How to Deal With Measurement Error in Measurement Models

Bertrand Wilden

2/23/23

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 latent variable values according their relative plausibility—not a single point estimate. Failure to propagate this uncertainty from the measurement model into

the theory-testing model, as I will show, can lead to biased and/or overly confident substantive conclusions.

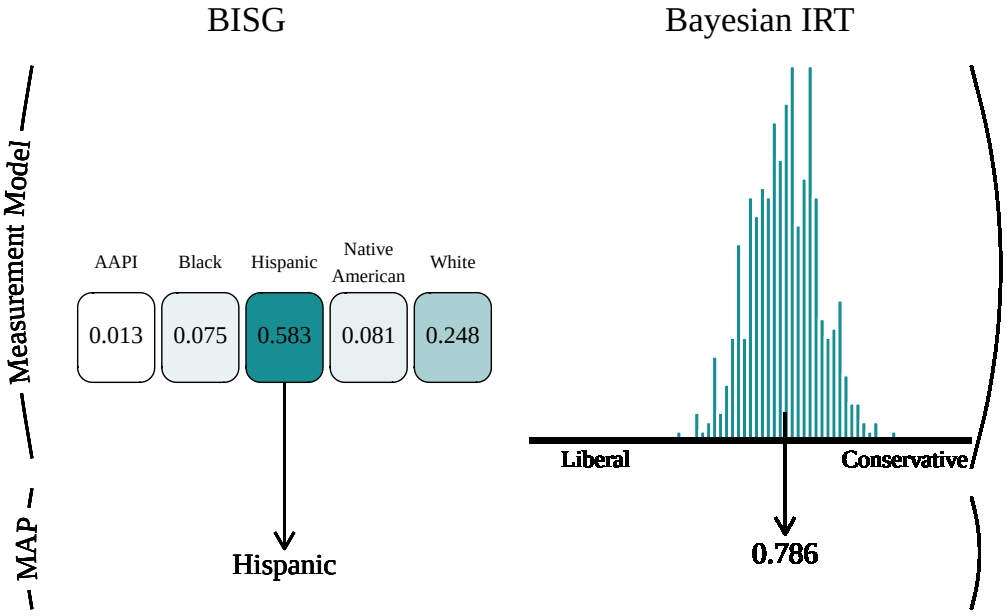
In this paper I develop two methods—one for continuous latent variables and one for discrete latent variables—which incorporate measurement model uncertainty into theory-testing models. I will use two running examples to help illustrate these methods. The first is the Bayesian Item-Response Theory (IRT) model used to measure political ideology (Clinton, Jackman, and Rivers 2004). Political ideology is typically imagined as existing on a latent left-right dimension. Bayesian IRT measurement models produce a posterior distribution of continuous values for each actor (e.g. member of Congress) on the left-right scale. The second illustrative model is Bayesian Improved Surname Geocoding (BISG), used to measure individuals' race and ethnicity in the absence of self-reported values for these characteristics (Elliott et al. 2009). Unlike political ideology, which could conceivably take any real-number value on the left-right scale, race and ethnicity are usually considered to be discrete, un-ordered, categories.[1] BISG measurement models, therefore, output a simplex for each individual corresponding to the posterior probability that they identify as a member of each potential racial or ethnic group.

Theory-testing research which uses these two measurement models typically reduces the posterior distributions down to a single *maximum a posteriori* (MAP) value. In the case of Bayesian IRT models, researchers select some statistic of central tendency from each posterior distribution to use in subsequent analyses, such as the mean, median, or mode. For BISG models, simply the race or ethnicity with highest posterior probability in the simplex is chosen. Figure 1 shows visually how

[1] For an alternative conception of race in social identity theory see Abdelal et al. (2006) .

information is lost when reducing the measurement model's posterior output to a single MAP value.

Figure 1: Ignoring Measurement Error in Measurement Models



While the two types of measurement models I consider: continuous and discrete, pose different practical challenges in how their outputs should be used in theory-testing models, the underlying logic in the method I propose applies to both. 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 values of the latent variable, θ_i based on the observations θ_i^*. The posterior estimates for θ_i from the measurement model are then used as data in the theory-testing

model $y_i \sim f(\cdot)$.

