Going Beyond Ideal Point Points: Modeling Measurement Model Measurement Error

Bertrand Wilden

2023-09-15

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

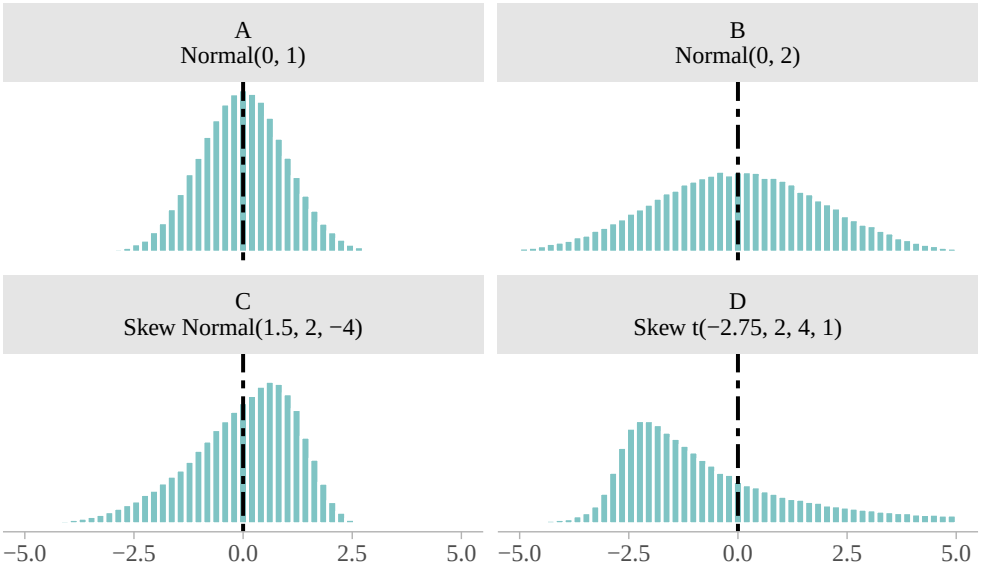
uncertainty from the measurement model into the theory-testing model, as I will show, can lead to mistaken conclusions regarding the underlying research question. And unlike so-called "classical" measurement error, whose attenuation bias is generally well known, the mistakes I investigate can lead to 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 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 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 widely.

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

Figure 1: Ignoring Measurement Error in Measurement Models



Four different measurement model posterior distributions with mean 0

in fact does. In other words, the method I propose can help increase the statistical power of theory tests. Panels C and D in Figure 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.

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, 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 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 used as data in the theory-testing model $f(\theta_i)$ using the outcome of interest y.

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

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. Because latent variables have no objective scale, measurement models can often be challenging to

fit in practice due to issues of identifiability.

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 $\bar{\theta}_i \sim N(\theta_i, \sigma^2_{\bar{\theta}[i]})$. The values $\bar{\theta}_i$ and $\sigma^2_{\bar{\theta}[i]}$ 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.

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 (z in Equation 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 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, 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 used as data in the linear regression theory-testing model (top) to estimate β_1—the coefficient of substantive interest.

$$\begin{aligned} y_i &\sim \text{Normal}(\beta_0 + \beta_1\theta_i, \sigma^2) \\ z_{ij} &\sim \text{Bernoulli}[\Phi(\gamma_j\theta_i + \xi_i)] \end{aligned} \tag{2}$$

As mentioned previously, however, estimating Equation 2 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 $\bar{\theta}_i$ 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 in order to estimate the latent θ_i for each observation.

$$\begin{aligned} y_i &\sim \text{Normal}(\beta_0 + \beta_1\theta_i, \sigma^2) \\ \bar{\theta}_i &\sim \text{Normal}(\theta_i, \sigma^2_{\theta[i]}) \end{aligned} \tag{3}$$

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]

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 1 helps fix them. First I will 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. 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 1 ameliorates this confounding bias. For both types of measurement error, I will use simulation studies to demonstrate how effective each modeling approach is at recovering known parameter values from the theory-testing model.

