the uncertainty from the measurement process into the final theory-testing analysis.

3.3 Case Study: Candidate Extremism and Electoral Success

The analysis in the previous section showed how a joint measurement theory-testing model can help avoid both attenuation and confounding bias stemming from measurement error. Simulation studies are powerful because we have complete control over the data generating process, and therefore know how well each model can recover the parameters which generated the outcome Y. The downside of simulation studies, however, is that they greatly simplify the complex social phenomena they aim to represent. With this in mind, I apply the proposed method to a real world example with ideology measurements from an IRT model.

Are ideologically extreme US House incumbents punished electorally? According to the widely cited Canes-Wrone, Brady, and Cogan (2002), the answer is yes. While these authors use interest group scores to measure ideology, they discuss how their results remain the same when using DW-NOMINATE scores of ideology. DW-NOMINATE is a measurement model similar to IRT, but whose uncertainty measurements are only given in the form of bootstrapped standard errors (Carroll et al. 2009). These standard error estimates could potentially be used as $\sigma^2_{\theta[i]}$ in the classical measurement error model $\theta_{obs[i]} \sim N(\theta_i, \sigma^2_{\theta[i]})$ since Bayesian posterior standard deviations and frequentist standard errors share similar qualities. But standard errors do not provide any information about error skewness. This makes them unsuitable for replication using the Bayesian method proposed here, so I instead re-estimate all US Representatives' ideology using the popular IRT model from the **pscl** R package.[4]

Because this research question relies on observational data, it is important to sketch out a

[4]https://github.com/atahk/pscl

causal graph of the system in order to understand how to isolate the main effect. Figure 3.8 is one plausible causal graph of this system. The main effect of interest is $\theta_{obs} \longrightarrow Y$: measured candidate ideology's effect on vote share in the general election. As in the simulation study examples, θ_{obs} is a function of the candidate's true ideology, θ and error e^{θ}. The confound U between e^{θ} and Y could represent a number of variables. Perhaps incumbent candidates who log-roll votes in Congress are seen as more, or less, effective representatives—thus influencing their future vote shares. But log-rolling would mean that the candidate's floor votes do not always represent their true ideology, thereby increasing the measurement error in the IRT model for that candidate.

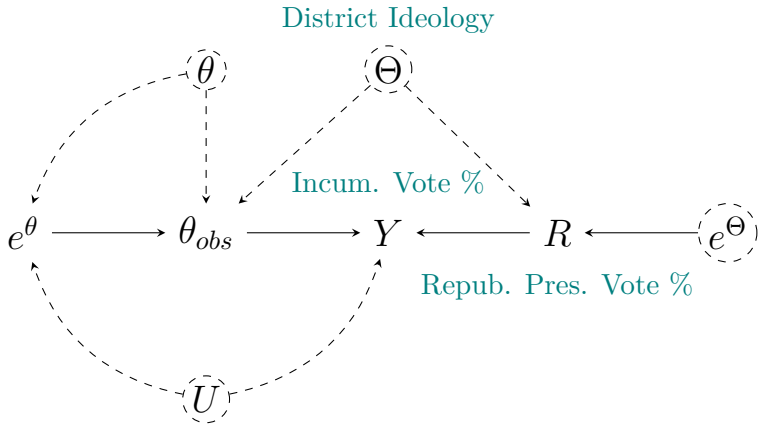

Figure 3.8: Isolating the Effect of Ideology on General Election Vote Share

The other key confound in this causal system is district ideology, Θ. Canes-Wrone, Brady, and Cogan (2002), and others who have asked similar research questions, are interested in whether ideologically extreme candidates, *relative to their district*, are punished electorally. A candidate who is considered extreme in one district might be considered moderate in another district. The effect $\Theta \longrightarrow \theta_{obs}$ represents this selection process, whereby candidates choose to run in districts with which they are already ideologically aligned. Unfortunately, district ideology, like candidate ideology, is not directly observed. Instead it is common practice to use an observed variable like district presidential vote share, R as a proxy for district ideology. Presidential vote share is in no

way a perfect measure of district ideology, hence the inclusion of e^{Θ} in Figure 3.8.

The data for this analysis come from a variety of sources. Vote View[5] provide congressional votes for each year used to estimate the IRT ideology models for House representatives. Then, using the maximum likelihood estimator in the R package **sn**[6] I calculate the Normal and Skew-Normal distributional parameters from the IRT posterior estimates. These data were then merged with candidate information from Volden and Wiseman (2014). which are in turn merged with presidential vote share data from POLIDATA.[7] The final unit of observation in the data set is candidate-election, with candidate ideology lagged one Congress session so as to reflect the fact that legislator's district electorates should be responding to their previous actions in the House.

