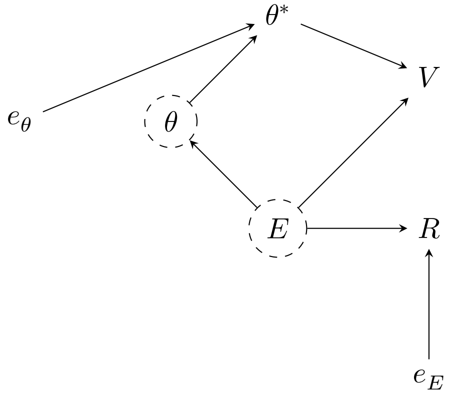
Going Beyond Ideal Point Points: Modeling Measurement Model Measurement Error

Bertrand Wilden

5/31/23



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.

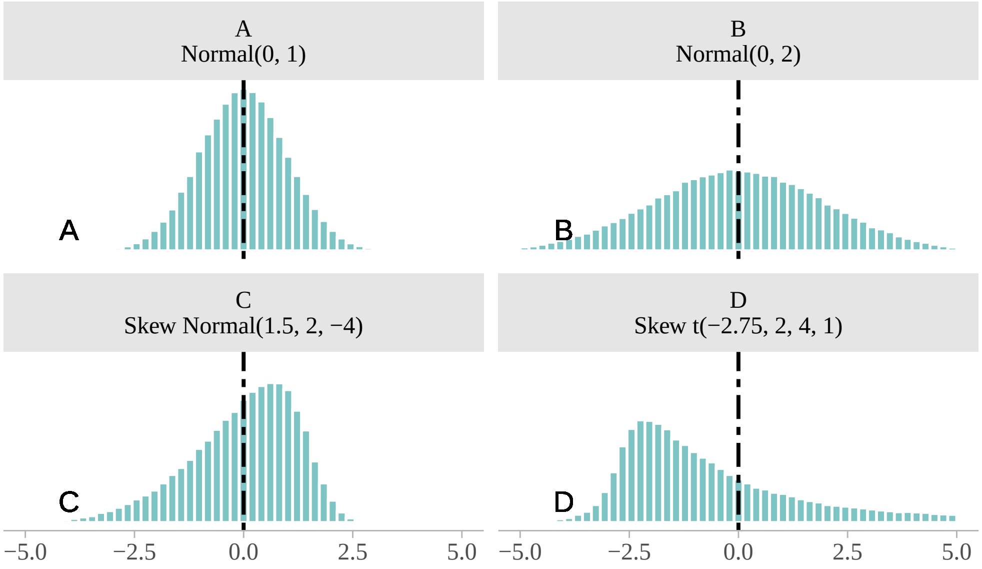
I will refer to this set of models as *theory-testing models*.

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 a Bayesian measurement model.

Theory-testing research which uses estimates from measurement models typically reduces the posterior distributions down to a single *maximum a posteriori* (MAP) 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 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 create distributions which vary to a large extent.

Figure 1: Ignoring Measurement Error in Measurement Models



Four different measurement model posterior distributions with mean zero

In Figure 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, as I will show, can lead to attenuation bias—or 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. Panels C and D in #fig-measurement-model 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.

1.1 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 can easily treat parameters and data interchangeably (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 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 y^*. The posterior estimates for θ_i from the measurement model $g(\cdot)$ are then used as data in the theory-testing model $f(\theta_i)$ using the outcome of interest y.

$$\begin{aligned} y_i &\sim f(\theta_i) \\ y_i^* &\sim g(\theta_i) \end{aligned} \tag{1}$$

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. Measurement models can often be challenging to fit in practice due to issues of identifiability.

My method overcomes the two problems outlined above by simplifying the measurement model step, $y_i^* \sim g(\cdot)$ 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 $\mathbb{E}[\theta_i] \sim N(\theta_i, \sigma_i)$. The values $\mathbb{E}[\theta_i]$ and σ_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.