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 $\bar{\theta}_i$, the more likely we are to make a False Negative

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

error in our theory-testing model.

Let's return to the Bayesian IRT measurement model used to estimate political ideology for legislators. Figure 2 shows the causal process which produces these ideal point estimates. The observed measurements of ideology, $\bar{\theta}$ come from an unobserved latent variable plus some measurement error. At least two variables can affect the measurement error, e^{θ} in Figure 2. First, 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 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 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 data on their true ideological preferences.

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

Now let's expand Figure 2 into a theory-testing model with Figure 4. While there may exist

Figure 2: IRT Measurement Model

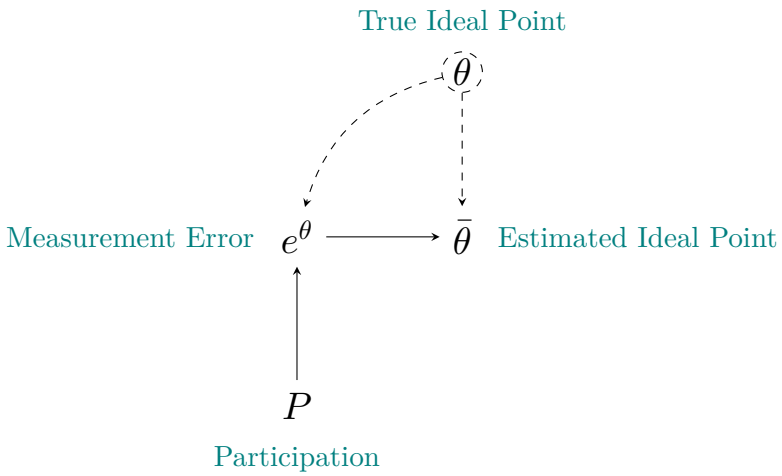
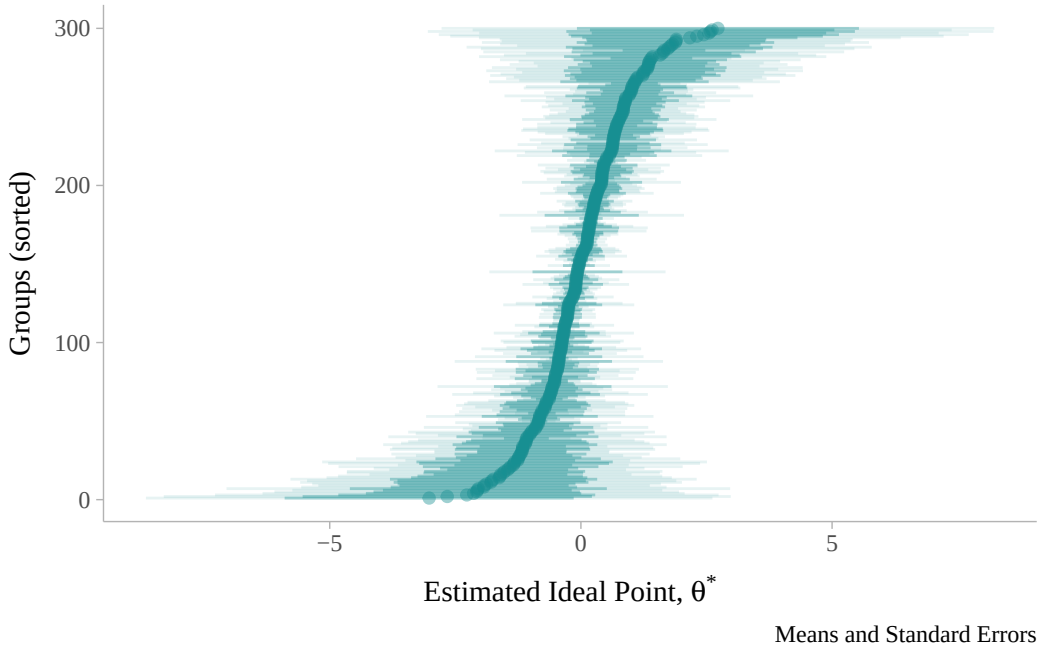
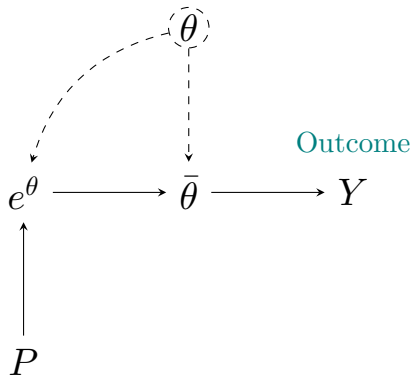