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

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 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. Applied researchers likely have much more knowledge about what their preferred theory-testing model looks like, rather than the intricacies involved in estimating latent variables. Measurement models can often be challenging to write in practice due to issues of identifiability.

My method overcomes the two problems outlined above by simplifying the measurement model step, $\theta_i^* \sim g(\cdot)$ in the joint model. Rather than estimating the latent variable from scratch, I take the uncertainty estimates already provided from the measurement models and use those as approximations in the full joint model. In the case of a categorical latent variable, such as race/ethnicity in BSIG, the measurement function is $\theta_i^* \sim \text{categorical}(\mathbf{p})$ where $\mathbf{p}$ is the simplex of race/ethnicity

probabilities given by the algorithm. And for continuous latent variables (e.g. Bayesian IRT ideal points), the measurement function is $\theta_i^* \sim \text{Normal}(\mu_i, \sigma_i)$ where μ_i and σ_i correspond to the mean and standard deviation of the posterior distribution for each ideal point. 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.

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.[2]

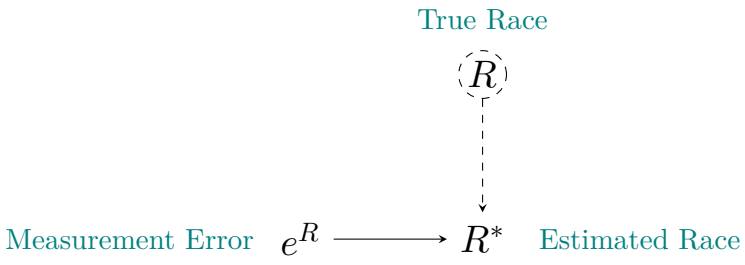
2.1 Measurement Error Confounding Bias

Figure 2 shows a simple DAG outlining the causal process implied by the BSIG race/ethnicity measurement model. The estimated race produced by the model, R^* is a function of an individual's "true" race, R and measurement error, e^R. The notion of a "true" race is somewhat misleading, however, because race is typically understood to be mutable and socially contingent—rather than biologically innate (Omi and Winant 2014; Fields and Fields 2014). The BSIG algorithm's training data come from the US Census, so what it is truly measuring is racial self-identification while filling out a Census survey. Self-reported race may differ in important ways from other conceptions of race, such as those observed or ascribed by others (Saperstein 2006). Ontological considerations of race aside, the important takeaway from Figure 2 is that the race variable produced by BSIG arises causally from a combination of some latent "true" race (which we do not observe), and from measurement error (which we observe, at least partially, in the form of the race probability simplex).

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

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

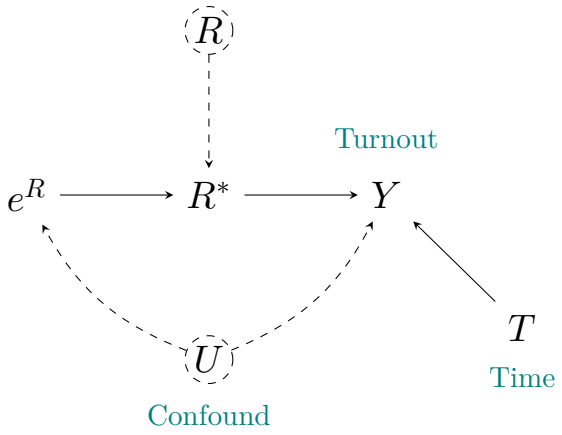
Figure 2: BSIG Measurement Error Model



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.

