Article

Sociological Methods & Research
2017, Vol. 46(3) 303-341
© The Author(s) 2015
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0049124115585360
journals.sagepub.com/home/smr

A Unified Approach to Measurement Error and Missing Data: Overview and Applications

Matthew Blackwell[1], James Honaker[1] and Gary King[1]

Abstract

Although social scientists devote considerable effort to mitigating measurement error during data collection, they often ignore the issue during data analysis. And although many statistical methods have been proposed for reducing measurement error-induced biases, few have been widely used because of implausible assumptions, high levels of model dependence, difficult computation, or inapplicability with multiple mismeasured variables. We develop an easy-to-use alternative without these problems; it generalizes the popular multiple imputation (MI) framework by treating missing data problems as a limiting special case of extreme measurement error and corrects for both. Like MI, the proposed framework is a simple two-step procedure, so that in the second step researchers can use whatever statistical method they would have if there had been no problem in the first place. We also offer empirical illustrations, open source software that implements all the methods described herein, and a companion article with technical details and extensions.

[1] Institute for Quantitative Social Science, Harvard University, Cambridge, MA, USA

Corresponding Author:
Gary King, Harvard University, Institute for Quantitative Social Science, 1737 Cambridge Street, Cambridge, MA 02138, USA.
Email: king@harvard.edu

Keywords

measurement error, missing data, modeling, inference, selection

Introduction

Social scientists routinely recognize the problem of measurement error in the context of data collection, but often choose to ignore it during their subsequent statistical analyses. In the most optimistic scenario, the bias induced by measurement error may be ignored if it is judged to be smaller than the effects being estimated. Some researchers act as if the analyses of variables with measurement error will still be correct on average, but this is untrue; others act as if the attenuation that occurs in simple types of random measurement error with a single explanatory variable holds more generally, but this too is incorrect. Sophisticated application-specific methods for handling measurement error exist, but they can be complicated to implement, require difficult-to-satisfy assumptions, or lead to high levels of model dependence; few such methods apply when error is present in more than one variable and none are widely used in applications, despite an active methodological literature. The corrections used most often are the easiest to implement but typically also require the strongest assumptions, about which more will be said subsequently (see Guolo 2008 and Stefanski 2000 for literature reviews).

We address here the challenge of creating an easy-to-use but more generally applicable method of dealing with measurement error. Our goal is to contribute to the *applied* statistics literature, to offer a statistically robust methodology that can be used for a wide range of applications. We do this through a unified approach to correcting for problems of measurement error and missing data in a single easy-to-use procedure. We extend multiple imputation (MI) for missing data to also accomodate measurement error (Cole, Chu, and Greenland 2006; Rubin 1987) and so that it handles an array of common social science analyses. We treat measurement error as partially missing information and completely missing values as an extreme form of measurement error. The proposed approach, which we call *multiple overimputation* (MO), enables researchers to treat data values as either observed without error, observed with (conditionally random) error, or missing. We accomplish this by constructing distributions for individual observations (or entire variables) with means equal to the observed values, if any, and variance for the three data types set to zero, a (chosen or estimated) positive real number, or infinity, respectively.

Like MI, MO requires two easy steps. First, analysts create multiple ($\approx$5) data sets by drawing missing and mismeasured values from their

posterior predictive distribution conditional on all available observation-level information. This procedure leaves the observed data constant across the data sets, imputes the missing values from their predictive posterior, and "overimputes," that is, overwrites the values or variables measured with error with draws from their predictive posterior, but informed by the observed measurement, other variables, and available assumptions. An especially attractive advantage of MO (like MI) is the second step, which enables analysts to run whatever statistical procedure they would have run on the completed data sets, as if all the data had been correctly observed. A simple procedure is then used to average the results from the separate analyses. The combination of the two steps enables scholars to overimpute their data set once and to then set aside the problems of missing data and measurement error for subsequent analyses.

As a companion to this article, we have modified a widely used MI software package known as "Amelia II: A Program for Missing Data," to also perform MO, and many related extensions (Honaker, King, and Blackwell 2010). Our basic approach to measurement error allows for random measurement error in any number of variables, or some values within some variables, in a data set. By building on the insights and procedures in MI, MO also inherits the attractive properties already proven in the extensive missing data literature. We also offer a companion article (Blackwell, Honaker, and King 2017, hereinafter BHK2), which gives mathematical details of the methodology, evidence of how many data sets need to be created, and when the technique is robust to error that is correlated with the dependent variable and the latent variable itself. We also show there that MO can be extended to handle heteroskedastic measurement error, and works well with categorical variables.

The second section describes our proposed framework in the context of multiple variables measured with random error. There, we generalize the MI framework, prove that a fast existing algorithm can be used to create imputations for MO, and offer Monte Carlo evidence that it works as designed. The third section goes further by deriving methods of estimating the measurement error variance so it need not be assumed. The fourth section provides a checklist of practical guidance for applying this technique in practice. The fifth section then offers three empirical illustrations and the sixth section concludes.

The MO Framework

To build intuition, we conceptualize the linkage between measurement error and missing data in two equivalent ways. In one, measurement error is a type

of missing data problem where observed proxy variables provide probabilistic prior information about the true unobserved values. In the other, missing values have an extreme form of measurement error where no such information exists. More simply, measurement error can be seen as a mitigated form of missing data, or conversely missing data can be seen as limiting special case of measurement error. Either way, they are linked methodological problems whose treatments go well together because variables measured with some degree of error fall logically between the extremes of observed without error and completely unobserved. This dual conceptualization also means that our MO approach to measurement error retains the advantages of MI in ease of use, and treatment for measurement error can be taken at the same time as treatment for missingness, which is often already seen as a necessary step in the analysis. Indeed, the same single run of software for one will now solve both problems.