Figure 3: IRT Model Posterior Distributions



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 $\bar{\theta} \longrightarrow Y$.[2] The purpose of this simplification is to highlight the consequences of random measurement error during parameter estimation of a theory-testing model.

Figure 4: IRT Measurement Model in Hypothetical Theory-Testing Model



2.1.1 Simulation Study: Attenuation Bias

Using the generative causal model in Figure 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 4 is the naive theory-testing model where $\bar{\theta}_i$ corresponds to a legislator's mean ideal point estimate from the Bayesian IRT measurement model. This is in contrast to the joint measurement theory-testing model in Equation 5, which models $\bar{\theta}_i$ as an outcome of the true ideology θ_i (an unobserved parameter for each observation) and $\sigma^2_{\bar{\theta}[i]}$ which is

[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.

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 the No-U-Turn (NUTS) MCMC sampler (Carpenter et al. 2017).

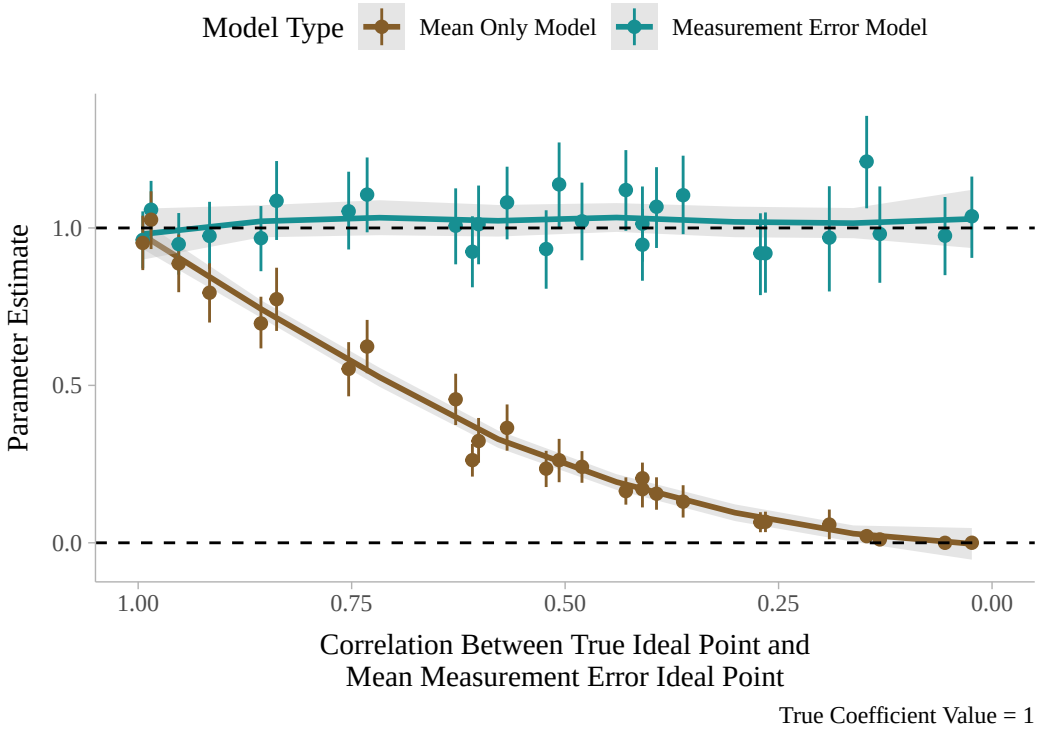
$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma^2) \\
\mu_i &= \beta_0 + \beta_1 \bar{\theta}_i \\
\beta_0, \beta_1 &\sim \text{Normal}(0, 2) \\
\sigma &\sim \text{Half Student t}(3, 0, 2)
\end{aligned}
\tag{4}
$$