1.2 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 (y^* in Equation 1) to produce a posterior distribution of continuous values for each actor (e.g. member of Congress) on this left-right scale. [*insert footnote about how these models can measure other types of latent variables*]

Let's say we wanted to estimate the effect of legislator ideology, θ, on some outcome, y. Using the format I developed in Equation 1, Equation 2 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 voted yes or no on a piece of legislation, y^*_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 used as data in the linear regression theory-testing model (top) to estimate β—the coefficient of theoretical interest.

$$\begin{aligned} y_i &\sim \text{Normal}(\alpha + \theta_i\beta, \sigma^2) \\ y^*_{ij} &\sim \text{Bernoulli}[\Phi(\gamma_j\theta_i + \eta_i)] \end{aligned} \tag{2}$$

As mentioned previously, however, estimating Equation 2 would not generally be feasible due to computational constraints. Instead, the IRT ideology measurement model should be fit beforehand.

Then, for each legislator's posterior distribution of θ, the values $\mathbb{E}[\theta_i]$ and $\sigma^2_{\theta i}$ should be calculated. These values can then used as data in the simplified measurement model in Equation 3 in order to estimate the latent θ_i for each observation.

$$\begin{aligned} y_i &\sim \text{Normal}(\alpha + \theta_i \beta, \sigma^2) \\ \mathbb{E}[\theta_i] &\sim \text{Normal}(\theta_i, \sigma^2_{\theta i}) \end{aligned} \tag{3}$$

If the posterior estimates of θ from the IRT measurement model are truly normally distributed for each legislator, then Equation 2 and Equation 3 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.

2 Measurement Error Models

In this section I will 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.

I will demonstrate my argument using the causal graph framework (Pearl 2000). Causal graphs are heuristic tools that 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]

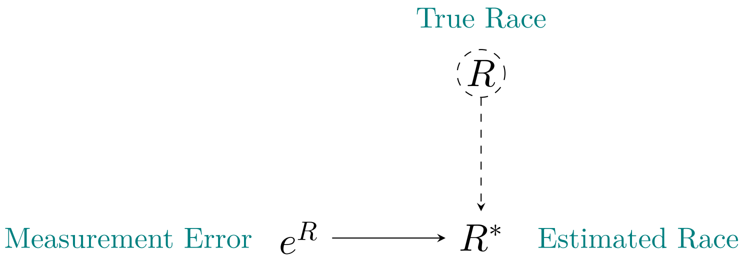
2.1 Measurement Error Confounding Bias

Figure 2 shows a simple DAG outlining the causal process implied by the IRT ideology measurement model. The estimated ideology variable produced by the model, θ^* is a function of an individual's true ideology, θ and measurement error, e^θ. The important takeaway from Figure 2 is that the ideology variable produced by IRT models arises causally from a combination of some latent true ideology (which we do not observe), and from measurement error (which we observe, at least partially, in the form of the posterior distribution for θ^*). There is likely also considerable unobserved measurement error in these models that can be due to a variety of factors. IRT models assume

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

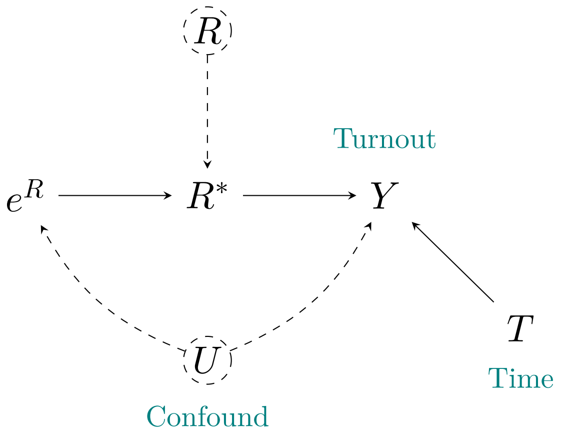
each legislator's decision to vote on bill is influenced solely by their innate ideology, rather than on strategic concerns. A violation of this assumption, therefore, will produce biased estimates of θ^*. There are also computational issues with fitting IRT models which may make the posterior estimates untrustworthy. For the purposes of this illustration, however, I will assume that the posterior distribution, e^θ for θ^* contains all relevant information about the measurement error for the true ideology θ.