The validity of the approach is also easy to understand within this framework. Consider the following thought experiment. Some small number of observations of a variable are known to be measured with error, so a researcher decides to discard those values and treat them as missing, but keep the gold standard or perfectly measured observations in the analysis. Under the same assumptions as MI for missing data (that the process is missing at random [MAR]), deleting data values with measurement error and using MI introduces no biases. However, this is obviously inefficient, as we know the mismeasured observations provide considerable information about the value's true location. When measurement error is relatively small, the true latent value will be close to the mismeasured value, where the relative meaning of "close" is determined by the degree of measurement error. Our goal is to correctly incorporate that information into the model, which we accomplish by running MI while also using observed values to help inform cell-level priors.[1]

Assume the value w_i is a combination of the true latent value we would like to record, x_i^*, and some degree of measurement error u_i drawn independently from some distribution. For example, if that distribution is normal with variance σ_u^2 we have:

$$\begin{array}{ccccc} & \text{observed} & \text{latent} & \text{measurement error} & \\ & \searrow & \downarrow & \swarrow & \\ w_i & = & x_i^* & + & u_i \\ & & & & \\ u_i|x_i^* & \sim & \mathcal{N}(0,\sigma_u^2) & \leftarrow & \text{measurement error variance} \end{array} \tag{1}$$

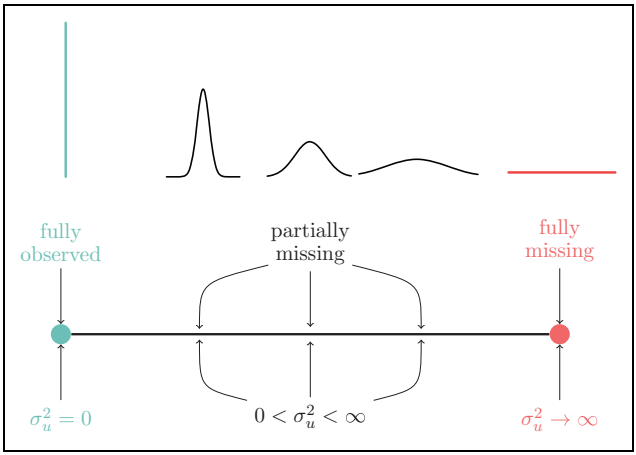

Figure 1. The continuum of measurement error, with observation-level priors illustrated in the top row.

Equation (1) represents the amount of measurement error, under this simple model, as σ_u^2. We can visualize the set of all σ_u^2 values as a line in Figure 1, with $\sigma_u^2 = 0$ at the left endpoint of this line, denoted in blue. At this extreme, we observe the latent variable perfectly. Above the σ_u^2 continuum, we sketch the distribution of the unobserved true value, conditional on the observed value. When $\sigma_u^2 = 0$, there is no uncertainty about the location of the true value, so our belief about the true value, x_i^*, is simply a spike at the observed location, w_i.

As measurement error begins to increase, σ_u^2 moves to the right in this continuum. Now we no longer know with certainty the location of the latent value from the observed value, but with small error distributions this distribution is still very tight and as σ_u^2 becomes larger the distribution has increased variance. At the extreme, as $\sigma_u^2 \to \infty$ the distribution of the latent variable becomes flat and we have no information about where the latent value is located from the mismeasured value. At this point, the observed value itself is entirely uninformative for the latent value, and thus it is a missing value in the data set. Missing data is thus merely the extreme limiting form of measurement error. Without a validation sample, classical missing data methods can only deal with values at $\sigma_u^2 = 0$ or $\sigma_u^2 \to \infty$, the red and the blue distributions, because they are unable to incorporate the additional information contained in the observed data when $0 < \sigma_u^2 < \infty$ into the missing data algorithm, even if we can estimate this variance from the data. Thus,

the MI framework forces us to think in terms of a false dichotomy between values that are fully observed and values that are fully missing, when a whole continuum of measurement error is possible.[2]

To move past the restrictions of MI, we add cell-level priors that incorporate the fact that the latent locations of some data values are neither spikes nor flat slabs, but (proper) distributions. This procedure combines two different sources of information about the unobserved true value of a given data point. First, if we rely only on the mismeasured observation, we have our prior that describes the latent value of that cell as a distribution, as shown in the top row of Figure 1. Second, if we treat that cell as if it were missing data, MI would construct a distribution conditional on all the rest of the observed data in the data set, from which we would classically draw random imputations. We combine these two approaches and jointly use the information from the mismeasured observation itself and the patterns in the rest of the observed data. Without measurement error in an observation, it degenerates to a spike and the observation remains unchanged. When measurement error is so extreme that we encounter a missing value, the prior is flat and the missing value is imputed solely from the observed data. However, now we can treat all the values with intermediate measurement error—neither $\sigma_u^2 = 0$ nor $\sigma_u^2 \to \infty$—by replacing the mismeasured value with a series of draws from the posterior for the true latent value. We call this framework MO. In the rest of this section, we provide a more technical explanation of MO as well as the MI foundation we generalize from; readers may skip this material, while those looking for even more specifics on the modeling assumptions and implementation details should also see BHK2, Section-Model and Estimation.