$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma^2) \\
\mu_i &= \beta_0 + \beta_1 \theta_i \\
\bar{\theta}_i &\sim \text{Normal}(\theta_i, \sigma^2_{\theta[i]}) \\
\theta_i &\sim \text{Normal}(\pi, \tau) \\
\beta_0, \beta_1 &\sim \text{Normal}(0, 2) \\
\sigma &\sim \text{Half Student t}(3, 0, 2) \\
\tau &\sim \text{Half Student t}(3, 0, 2) \\
\pi &\sim \text{Normal}(0, 1)
\end{aligned}
\tag{5}
$$

Figure 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 4 rapidly attenuate towards zero. This is in contrast to the estimates from the joint measurement theory-testing model in Equation 5 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 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 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.

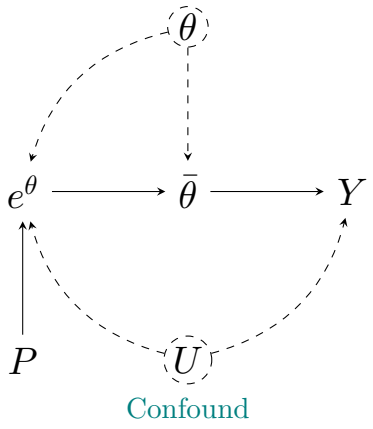
2.2 Measurement Error Confounding Bias

The previous discussion of highlighted how failing to account for random measurement error could lead to attenuation bias. Next I will turn to the problem of 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 Figure 4, let's now consider the hypothetical causal graph shown in Figure 6. As before, the main causal effect of interest is represented by the path $\bar{\theta} \longrightarrow Y$. In order to get an unbiased estimate of this causal effect we need to close all backdoor paths leading from $\bar{\theta}$ to Y, which in this case, flows through the unobserved variable U. This confound 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 6 it may be possible to directly condition on some variables in U in order to obtain an unbiased estimate of $\bar{\theta} \longrightarrow Y$. But with something as multi-

Figure 6: IRT Measurement Model in Hypothetical Theory-Testing Model with Confounding



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 with U at all. This is because the measurement error, e^{θ} in Figure 6 is part of the backdoor path from $\bar{\theta}$ to Y. Therefore when we explicitly incorporate e^{θ} into a model estimating $\bar{\theta} \longrightarrow Y$ 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.

2.2.1 Simulation Study: Confounding Bias

In order 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 Section 2.1.1. Given the causal graph Figure 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. $\bar{\theta}$ 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 6.

$$\frac{2}{\omega\sqrt{2\pi}} e^{-\frac{(x-\xi)^2}{2\omega^2}} \int_{-\infty}^{\alpha\left(\frac{x-\xi}{\omega}\right)} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \, dt \tag{6}$$

Equation 6 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, we fit two models: the ordinary linear regression using only $\bar{\theta}$ values as used previously in the attenuation bias example (Equation 4), and a modified version of Equation 5 which substitutes the Normal distribution for the Skew Normal distribution (Equation 7). As before, the key parameter of theoretical interest is β_1—which, according to the simulated data, should equal zero if unconfounded.