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

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



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

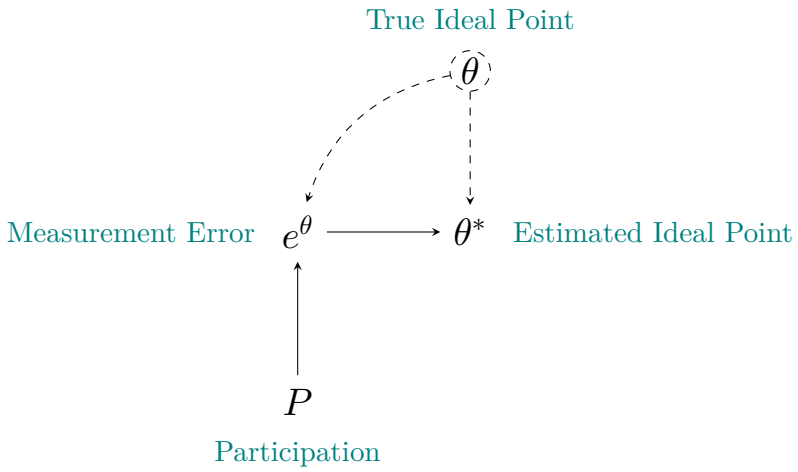
The previous discussion of BSIG race estimates highlighted how failing to account for measurement error could lead to bias via backdoor confounding. This general issue is known as nonrandom, or unignorable, measurement error (Blalock 1970). In the political science methodology literature, methods such 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

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

Figure 4: Bayesian IRT Measurement Model



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.

While there may exist some backdoor paths through e^{θ} and P in a hypothetical theory-testing

Figure 5: Bayesian IRT Posterior Distributions

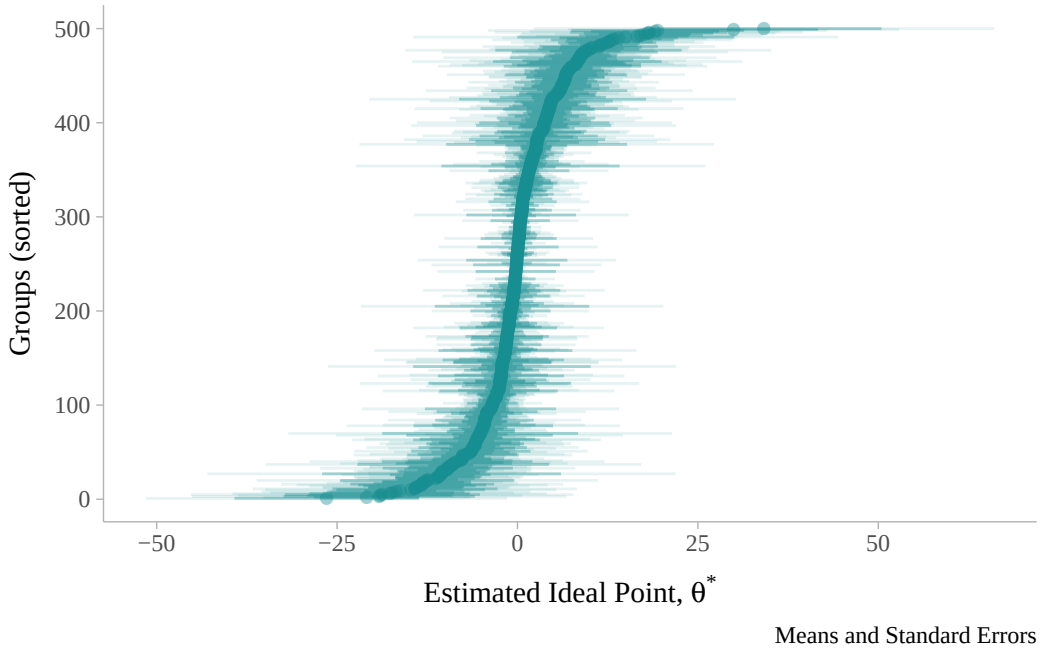
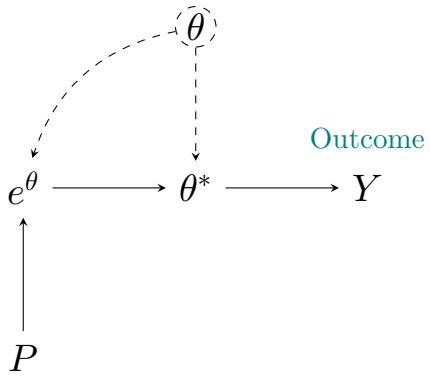


Figure 6: Bayesian IRT Measurement Model in Hypothetical Theory-Testing Model



model, I will assume that the outcome of interest, Y is unaffected by anything other than the direct causal effect $\theta^* \longrightarrow Y$.[3] The purpose of this simplification is to highlight the consequences of random 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 2 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 3, which models $X_{\text{MEAS},i}$ as an outcome of $X_{\text{TRUE},i}$ (an unobserved parameter for each observation)[4] 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.