The Foundation: An MI Model

MO builds on MI, which we now review. MI involves two steps. First, we use a model to generate multiple, $B \approx 5$, imputations for each of the missing values in the data set. The imputations are predictions from a model that uses all the observed data available (we describe this model subsequently). Then we make B copies of our data set with all the observed cell values identical in every copy, and the imputations for the missing cells varying. When our imputation model indicates that we can predict a missing cell value well, that cell value's imputation doesn't vary much over the B data sets; model predictions that are more uncertain reflect this uncertainty by varying more. In this way, both our predictive ability (of our data and the model) and a fair characterization of its uncertainty are reflected in the completed data sets.

Then for each of the B completed data sets, the analyst runs whatever statistical method they would have if the data were fully observed, and one of two simple procedures is used to combine the results from the separate analyses. We first describe how to combine the separate analyses and then return to the model that generates the imputations and some more detail about estimation.

Combining Rules. For the first method, consider some quantity of interest, Q such as a first difference, risk ratio, probability, and so on. Let $q_1, \ldots, q_B$ denote the separate estimates of Q, which come from applying the same analysis model to each of the overimputed data sets. The overall point estimate $\bar{q}$ of Q is simply the average $\bar{q} = \frac{1}{B}\sum_{b=1}^{B} q_b$. As shown by Rubin (1978), an estimate of the variance of the MO point estimate is the average of the estimated variances from within each completed data set, plus the sample variance in the point estimates across the data sets (multiplied by a factor that corrects for bias because $B < \infty$): $\bar{s}^2 = \frac{1}{B}\sum_{b=1}^{B} s_b^2 + S_b^2(1 + 1/B)$, where s_b is the standard error of the estimate of q_b from the analysis of data set b and $S_b^2 = \sum_{b=1}^{B} (q_b - \bar{q})/(B - 1)$.

A second procedure for combining estimates is useful when simulating quantities of interest, as in King, Tomz, and Wittenberg (2000) and Imai, King, and Lau (2008). To draw B simulations of the quantity of interest, we merely draw $1/B$ of the needed simulations from each of the overimputed data sets. Then we would treat the set of all these simulations as we would if they were all coming from the same model, such as taking the mean and standard deviation as a point estimate and standard error, respectively, or plotting a histogram as an estimate of the posterior distribution.

Imputation Model. For expository simplicity, consider a simple special case with only two variables, y_i and x_i ($i = 1, \ldots, n$), where only x_i contains some missing values (BHK2 provides further details). These variables are not necessarily dependent and independent variables, as they each play any role in the subsequent analysis model. The analysis of this section applies to any number of variables and with missingness in any or all of the variables (Honaker and King 2010).

We now write down a common model that could be used to apply to the data if they were complete and then afterward explain how to use it to impute any missing data scattered through the input variables. This model assumes that the joint distribution of y_i and x_i, $p(y_i, x_i|\mu, \Sigma)$, is multivariate normal:

$$(y_i, x_i) \sim \mathcal{N}(\mu, \Sigma), \quad \mu = (\mu_y, \mu_x), \quad \Sigma = \begin{pmatrix} \sigma_y^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_x^2 \end{pmatrix}, \tag{2}$$

where the elements of the mean vector μ and variance matrix Σ are constant over the observations. This model is deceptively simple: As there is no i subscript on the scalar means μ_x and μ_y, it may appear as though only the marginal means are used to generate imputations. In fact, its joint distribution implies that a prediction is always based on a regression (the conditional expectation) of that one variable on *all* the others, with the population values of the coefficients in the regression being a deterministic function of μ and Σ. This is extremely useful in missing data problems for predicting a missing value conditional on observed values. For instance, given model (2), the conditional expectation of x_i given y_i is a regression $E[x_i|y_i] = \gamma_0 + \gamma_1(y_i - \mu_y)$, where $\gamma_0 = \mu_x$ and $\gamma_1 = \sigma_{xy}/\sigma_x$. The conditional expectation of y_i given x_i is also a regression, and both regressions are implied by, with parameters directly calculable from, μ and Σ in equation (2).

Researchers have repeatedly demonstrated that this imputation model gives similar point estimates and provides adequate confidence interval coverage compared to complicated nonlinear and non-normal alternatives even for ordinal or categorical variables, and even when more sophisticated models are preferred at the analysis stage (Bernaards, Belin, and Schafer 2007; King et al. 2001; K. J. Lee and Carlin 2010; Rubin and Schenker 1986; Schafer 1997; Schafer et al. 1996). And so even though the model appears very simple, it is indeed very powerful and generalizes to large numbers of variables.

To estimate the regression of each variable in turn on all the others, we only need to estimate the elements of μ and Σ. If no data were missing, the results would be equivalent to running each of the separate regressions (y_i on x_i and x_i on y_i). But how can we run either of these regressions with variables containing arbitrary patterns of missing data? The trick is to find a single set of estimates of μ and Σ from data with scattered missingness and then to use these to deterministically compute the coefficients of all the separate regressions.