$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma^2) \\
\mu_i &= \beta_0 + \beta_1 \theta_i \\
\bar{\theta}_i &\sim \text{Skew Normal}(\theta_i, \omega_{\theta[i]}, \alpha_{\theta[i]}) \\
\theta_i &\sim \text{Normal}(\pi, \tau) \\
\beta_0, \beta_1 &\sim \text{Normal}(0, 2) \\
\sigma &\sim \text{Half Student t}(3, 0, 2) \\
\tau &\sim \text{Half Student t}(3, 0, 2) \\
\pi &\sim \text{Normal}(0, 1)
\end{aligned}
\tag{7}
$$

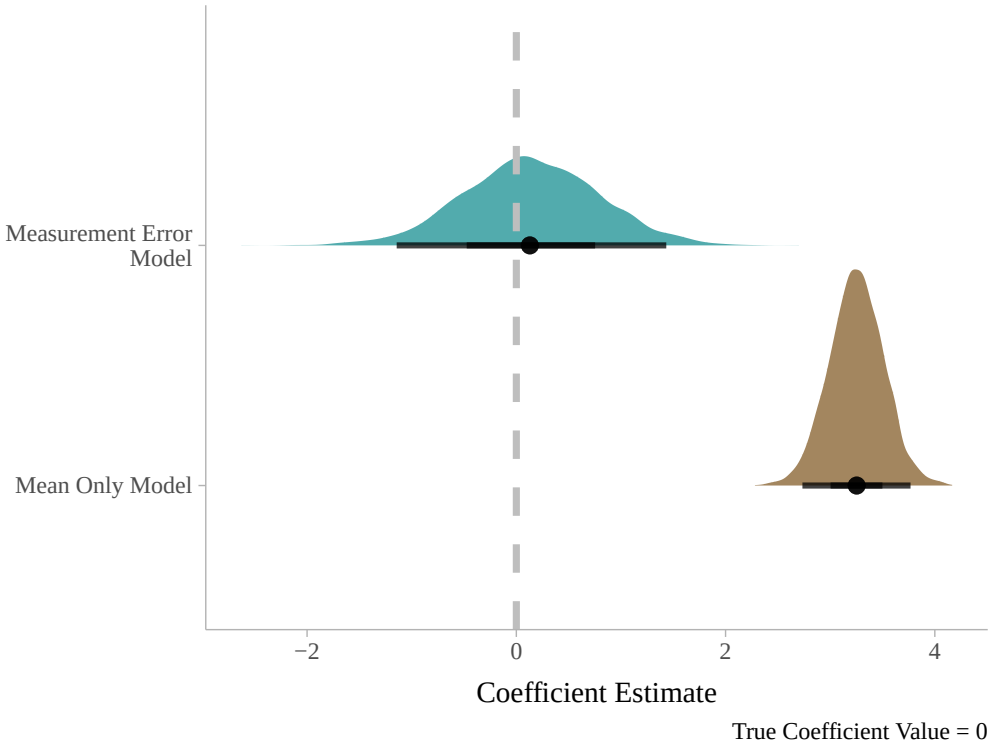
Figure 7 shows how well each model estimates β_1. As expected, the bottom model from Equation 4 using only the $\bar{\theta}$ values produces a biased estimate of β_1. The open backdoor path $\bar{\theta} \longleftarrow e^{\theta} \longleftarrow U \longrightarrow Y$ from Figure 6 confounds the causal effect $\theta \longrightarrow Y$ if the latent variable was measured perfectly. In contrast, the top model from Equation 7 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 error model is faithfully propagating the uncertainty from the measurement process into the final theory-testing analysis.

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

[3]The scale parameter $\omega_{\theta[i]}$ is also included in Equation 7 to help avoid attenuation bias.