Figure 2: IRT Ideology Measurement Error Model



Now consider a hypothetical theory-testing model constructed to determine whether there were racial disparities in voter turnout, Y after the enactment of voter identification laws at time, T. Suppose we do not have data on individuals' self-reported race, and therefore must use BSIG to estimate R^*. Figure 3 shows a potential causal graph for this system. The main causal effect of interest is represented by the path $R^* \longrightarrow Y$, moderated by T. In order to get an unbiased estimate of this causal effect we need to close all backdoor paths leading from R^* to Y, which in this case, flows through the unobserved variable U. This confound represents anything that is a common cause of both the BSIG measurement error and voter turnout. Argyle and Barber (2022) systematically review BSIG misclassification rates and find that socio-economic status and geography affect the amount of measurement error in the algorithm. It is also highly plausible that these variables have an independent effect on turnout rates in the population.

Figure 3: BSIG Measurement Error in a Hypothetical Theory-Testing Model



For the theory-testing model in Figure 3 it may be possible to condition on some variables in U in order to obtain an unbiased estimate of $R^* \longrightarrow Y$. But with something as multi-faceted as socio-economic status, there is always the 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 with U at all. This is because the measurement error, e^R in Figure 3 is part of the backdoor path from R^* to Y. Therefore when we explicitly incorporate e^R into a model estimating $R^* \longrightarrow Y$ we can obtain an unbiased estimate of the causal effect of race on turnout.

While I have only presented a discussion of a single theory-testing model—the impact of race on turnout—incorporating e^R into any model using BSIG race estimates will be valuable. Race is often thought to be an “un-caused cause” because it is assigned from an individual’s birth, hence unbeholden to influence from other variables. (Sen and Wasow 2016). This removes the need to worry about potential confounds because there cannot be any backdoor paths through the variable race. But as Figure 3 shows, BSIG *estimates* of race, R^* always have the potential to have open backdoor paths through measurement error e^R. Simply including e^R directly in a joint measurement theory-testing model blocks this backdoor path going into R^*, which helps identify the direct effect of race on the outcome of interest.

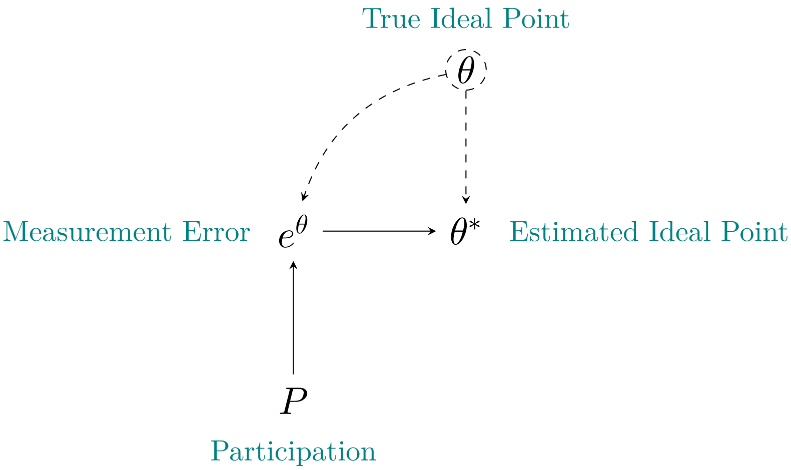
2.2 Measurement Error Attenuation Bias

The previous discussion of BSIG race estimates highlighted how failing to account for measurement error could lead to bias via backdoor confounding. This general issue is 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.