Estimation. The "complete-data" likelihood (i.e., still assuming no missing data) is simply the product of model (2) over the n observations:

$$\mathcal{L}(\theta|y, x) \propto \prod_i p(y_i, x_i|\theta) \tag{3}$$

$$= \prod_i p(x_i|y_i, \theta) p(y_i|\theta), \tag{4}$$

where $\theta = (\mu, \Sigma)$. (We use variables without an i subscript to denote the vector of observations, so $y = (y_1, \ldots, y_n)$.) This likelihood is not usable as is, because it is a function of the missing data, which we do not observe. Thus, we integrate out whatever missing values happen to exist for each observation to produce the actual (observed-data) likelihood:

$$\mathcal{L}\left(\theta|y, x^{\text{obs}}\right) \propto \prod_i \int p(x_i|y_i, \theta)p(y_i|\theta)\mathrm{d}x^{\text{mis}} \tag{5}$$

$$= \prod_{i \in x^{\text{mis}}} p(y_i|\theta) \prod_{j \in x^{\text{obs}}} p\left(x_j|y_j, \theta\right)p\left(y_j|\theta\right), \tag{6}$$

where x^{obs} denotes the set of values in x that are observed and x^{mis}, the set that are missing. That we partition the complete data in this way is justified by the standard "MAR" assumption that the missing values may depend on observed values in the data matrix but not on unobservables (Rubin 1976; Schafer 1997). The key advantage of this expression is that it appropriately assumes that we only see what is actually observed, x^{obs} and y, but still estimate μ and Σ.[3]

This result enables one to take a large data matrix with scattered missingness across any or all variables and impute missing values based on the regression of each variable on all of the others. The actual imputations are based on the regression predicted values, their estimation uncertainty (due to the fact that μ and Σ, and thus the calculated coefficients of the regression, are unknown), and the fundamental uncertainty (as represented in the multivariate normal in equation (2) or, equivalently, the regression error term from each conditional expectation). MI works by imputing as few as five values for each missing data point (or more for data sets with unusually high missingness), creating "completed" data sets for each, running whatever analysis model we would have run on each completed data set as if there were no missing values, and averaging the results using a simple set of rules (see previous Subsection-Combining Rules). The assumption necessary for most implementations of MI to work properly is that the missing values are MAR. This is considerably less restrictive than, for example, the "missing completely at random" assumption required to avoid bias in listwise deletion, which is equivalent to assuming that missingness is determined by only random coin flips.

Incorporating Measurement Error

We begin with some simple notation. First, define the data matrix as d, with representative element d_{ij}, as including all the (dependent and independent) variables from the analysis stage. It may also include other variables not to be

explicitly used in the analysis stage but which help improve the imputations. For our simple running example, $d = \{x, y\}$. All the cell values of d exist, but the extent to which we observe them is "assigned" (by the data generating process) to one of three types denoted by the variable m_{ij}. In type $m_{ij} = 0$, d_{ij} is fully *observed*. In type $m_{ij} = 1$, the true value d_{ij} exists but in our data it is *missing and measured with error* using an available unbiased proxy w_{ij}. And in type $m_{ij} = 2$, the true value d_{ij} is *missing* entirely and no proxy exists. Thus, m can be thought of as the assignment of a "measurement mechanism" to each cell of the data matrix.

To this notation, we add two assumptions. First is the *Ignorable Measurement Mechanism Assignment* (IMMA). IMMA says simply that the value of m can be created or influenced by a random draw from a probability distribution, by observed values in d, or by the (observed) values of the proxy values w, but *not* by unobserved cell values in d. This is an optimistic assumption since the unknown assignment mechanism is assumed to be a function of only objects we know, the rest being "ignorable." (For a formal version of this assumption, see BHK2, Subsection- Assumptions). In the fourth section, we offer practical advice to help ensure that IMMA holds in applications.

IMMA can be understood as a direct generalization of MI's MAR assumption, applying to the three categories of m rather than only the two categories of missing or observed used in MI. In fact, since MAR does not use the proxy variables, an approach that may meet the MAR assumption in the presence of measurement error is to simply ignore the proxy variable values and to treat any cell measured with error as fully missing; that is, we reduce the three-category measurement mechanism variable to two categories: $m_{ij} = 0$ and $m_{ij} > 0$. Of course, this will usually represent a complete waste of all the information in the proxy variable w. (Even when the stronger MAR assumption holds, BHK2, Subsection- Robustness to Violating Assumptions, shows that MO outperforms MI due to the information it adds.)

Our second assumption is a choice of a specification for the measurement model that generates the proxy values w_{ij}. For example, one possible choice of a data generation process for w is random normal measurement error around the true value, $w_i \sim \mathcal{N}\left(x_i^*, \sigma_u^2\right)$, with σ_u^2 set to a chosen or estimated value (we discuss interpretation and estimation of σ_u^2 in the third section). The value of σ_u^2 places the observation on the continuum in Figure 1.

Other possible choices for this assumption allow for heteroskedastic measurement error, such as might occur with gross domestic product from a country where a government's statistical office is professionalizing over time; mortality statistics from countries with and without death registration systems; or survey responses from a self-report versus elicited about that

person from someone else in the same household. In the Social Ties and Opinion Formation subsection, we investigate a situation with known heteroskedastic measurement error: Variables that are aggregations of different amounts of randomly selected individual-level data where the sample size completely determines the degree of measurement error. (We also discuss possibilities for heteroskedastic measurement error more generally in BHK2, Subsection- Heteroskedastic Measurement Error).

The choice for this assumption can include biased measurement error, where $E\left[w_i|x_i^*\right] = a_i + x_i^*$, so long as the bias, a_i, is known or estimable. For instance, if validation data are available, a researcher could estimate the bias of the measure or use a model to estimate how the offset changes with observed variables. From our perspective, an observation that does not possess at least this minimally known set of relationships to its true value could more easily be considered a new observation of a different variable rather than a proxy for an unobserved one. Another way of stating this is to say that any statistical method which uses a proxy value w_{ij} to measure something different (i.e., w_{ij}) requires an assumption of some kind. This type of external information is a requirement of every proper statistical approach to measurement error (Stefanski 2000). (In the extreme situation when the bias is not known and cannot be estimated, we can sometimes narrow inferences to a bounded range rather than a point estimate and use sensitivity analysis, or "robust Bayes," to make further progress; see Berger 1994; King and Zeng 2002.)