Figure 7: Parameter Recovery Under Confounding



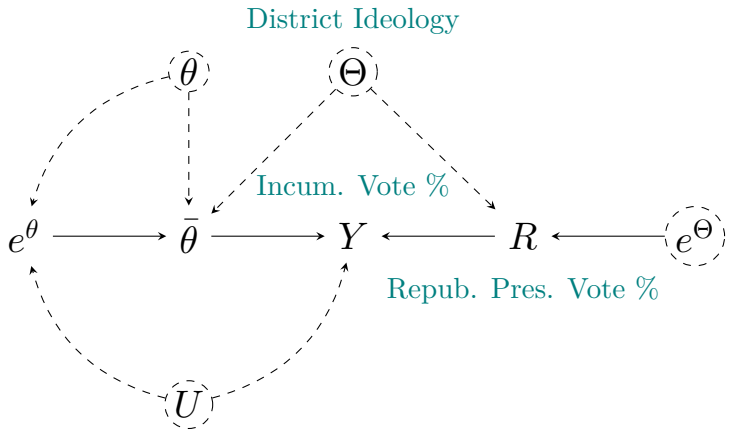
are powerful because we have complete control over the data generating process, and therefore know how well each model can recover the parameters which generated the outcome Y. The downside of simulation studies, however, is that they greatly simplify the complex social phenomena they aim to represent. 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 $\bar{\theta}_i \sim N(\theta_i, \sigma^2_{\theta[i]})$ 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 (Jackman et al. 2020).

Because this research question relies on observational data, it is important to sketch out a causal graph of the system in order to understand how to isolate the main effect. Figure 8 is one plausible causal graph of this system. The main effect of interest is $\bar{\theta} \longrightarrow Y$: measured candidate ideology's effect on vote share in the general election. As in the simulation study examples, $\bar{\theta}$ 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 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 $\Theta \longrightarrow \bar{\theta}$ 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 8.

The data for this analysis came from a variety of sources. Lewis et al. (2023) provided 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** (Azzalini 2022) 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 Bensen (2016). 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 positive relationship between ideology and vote share as district electorates favoring moderate candidates, whereas a positive relationship for Republican candidates would mean district electorates favor more extremists.

$$
\begin{aligned}
\text{VotePct}_i &\sim \text{Normal}(\mu_i, \sigma^2) \\
\mu_i &= \beta_0 + \beta_1 \bar{\theta}_i + \beta_2 \text{R}_i \\
\beta_0 &\sim \text{Normal}(50, 5) \\
\beta_1, \beta_2 &\sim \text{Normal}(0, 2) \\
\sigma &\sim \text{Half Student t}(3, 0, 2)
\end{aligned}
\tag{8}
$$

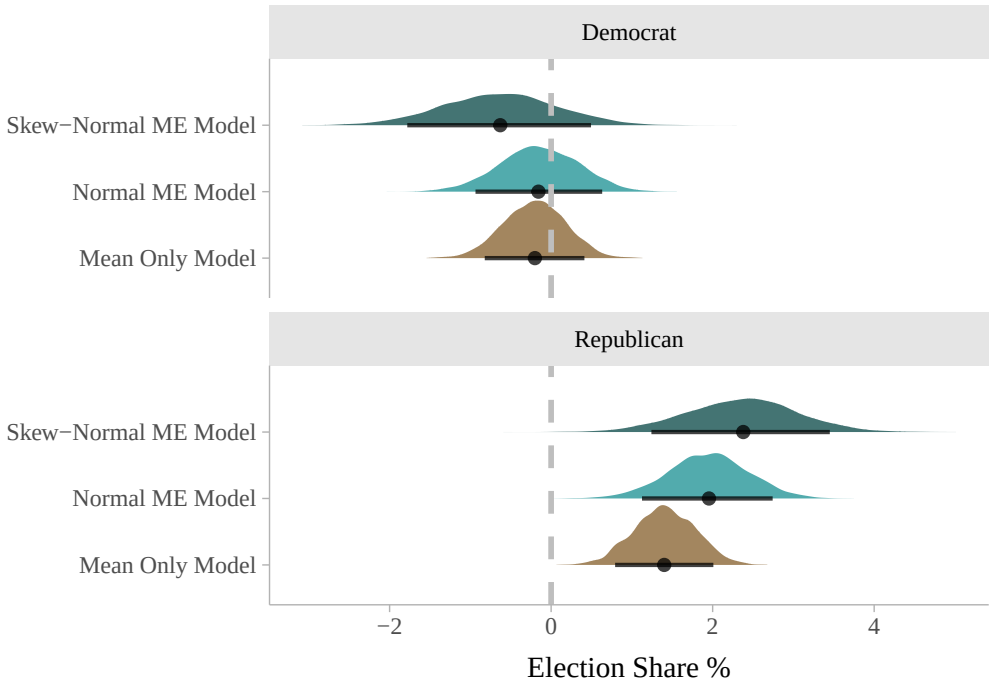
Equation 8 shows the simple linear regression without using a measurement error model for candidate ideology. Equation 9 models the measurement error using the Skew-Normal distribution as discussed in the previous section. This attempts to handle both attenuation bias and confounding from U in Figure 8. Both models control for district Republican presidential vote share, R to close the backdoor path $\bar{\theta} \longleftarrow \Theta \longrightarrow R \longrightarrow Y$. 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.