Measurement model errors can also take the form of classical measurement error. This is where 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. As I demonstrate via simulations, my method can help correct this kind of bias in theory-testing models as well.

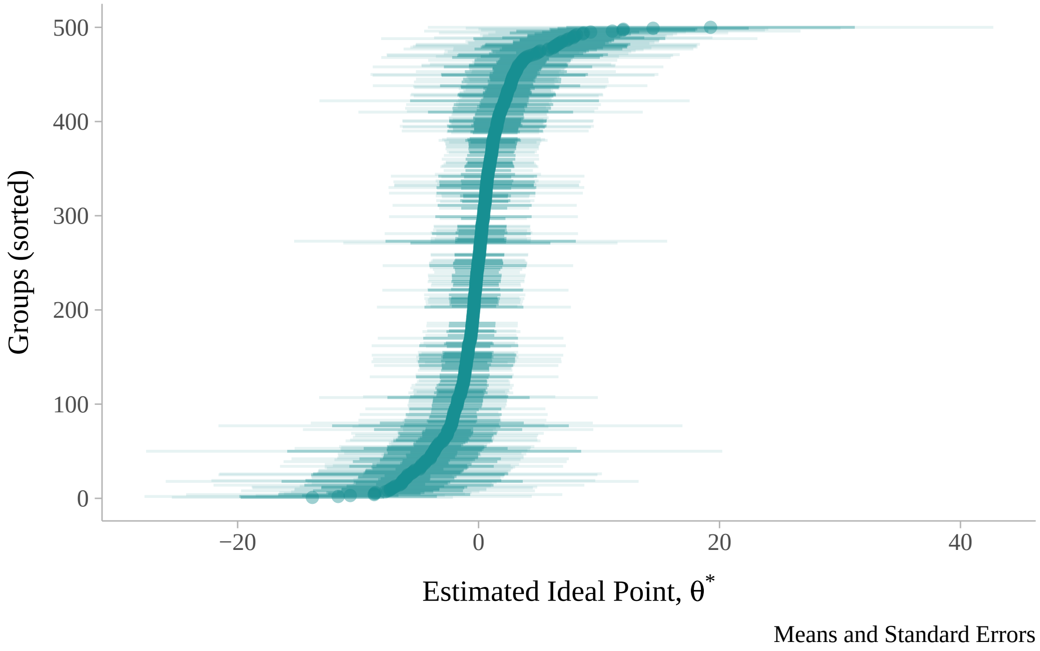
3 Simulation Study - Bayesian IRT

Figure 4: Bayesian IRT Measurement Model



In this section I will demonstrate, through simulations, how my method of constructing a joint measurement theory-testing model can help mitigate attenuation bias that arises from classic measurement error. Figure 4 shows the causal process which produces ideal point estimates, θ^* in a Bayesian IRT model. As in the BSIG case, our observed measurements come from an unobserved latent variable plus some measurement error. At least two variables can affect the measurement error, e^{θ} in Figure 4. 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 5 shows an example of what the distribution of posterior estimates from a Bayesian IRT model look like. Groups further from the center have much wider ideal point posterior distributions. The second variable affecting e^{θ} is group participation, P—how often the group signals its position among the items in the model. Groups that signal more positions on items will have smaller levels of measurement error compared to groups that signal fewer positions because we have more data on their true ideological preferences.