Implementation

Honaker and King (2010) propose a fast and computationally robust MI algorithm that allows for informative Bayesian priors on individual missing values. The algorithm is known as EMB, or EM with bootstrapping. They use this algorithm to incorporate qualitative case-specific information about missing cells to improve imputations. To make it easy to implement our approach, we prove in BHK2, Section- Model and Estimation, that the same algorithm can be used to estimate our model. The statistical duality property assumed there enables us to turn the data generation process for w_i into a prior on the unobserved value x_i^*, without changing the mathematical form of the density.[4] For example, in the simple random normal error case, the data generation process for w_i is $\mathcal{N}\left(w_i|x_i^*, \sigma_u^2\right)$ but, using the property of statistical duality of the normal, this is equivalent to a prior density for the unobserved x_i^*, $\mathcal{N}\left(x_i^*|w_i, \sigma_u^2\right)$.[5]

This strategy also offers important intuitions: We can interpret our approach as treating the proxy variables as informative, observation-level

prior means for the unobserved missing values. Our imputations of the missing values, then, will be precision-weighted combinations of the proxy variable and the predicted value from the conditional expectation (the regression of each variable on all others) using the missing data model. In addition, the parameters of this conditional expectation (computed from μ and Σ) are informed and updated by the priors on the individual values.

Under our approach, then, all values in the data matrix with measurement error are replaced—overwritten in the data set, or *overimputed* in our terminology—with MOs that reflect our best guess and uncertainty in the location of the latent values of interest, x_i^*. These overimputations include the information from our measurement error model, or equivalently the prior with mean set to the observed proxy variable measured with error, as well as all predictive information available in the observed variables in the data matrix. As part of the process, all missing values are imputed as usual with MI. The same procedure is used to fill in multiple completed data sets; usually about 10 to 25 data sets are sufficient, but more may be necessary with large fractions of missing values or high degrees of measurement error. Imputations and overimputations vary across the multiple completed data sets—with more variation when the predictive ability of the model is smaller and measurement error is greater—while correctly observed values remain constant.

Researchers create a collection of completed data sets once and then run as many analyses of these as desired. The same analysis model is applied to each of the completed (imputed and overimputed) data sets as if it were fully observed. A key point is that the analysis model need not be linear-normal even though the model for missing values and measurement error overimputation is (Meng 1994). The researcher then applies the usual MI rules for combining these results (see previous Subsection Combining Rules).

Monte Carlo Evidence

We now offer Monte Carlo evidence for MO, using a data generation process that would be difficult for most prior approaches to measurement error. We use two mismeasured variables, a non-normal dependent variable, scattered missing data, and a nonlinear analysis model. The measurement error accounts for 25 percent of the total variance for each proxy, meaning these are reasonably noisy measures. In doing so, we attempt to recreate a difficult but realistic social science data analysis, with the addition of the true values so we can use them to validate the procedure.

We generated proxies x and z for the true variables x^* and z^*, respectively, using a normal data generation process with the true variables as the mean

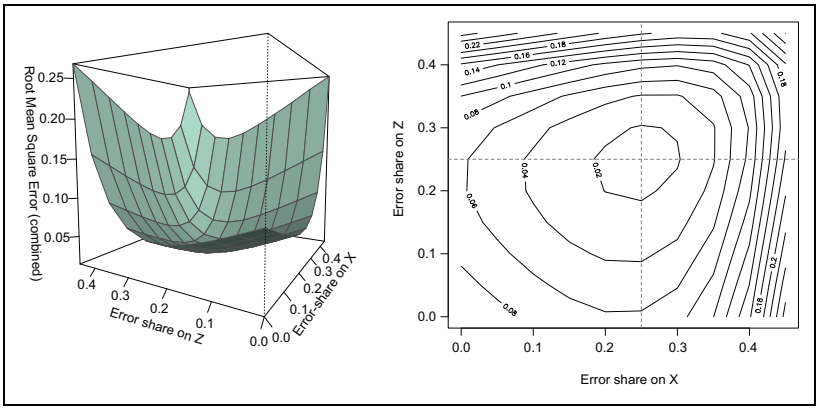

Figure 2. On the left is a perspective plot of the mean square error of a logit analysis model estimates after multiple overimputation (MO) with various assumptions about the measurement error variance. The right shows the same information as a contour plot. Note that the axes here are the share of the observed variance due to measurement error which has a true value of 0.25, which is where the mean square error (MSE) reaches a minimum.

and a variance equal to $\sigma_u^2 = \sigma_v^2 = 0.5$.[6] To mimic real applications, we run MO with 20 imputed data sets under various (sometimes incorrect) assumptions about the error variances. At each combination of σ_u^2 and σ_v^2, we calculate the mean square error (MSE) for the logit coefficients of the overimputed latent variables, using the standard rules for combining MI results and repeated this whole process for 10,000 simulations. We took the average MSE across these coefficients and present the results in Figure 2. On the left is the MSE surface with the error variances on the axes along the floor and MSE on the vertical axis; the right graph shows the same information viewed from the top as a contour plot.