$$
\begin{aligned}
\text{VotePct}_i &\sim \text{Normal}(\mu_i, \sigma^2) \\
\mu_i &= \beta_0 + \beta_1 \theta_i + \beta_2 \text{R}_i \\
\bar{\theta}_i &\sim \text{Skew Normal}(\theta_i, \omega_{\theta[i]}, \alpha_{\theta[i]}) \\
\theta_i &\sim \text{Normal}(\pi, \tau) \\
\beta_0 &\sim \text{Normal}(50, 5) \\
\beta_1, \beta_2 &\sim \text{Normal}(0, 2) \\
\sigma &\sim \text{Half Student t}(3, 0, 2) \\
\tau &\sim \text{Half Student t}(3, 0, 2) \\
\pi &\sim \text{Normal}(0, 2)
\end{aligned}
\tag{9}
$$

Figure 9 shows the posterior distributions for β_1 in the above models. For Democrats (top panel), there does not appear to be a strong relationship between incumbent ideology and their electoral success. Both the Normal measurement error model and mean-only model coefficients are centered around zero. Using the Skew-Normal measurement error model, however, there appears to be some evidence that centrist Democrats perform worse electorally compared to more liberal Democrats. For Republicans (bottom panel), all models agree that more extremist candidates perform better.

The fact that both measurement error models predict larger electoral gains from moving rightward suggests that there may be some attenuation bias taking place in the mean-only model.

Figure 9: The Effect of Legislator Ideology on Vote Share in Next Election



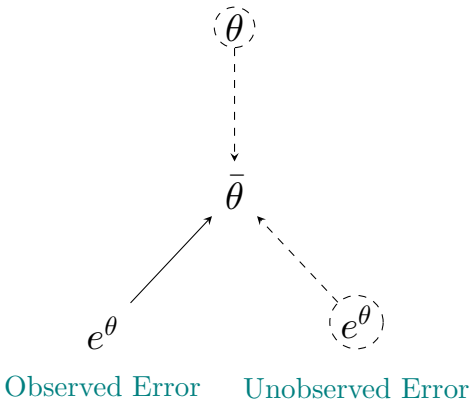
Independent variable is measured such that lower values correspond to more liberal, and higher values correspond to more conservative

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 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 10: Total Measurement Error



Social scientists should therefore 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 full posterior distribution of plausible values the latent variable could take. All of DW-NOMINATE's uncertainty estimates must come from bootstrap procedures which produce, at best, only standard errors. 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.

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.

References

Azzalini, Adelchi. 2022. *Sn: The Skew-Normal and Related Distributions Such as the Skew-t and the SUN*. http://azzalini.stat.unipd.it/SN/.

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.

Bensen, Clark H. 2016. "Presidential Results by Congressional District (PRCD)."

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.

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.

Jackman, Simon, Alex Tahk, Achim Zeileis, Christina Maimone, and Jim Fearon. 2020. *Pscl: Political Science Computational Laboratory*. http://github.com/atahk/pscl.

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.

Lewis, Jeffrey B., Keith T. Poole, Howard Rosenthal, Adam Boche, Aaron Rudkin, and Luke Sonnet. 2023. "Voteview: Congressional Roll-Call Votes Database."

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.

Volden, Craig, and Alan E. Wiseman. 2014. *Legislative Effectiveness in the United States Congress:*

The Lawmakers. New York, NY: Cambridge University Press.

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/.