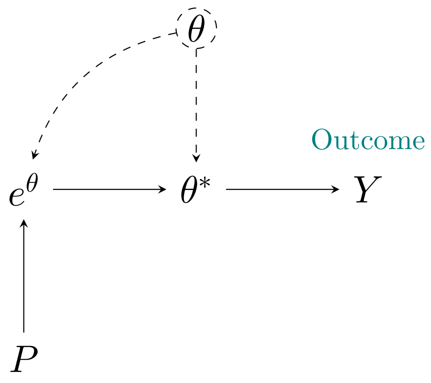
Figure 5: Bayesian IRT Posterior Distributions



While there may exist some backdoor paths through e^{θ} and P in a hypothetical theory-testing model, I will assume that the outcome of interest, Y is unaffected by anything other than the direct causal effect $\theta^* \longrightarrow Y$.[2] The purpose of this simplification is to highlight the consequences of random

[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 6: Bayesian IRT Measurement Model in Hypothetical Theory-Testing Model



measurement error during parameter estimation of a theory-testing model. Using the generative causal model in Figure 6 we can simulate data with a known parameter for the effect of group ideal point on the outcome Y. Then we fit two linear regression models to estimate this parameter, β. Equation 4 is the naive theory-testing model where $X_{\text{MEAS},i}$ corresponds to a group's mean ideal point estimate from the Bayesian IRT measurement model (see right panel of Figure 1). This is in contrast to the joint measurement theory-testing model in Equation 5, which models $X_{\text{MEAS},i}$ as an outcome of $X_{\text{TRUE},i}$ (an unobserved parameter for each observation)[3] and $X_{\text{SE},i}$ which is the observed standard deviation of the posterior distribution for each groups' ideal point. The estimates of $X_{\text{TRUE},i}$ are then used in the linear model which predicts the outcome y_i.

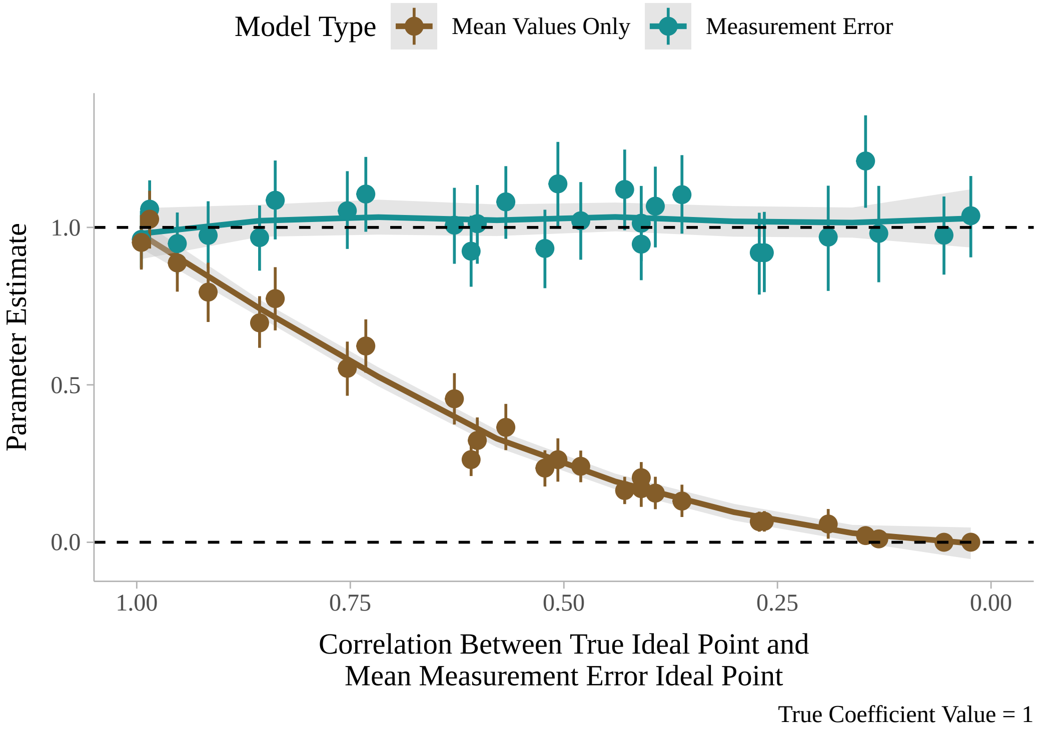
[3] The parameters $X_{\text{TRUE},i}$ should not be confused with the true ideal points produced in the simulation. These true values are never seen by either model.