The figure shows that when we assume the absence of measurement error (i.e., $\sigma_u^2 = \sigma_v^2 = 0$), as most researchers do, we are left with high MSE values. As the assumed amount of measurement error grows, we see that the MO lowers the MSE smoothly. The MSE reaches a minimum at the true value of the measurement error variance (the gray dotted lines in the contour plot).[7] Assuming values that are much too high also leads to larger MSEs, but the figure reveals that MO can improve MSE even when the measurement error variance is not precisely known. We discuss this issue further subsequently.

In results shown in table 6 of BHK2, we also find that there are benefits to increasing the number of imputations, in terms of both MSE and confidence

interval coverage, but that these benefits are relatively minor after 10 imputed data sets. Indeed, at the correct value of the measurement error variance, the MO estimator achieved close to nominal coverage for 95 percent confidence intervals on the logit coefficients. Of course, these results are one simulation and different choices of parameters can lead to different performance, but we attempt to augment this with other simulations subsequently and in BHK2. Indeed, BHK2, Section- Robustness to Categorical Variables Measured with Error, shows that naively applying MO to ordered categorical data leads to similar results as here.

Comparison to Other Techniques

As measurement error is a core threat to many statistical analyses, many approaches have been proposed. In fact, MI has previously been extended to measurement error in the specific instance where validation subsamples are available. This is when researchers observe both the mismeasurement and the true latent variable for a subset of observations (Brownstone and Valletta 1996; Cole et al. 2006; Guo and Little 2011; Guo, Little, and McConnell 2012; He and Zaslavsky 2009; Wang and Robins 1998). This type of data is relatively rare in the social sciences, but the results hint at what might be possible in our more general approach: In this special data type, the approach outperforms maximum likelihood (Messer and Natarajan 2008) and is robust to measurement error correlated with the dependent variable (Freedman et al. 2008). We build on the insights in this approach and extend it to a wider range of more commonly observed types of data, analyses, and available information.[8]

Other measurement error solutions broadly fall into two camps: general-purpose methods and application-specific methods. General-purpose methods are easily implemented across a wide variety of models, while application-specific methods are closely tailored to a particular context. These approaches use a variety of assumptions that are, in different ways, more and also less restrictive than our approach. See Fuller (1987), Carroll, Ruppert, and Stefanski (1995), and Imai and Yamamoto (2010) for formal definitions and citations.

The first general-purpose method, *regression calibration* (Carroll and Stefanski 1990), is similar in spirit to MO in that it replaces the mismeasured variable with an estimate of the underlying unobserved variable and then performs the desired analysis on this "calibrated data." This estimate is typically in the form of a regression of true, validated data on the mismeasurements, making it similar to a single imputation technique. Cole et al. (2006) compared the performance of regression calibration to that of MI with the

same type of validation subsample and found that MI sometimes outperformed regression calibration and subsequent research has shown that either can outperform the other, depending on the data generating process (Freedman et al. 2008; Messer and Natarajan 2008). As White (2006) points out, both methods rely on validation data, but regression calibration uses conditional means in the analysis step, even when the true data are available. MO combines the best parts of each of these approaches by utilizing all information when it is available and extends their applicability beyond situations with validation samples.

The easiest technique to implement is a simple method-of-moments estimator, which simply corrects a biased estimate of a linear regression coefficient by dividing it by the reliability ratio, $\sigma^2_{x^*}/\sigma^2_w$. This technique depends heavily on the estimate of the measurement error variance and, as shown in our simulations in the third section, has poor properties when this estimate is incorrect. Further, the method-of-moments technique requires the analysis model to be linear.

Other general approaches to measurement error include simulation-extrapolation, or SIMEX (Cook and Stefanski 1994; Hopkins and King 2010), and minimal-assumption bounds (Black, Berger, and Scott 2000; Klepper and Leamer 1984; Leamer 1978). These are both excellent approaches to measurement error, but they both have features that limit their general applicability. SIMEX simulates the effect of adding *additional* measurement error to a single mismeasured variable and then uses these simulations to extrapolate back to the case with no measurement error. With multiple mismeasured variables, SIMEX becomes harder to compute and more dependent on the extrapolation model. The minimal-assumption bounds specify a range of parameter values consistent with a certain set of assumptions on the error model. Bounds typically require fewer assumptions than our MO model, but cannot reveal how estimates change within those bounds. Even if we lack any information on the measurement error variance, we can use MO to perform a sensitivity analysis to quantify the effects of various assumptions about measurement error.

Structural equation modeling (SEM) attempts to alleviate the measurement error by finding latent dimensions that could have generated a host of observed measures.[9] Our goal, however, is to rid a particular variable (or variables) of its measurement error. While discovering and measuring latent concepts is a useful and common task in the social sciences, we often want to measure the effect of a specific variable, and measurement error stands in the way. Without strong structural assumptions, SEM would sweep that variable up into a larger construct and perhaps muddle the question at hand. Thus, MO and SEM tackle different sets of substantive questions.

Furthermore, MO can easily handle gold standard and validation data when it is unclear how to incorporate these into an SEM framework.

Specifying or Estimating the Measurement Error Variance

Under MI, researchers must indicate *which* observations are missing. Under the approach here, researchers must instead indicate or estimate *how much* measurement error exists in each cell in the data set. In this sense, MI is a limiting special case of MO where the only acceptable answers to the "how much" question is all or none. Under both, we can parameterize the information we need in terms of a measurement error variance.