I also split the data by party to make the main effects easier to interpret. Lower values from the IRT ideology measurement model correspond to more left-wing candidates, whereas higher values correspond to more right-wing candidates. So, for Democrats we interpret a *negative* relationship between ideology and vote share as district electorates favoring extremist candidates, whereas a *positive* relationship for Republican candidates would mean district electorates favor more extremists.

$$
\begin{aligned}
\text{VotePct}_i &\sim \text{Normal}(\mu_i, \sigma^2) \\
\mu_i &= \beta_0 + \beta_1 \theta_{obs[i]} + \beta_2 \text{R}_i \\
\beta_0 &\sim \text{Normal}(50, 5) \\
\beta_1, \beta_2 &\sim \text{Normal}(0, 2) \\
\sigma &\sim \text{Half Student t}(3, 0, 2)
\end{aligned}
\tag{3.7}
$$

Equation 3.7 shows the simple linear regression without using a measurement error model for candidate ideology. Equation 3.8 models the measurement error using the Skew-Normal distribution

[5]Lewis, Jeffrey B., Keith Poole, Howard Rosenthal, Adam Boche, Aaron Rudkin, and Luke Sonnet (2021). *Voteview: Congressional Roll-Call Votes Database*. https://voteview.com/

[6]https://cran.r-project.org/web/packages/sn/index.html

[7]Bensen, Clark H., "Presidential Results by Congressional District (PRCD)", POLIDATA ® Demographic & Political Guides [Distributor]

as discussed in the previous section. This attempts to handle both attenuation bias and confounding from U in Figure 3.8. Both models control for district Republican presidential vote share, R to close the backdoor path $\theta_{obs} \longleftarrow \Theta \longrightarrow R \longrightarrow Y$. I also fit a random measurement error model using a Normal distribution in place of the Skew-Normal distribution in case there is no confounding from U, but omit the model equation for brevity.

$$
\begin{aligned}
\text{VotePct}_i &\sim \text{Normal}(\mu_i, \sigma^2) \\
\mu_i &= \beta_0 + \beta_1 \theta_i + \beta_2 \text{R}_i \\
\theta_{obs[i]} &\sim \text{Skew Normal}(\theta_i, \omega_{\theta[i]}, \alpha_{\theta[i]}) \\
\theta_i &\sim \text{Normal}(\pi, \tau) \\
\beta_0 &\sim \text{Normal}(50, 5) \\
\beta_1, \beta_2 &\sim \text{Normal}(0, 2) \\
\sigma &\sim \text{Half Student t}(3, 0, 2) \\
\tau &\sim \text{Half Student t}(3, 0, 2) \\
\pi &\sim \text{Normal}(0, 2)
\end{aligned}
\tag{3.8}
$$

Figure 3.9 shows the posterior distributions for β_1 in the above models (which represents $\theta_{obs} \longrightarrow Y$: measured candidate ideology's effect on vote share in the general election). All the results point towards both Democrats and Republicans favoring more ideologically extreme candidates. The negative coefficient values in the top panel for Democrats suggest that more left-wing candidates perform better, and the positive coefficients in the bottom panel suggest that more right-wing Republican candidates perform better. In this example, the different models only disagree about the magnitudes and uncertainty of these effects. The mean-only models have coefficient values closer to zero compared to the measurement error models. This points to some amount of attenuation bias taking place. Furthermore, the skew-measurement error models show larger electoral effects for extremist candidates, which could mean that there is some residual confounding in Figure 3.8 that the other models are not taking into account.