$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta X_{\text{MEAS},i} \\
\alpha &\sim \text{Normal}(0, 2) \\
\beta &\sim \text{Normal}(0, 2) \\
\sigma &\sim \text{Half Student t}(3, 0, 2)
\end{aligned}
\tag{4}
$$

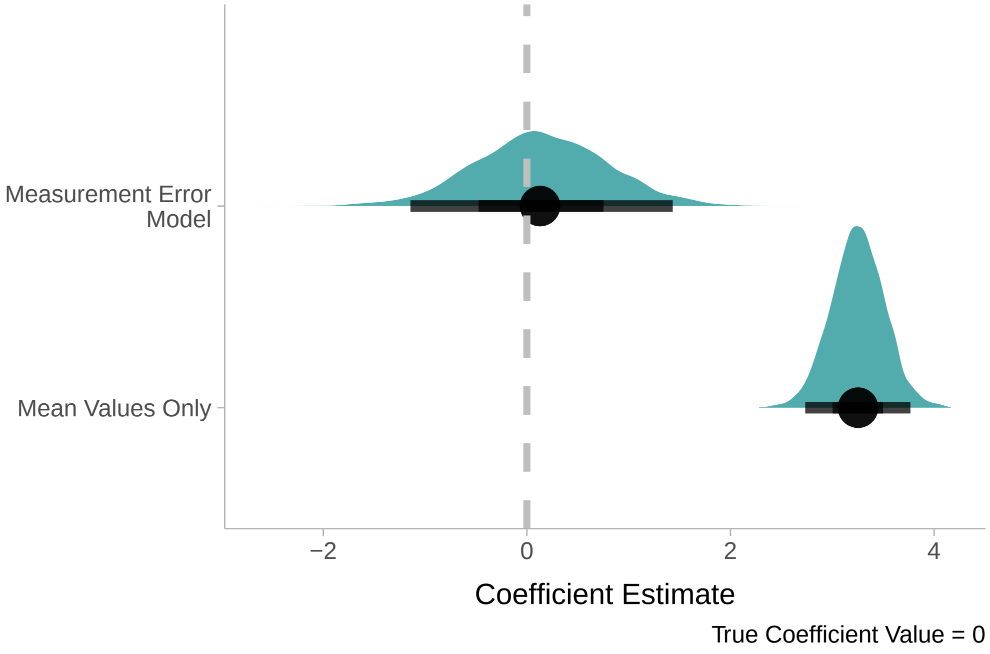
$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta X_{\text{TRUE},i} \\
X_{\text{MEAS},i} &\sim \text{Normal}(X_{\text{TRUE},i}, X_{\text{SE},i}) \\
X_{\text{TRUE},i} &\sim \text{Normal}(0, \tau) \\
\alpha &\sim \text{Normal}(0, 2) \\
\beta &\sim \text{Normal}(0, 2) \\
\sigma &\sim \text{Half Student t}(3, 0, 2) \\
\tau &\sim \text{Half Student t}(3, 0, 2)
\end{aligned}
\tag{5}
$$

Figure 7 shows how well each model recovers the true parameter value for β: the effect of the true ideal point on the simulated outcome. Each model was fit 40 times across a range of increasing random error levels (shown on the horizontal axis as the correlation between the simulated true ideal point and mean measurement error value approach zero). The mean posterior estimates of each model's β parameter were then plotted using a loess fit. With little-to-no measurement error (left side of the graph), both models reliably recover the true β value of 1. But as the random measurement error increases, the β estimates from the naive model from Equation 4 rapidly attenuate towards zero. This is in contrast to the estimates from the joint measurement theory-testing model in Equation 5 which attenuate slightly but remain much closer to the true β value even after there is essentially zero correlation between the true ideal points and means from the ideal points with measurement error.