In Section- Directly Estimating Measurement Error Variances of BHK2, we show how to directly estimate this measurement error variance either from the correlation of multiple proxies for the same variable or from the relationship between a variable with measurement error and a small subset of validated or gold standard observations. These are the settings that almost all of the other models for measurement error reviewed in the Comparison to Other Techniques subsection also rely on.

When these extra sources of information are not available, the variance and thus the quantity of interest is not point identified under our approach and others (Stefanski 2000). However, we are able to offer a simple way around the problem. To do this, we reparametrize σ_u^2 to a scale that is easy to understand and then enable researchers to provide uncertainty bounds on the quantity of interest.

The Monte Carlo Evidence subsection shows that using the true measurement error variance σ_u^2 with MO will greatly reduce the bias and MSE relative to the usual procedure of make-believing measurement error does not exist (which we refer to as the "denial" estimator). Moreover, in the simulation presented there (and in others like it), the researcher needs to only have a general sense of the value of these variances to greatly decrease the bias of the estimates. Of course, knowing the value of σ_u^2 (or σ_u) is not always immediately obvious, especially on its given scale. In this section, we deal with this problem by reparameterizing it into a more understandable quantity and then putting bounds on the ultimate quantity of interest.

The alternative parametrization we have found useful is the *proportion of the proxy variable's observed variance due to measurement error*, which we denote by $\rho = \frac{\sigma_u^2}{\sigma_{x^*}^2 + \sigma_u^2} = \frac{\sigma_u^2}{\sigma_w^2}$, where σ_w^2 is the variance of our proxy. This is easy to calculate directly if the proxy is observed for an entire variable (or at least more than one observation). Thus, if we know the extent of the

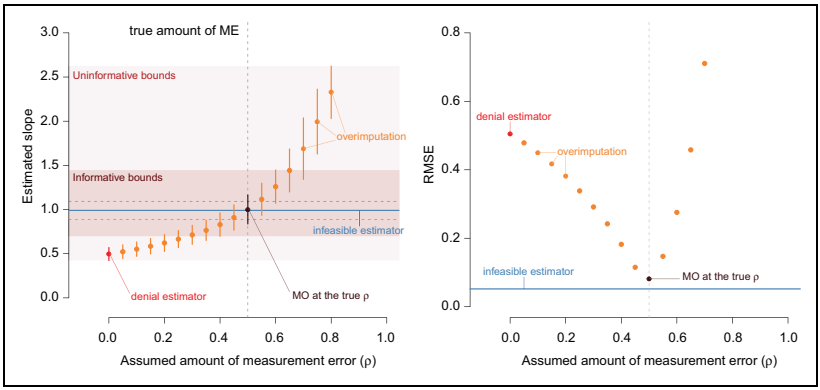

Figure 3. Simulation results using the denial estimator (that assumes no measurement error, in red), the complete-data, infeasible estimator (in blue), and the multiple overimputation (MO) estimator (in orange), with varying assumptions about the degree of mismeasurement. The MO estimator at the correct value of ρ is in dark red. The left panel shows estimates of the coefficients of interest along with confidence bands. In the background, the light tan area shows the minimal-assumption bounds and the dark tan region gives bounds assuming $\rho \in [0.05, 0.6]$. The right panel shows mean square error (MSE) for the same range of estimates.

measurement error, we can create an estimated version of $\hat{\sigma}_u^2 = \rho\hat{\sigma}_w^2$ and substitute it for σ_u^2 in the complete-data likelihood (3).

In Figure 3, we present Monte Carlo simulations of how our method works when we alter our assumptions on the scale of ρ rather than σ_u^2.[10] More importantly, it shows how providing little or no information about the measurement error can bound the quantities of interest. Leamer (1978:238-43) showed that we can use a series of reverse regressions in order to bound the true coefficient without making any assumptions about the amount of measurement error. We compare these "minimal-assumption" bounds to the more model-based MO bounds. The vertical axis in the left panel is the value of the coefficient of a regression of the overimputed w on y. The orange points and vertical lines are the estimates and 95 percent confidence intervals from overimputation as we change our assumption about ρ on the horizontal axis.

We can see that the denial estimator, which treats w as if it were perfectly measured (in red), severely underestimates the effect calculated from the complete data (solid blue horizontal line), as we might expect from the standard attenuation result. As we assume higher levels of ρ with MO, our estimates move smoothly toward the correct inference, hitting it right when

ρ reaches its true value (denoted by the vertical dashed line). Increasing ρ after this point leads to overcorrections, but one needs to have a very bad estimate of ρ to make things worse than the denial estimator. The root MSE leads to a similar conclusion and is thus also minimized at the correct value of ρ.

A crucial feature of MO is that it can be informative even if one has highly limited knowledge of the degree of measurement error. To illustrate this, the left panel of Figure 3 offers two sets of bounds on the quantity of interest, each based on different assumptions about ρ. We use the reverse regression technique of Leamer (1978) to generate minimal-assumption bounds, which make no assumptions about ρ (the mean of these bounds are in light tan). In practice, it would be hard to justify using a variable with more than half of the variance due to measurement error, but even in the extreme situation of 80 percent error, the bounds on the quantity of interest still convey a great deal of information. They indicate, for example, that the denial estimator is an underestimate of the quantity of interest and almost surely within approximately the range [0.5, 1.75]. Note that all of our MO estimates are within these bounds. In simulations in which we lowered the true ρ, we found that even dramatic overestimates of ρ still lead to MO estimates that obey these bounds.[11]