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

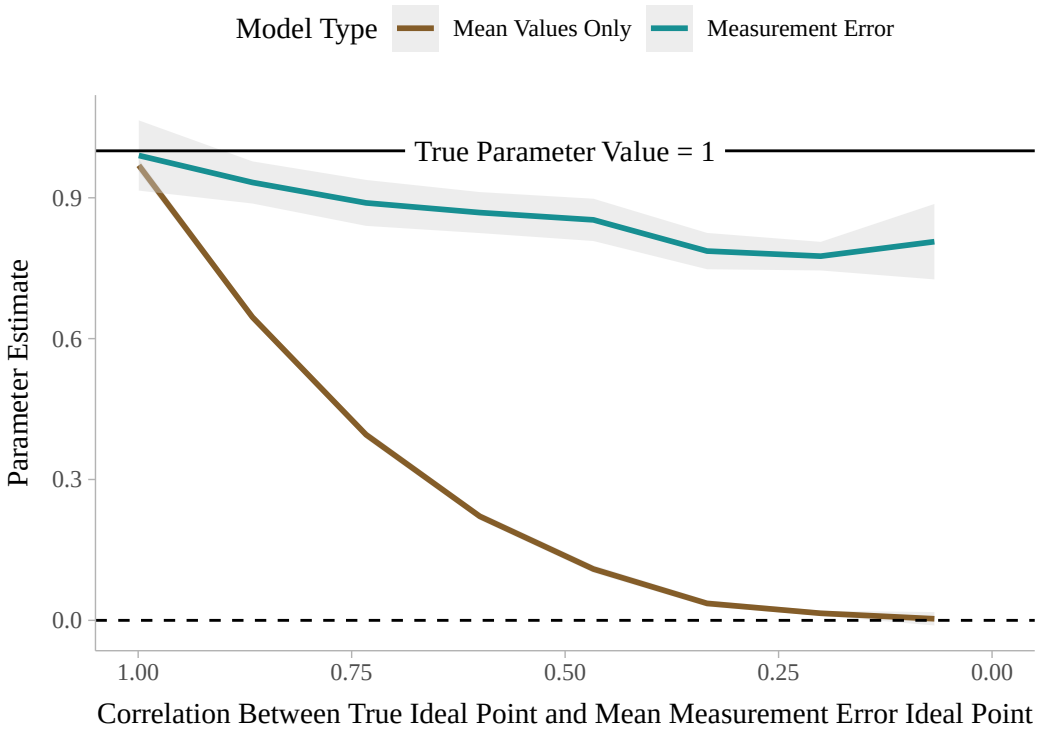
[3] In principle, we could close any backdoor paths through P by conditioning on it directly because the level of group participation is directly observed.

[4] 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.

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

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 2 rapidly attenuate towards zero. This is in contrast to the estimates from the joint measurement theory-testing model in Equation 3 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



[*I am still working on how to code up the model using the categorical measurement error. But once I get it done I hope to use it to show how this method can be used to deal with non-random, confounding, measurement error I discuss in the BSIG section*]

References

Abdelal, Rawi, Yoshiko M. Herrera, Alastair Iain Johnston, and Rose McDermott. 2006. "Identity as a Variable." *Perspectives on Politics* 4 (04). https://doi.org/10.1017/S1537592706060440.

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

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.

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.

Fields, Karen E., and Barbara Jeanne Fields. 2014. *Racecraft: The Soul of Inequality in American*

Life. London: Verso.

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.

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.

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

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.

Saperstein, A. 2006. "Double-Checking the Race Box: Examining Inconsistency Between Survey Measures of Observed and Self-Reported Race." *Social Forces* 85 (1): 57–74. https://doi.org/10.1353/sof.2006.0141.

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. 2022. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. *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/.