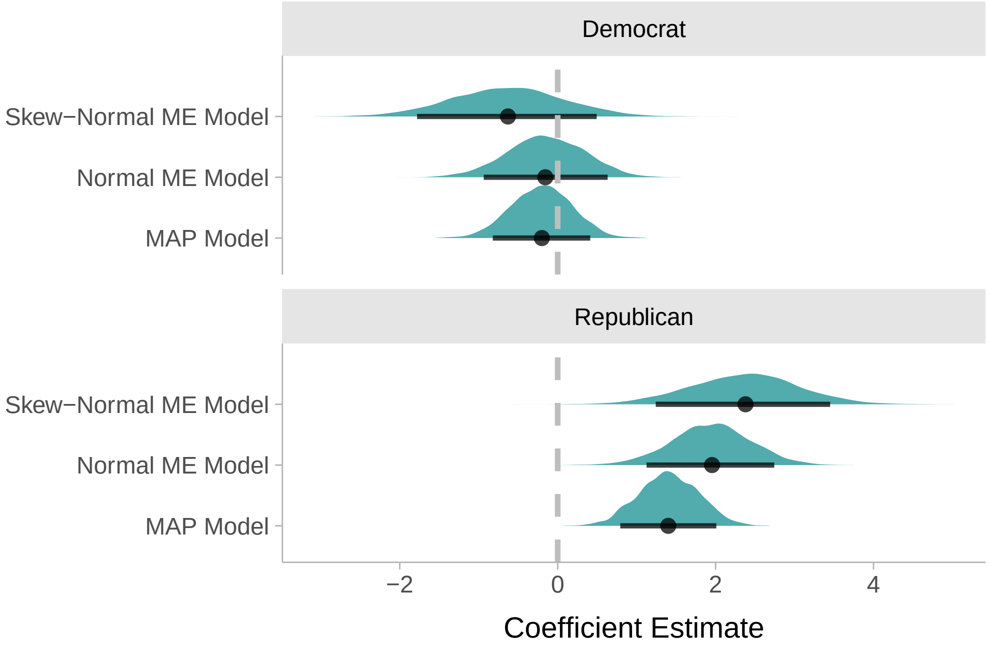
Figure 7: Parameter Recovery as Measurement Error Increases



4 Skew-Normal Distribution



5 Case Study: Does Political Extremism Affect Election Outcomes?



5.1 Extension Skew-t

5.2 Importance of gibbs vs HMC vs no uncertainty

References

Argyle, Lisa, and Michael Barber. 2022. "Misclassification and Bias in Predictions of Individual Ethnicity from Administrative Records."

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.

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.

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.

Gabry, Jonah, and Rok Češnovar. 2022. *Cmdstanr: R Interface to CmdStan*.

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.

Kay, Matthew. 2021. *Ggdist: Visualizations of Distributions and Uncertainty.*

Landau, William Michael. 2022. *Targets: Dynamic Function-Oriented Make-Like Declarative Workflows.* https://CRAN.R-project.org/package=targets.

Martin, Andrew D., Kevin M. Quinn, Jong Hee Park, Ghislain Vieilledent, Michael Malecki, Matthew Blackwell, Keith Poole, et al. 2022. *MCMCpack: Markov Chain Monte Carlo (MCMC) Package.* https://CRAN.R-project.org/package=MCMCpack.

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.

Mills, Blake Robert. 2022. *MetBrewer: Color Palettes Inspired by Works at the Metropolitan Museum of Art.* https://CRAN.R-project.org/package=MetBrewer.

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.

Pedersen, Thomas Lin. 2020. *Patchwork: The Composer of Plots.* https://CRAN.R-project.org/package=patchwork.

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.

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.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. 2023. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wilke, Claus O., and Brenton M. Wiernik. 2022. *Ggtext: Improved Text Rendering Support for Ggplot2*. https://wilkelab.org/ggtext/.