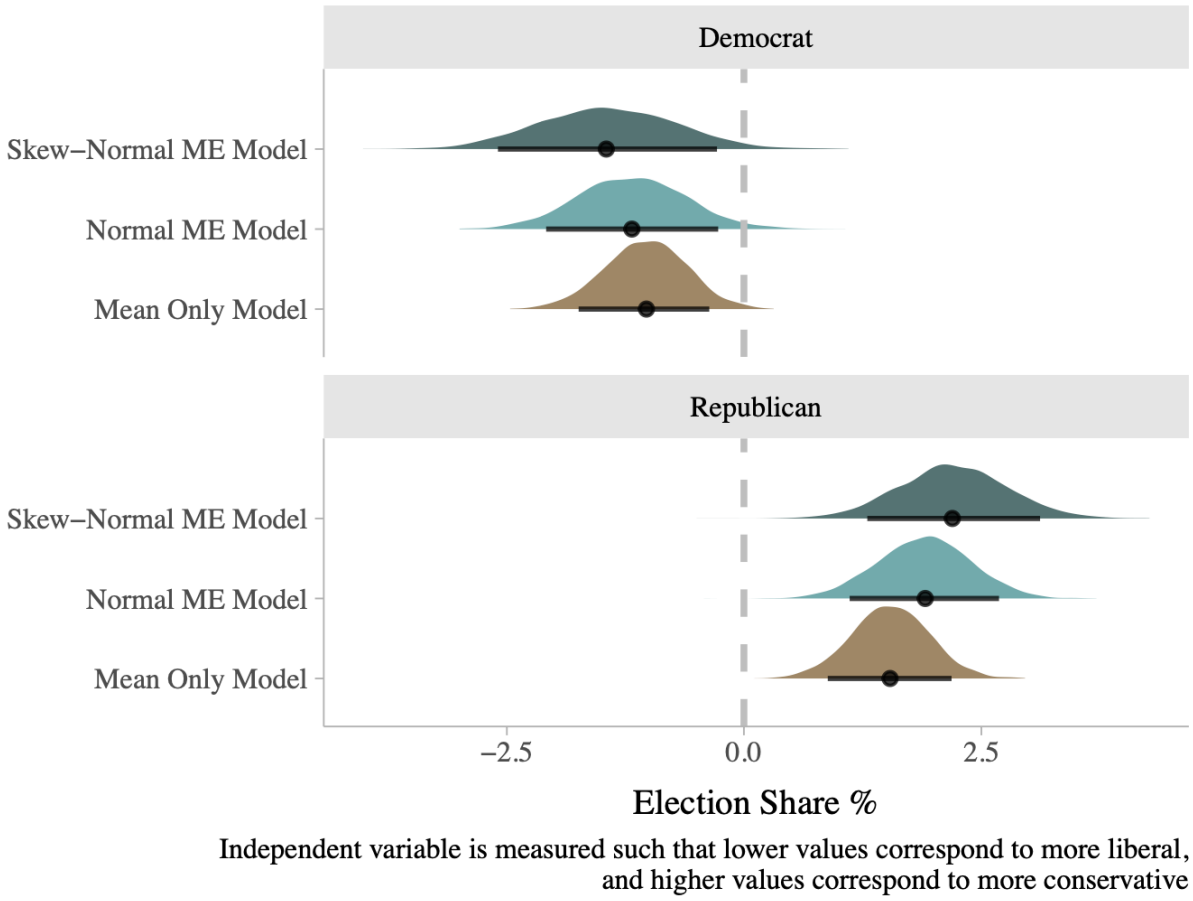

Figure 3.9: The Effect of Legislator Ideology on Vote Share in Next Election

3.4 Measurement Error Validity

One of the key takeaways from this project is the importance of accounting for measurement error in theory testing. The posterior distribution generated by measurement models contains valuable information about this error but is often discarded. Unobserved measurement error, however, can also have an independent effect on latent variables (see Figure 3.10). If this unobserved error outweighs the observed error, the effectiveness of the joint measurement theory-testing method proposed in this paper may be diminished. Therefore, it is important to estimate the measurement model posterior distribution as accurately as possible.

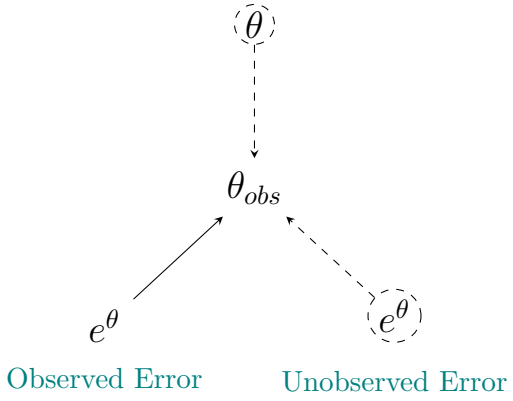

Figure 3.10: Total Measurement Error

Social scientists should be aware of recent computational advancements in Bayesian posterior estimation, particularly the superiority of Hamiltonian Monte Carlo (HMC) samplers over traditional Gibbs samplers (such as the one used in the **pscl** R package). HMC methods enable more accurate and efficient handling of high dimensional parameter spaces (Betancourt 2018)—of which IRT models are a prime example. Gibbs samplers also lack the sophisticated suite of diagnostic tools that come with HMC. This means that convergence issues, and therefore poor posterior exploration, may go undetected. These makes the observed error estimates in the posterior less trustworthy.

In the same vein as using the best computational sampling methods, this project highlights

why latent variables whose measurement models produce rich uncertainty estimates should be preferred over those that do not. The methodological competitor to IRT models for measuring political ideology is DW-NOMINATE, which uses multidimensional scaling rather than Bayesian estimation. This optimization method does not provide explicit estimates of uncertainty, much less a rich posterior distribution of values the latent variable could take. All of DW-NOMINATE's uncertainty estimates must come from bootstrap procedures which produce, at best, only standard error estimates. This means that more of this model's measurement error is in the unobserved category, which in turn raises concerns about attenuation bias and/or confounding bias when using DW-NOMINATE values in a theory-testing model.

3.5 Conclusion

This project highlights the problems associated with ignoring measurement error when testing theories which rely on measurement model variables. Doing so will often lead to attenuation bias, which could lead to the mistaken conclusion that no relationship between the independent variable and dependent variable exist, when in fact it does. And in even more problematic cases, ignoring measurement model error can introduce confounding bias into the theory-testing analysis. The method proposed here: to simultaneously estimate the theory-testing model and measurement model at once, helps fix these two issues to the extent that the posterior distribution of the measurement model is valid.

3.6 References

Azzalini, Adelchi, and Antonella Capitanio. 2014. *The Skew-Normal and Related Families.* Institute of Mathematical Statistics Monographs 3. Cambridge: Cambridge University Press.

Bafumi, Joseph, Andrew Gelman, David K. Park, and Noah Kaplan. 2005. "Practical Issues in Implementing and Understanding Bayesian Ideal Point Estimation." *Political Analysis* 13 (2): 171–87. https://doi.org/10.1093/pan/mpi010.

Betancourt, Michael. 2018. "A Conceptual Introduction to Hamiltonian Monte Carlo." *arXiv*, 60.

Blackwell, Matthew, James Honaker, and Gary King. 2017. "A Unified Approach to Measurement Error and Missing Data: Overview and Applications." *Sociological Methods & Research* 46 (3): 303–41. https://doi.org/10.1177/0049124115585360.

Blalock, H. M. 1970. "A Causal Approach to Nonrandom Measurement Errors." *American Political Science Review* 64 (4): 1099–1111. https://doi.org/10.2307/1958360.

Canes-Wrone, Brandice, David W. Brady, and John F. Cogan. 2002. "Out of Step, Out of Office: Electoral Accountability and House Members' Voting." *American Political Science Review* 96 (1): 15.

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "*Stan* : A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1). https://doi.org/10.18637/jss.v076.i01.

Carroll, Royce, Jeffrey B. Lewis, James Lo, Keith T. Poole, and Howard Rosenthal. 2009. "Measuring Bias and Uncertainty in DW-NOMINATE Ideal Point Estimates via the Parametric Bootstrap." *Political Analysis* 17 (3): 261–75. https://doi.org/10.1093/pan/mpp005.

Cinelli, Carlos, Andrew Forney, and Judea Pearl. 2020. "A Crash Course in Good and Bad Controls." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3689437.

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98 (2): 355–70. https://doi.org/10.1017/S0003055404001194.

Gallop, Max, and Simon Weschle. 2019. "Assessing the Impact of Non-Random Measurement Error on Inference: A Sensitivity Analysis Approach." *Political Science Research and Methods* 7 (2): 367–84. https://doi.org/10.1017/psrm.2016.53.

Imai, Kosuke, and Teppei Yamamoto. 2010. "Causal Inference with Differential Measurement Error: Nonparametric Identification and Sensitivity Analysis." *American Journal of Political Science* 54 (2): 543–60. https://doi.org/10.1111/j.1540-5907.2010.00446.x.

McElreath, Richard. 2020. *Statistical Rethinking: A Bayesian Course with Examples in r and Stan.*

2nd ed. CRC Texts in Statistical Science. Boca Raton: Taylor; Francis, CRC Press.

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Analytical Methods for Social Research. New York: Cambridge University Press.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge, U.K. ; New York: Cambridge University Press.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66 (5): 688–701. https://doi.org/10.1037/h0037350.

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