Alternatively, we might consider making a more informative (and reasonable) assumption about ρ. Suppose that we know that there is some positive measurement error, but that less than 70 percent of the observed variance is due to measurement error. These are informative assumptions about ρ and allow MO to estimate bounds on the estimated coefficient. The result is that the bounds shrink (in dark tan, marked "MO-based") closer around the truth. MO thus tells us about how various assumptions about measurement error affect our estimates.[12] The MO-based bounding approach to measurement error shifts the burden from choosing the correct share of measurement error to choosing a range of plausible shares. Researchers may feel comfortable assuming away higher values of ρ since we may legitimately consider a variable with, say, 80 percent measurement error as a different variable entirely. The lower bound on ρ can often be close to 0 in order to allow for small amounts of measurement error.[13]

This figure also highlights the dangers of incorrectly specifying ρ. As we assume that more of the proxy is measurement error, we eventually overshoot the true coefficient and begin to see increased MSE. Note, though, that there is again considerable robustness to incorrectly specifying the prior in this case. Any positive value ρ does better than the naive estimator until we assume that almost 70 percent of the proxy variance is due to error. This result will vary, of course, with the true degree of measurement error and the

model under study. In fact, we show in BHK2, Section- The Number of Datasets to Overimpute that the confidence intervals from MO can be conservative when the measurement error dominates the variance of the latent variable. We do find, though, that the MO estimator is far less sensitive to misspecifications of ρ than the method-of-moments approach, as shown by their respective MSEs. One reason for this is that MO combines information from both the error variance and the observed covariates and so is less dependent on either for performance. Additionally, in BHK2, Section- Robustness to Categorical Variables Measured with Error, we show that these results hold for categorical variables with measurement error.

A Practical Checklist

We offer here a practical, "best practices" checklist that researchers can follow in applying MO to social science analyses. Since the application of MO parallels that of MI, a great deal of intuition follows from the now well-known MI technique. We discuss these but emphasize the unique aspects of MO as well. In the following empirical example, we reference these steps to show how to implement them.

1. *Collect data*. Three types of data are especially useful in using MO to correct measurement error. First, "gold standard data" are measurements of the proxy known to have no error for at least some (known) observations. Researchers can approximate this situation when an expensive high-quality measurement device was used for some subjects or sometimes for measures that improve over time. Second, "validation data" involve both a proxy variable, measured with some error for all units, and measures of the true, latent variable for a subset of data. The units for which we observe both the proxy and the true value constitute what is known as the validation subset. These sources are important because they allow us to identify the relationship between the latent variable and the other covariates and to estimate the variance of the measurement error. Finally, "multiple measures" of the same latent variable can help us estimate the measurement error variance. For example, in survey research, analysts searching for a way to reduce the complexity of the numerous variables available often focus on only the "best" of the available proxies. Instead, a better practice is to use MO, along with *all* available measures of the true but unobserved construct.
2. *Choose variables for the overimputation model.* Including appropriate variables helps us satisfy MO's IMMA assumption. To do this,

three rules are useful to remember. First, as in MI, include all variables to be used in the analysis model (e.g., Meng 1994). Since the overimputation model is linear, any important nonlinearities in the analysis model should also be included in the MO model, such as via squared terms or interactions. Second, include any variables that help predict the true value of the latent variable. For instance, when overimputing income, include available variables that typically correlate highly with income, such as wealth, real estate values, occupation, investments, geographic location, and so on. Similar to MAR in the missing data case, IMMA might hold approximately, but small deviations from the assumption are unlikely to greatly affect estimates. For example, we show in BHK2, Subsection- Robustness to Violating Assumptions, that MO is robust to high levels of correlated measurement error, a clear violation of IMMA. Better predictors can help to ensure this robustness. Third, include variables that will help predict missingness and measurement error even if not used in the analysis model. For example, include variables in the overimputation model even if they would induce post-treatment bias if included in the causal analysis model; these extra variables can create "super-efficient estimates" (i.e., efficiency higher than the highest efficiency in an optimal application-specific analysis model).

3. *Transform non-normal variables*. With normal-based overimputation models, use standard approaches as we would in linear regression. For example, transform skewed variables to approximately symmetric (such as taking the log for income); recode ordered variables to approximately interval; use variance stabilizing transformations (such as a square root for counts).
4. *Select or estimate measurement error variance.* With gold standard data, validation data, or multiple proxies, it is straightforward to estimate the variance of the measurement error (as described in the third section). When none of these data sources are available, we must select the measurement error variance or a range of variances to investigate how quantities of interest depend on this choice.
5. *Choose the number of imputations*. The number of imputations needed to estimate parameters depends on the severity of the measurement error or missing data. Larger measurement error variances require more imputations in order for analysis model estimates to have good properties. When there is gold standard data, one could always ignore the mismeasurements, treat them as missing, and impute them using MI. Thus, in these circumstances, MO will require generally require fewer imputed data sets than MI. Graham, Olchowski,

and Gilreath (2007) recommend a minimum of 10 imputed data sets to avoid a drop in the relative power of imputation estimators. In our own simulations in BHK2, Section- The Number of Datasets to Overimpute, we find for MO that the increases in MSE and confidence interval coverage tend to dissipate after 10 to 25 imputations, with as few as 5 imputations usually being sufficient. Of course, the general principle remains: More missingness or error requires more imputations and so more measurement error also requires more imputations. Since the cost of additional imputations has dropped dramatically with hardware and sofware advances, it is easy to add more imputations if you are unsure.

6. *Run analysis model*. Choose whatever analysis model you would have run if all the data had been observed without error. Apply it to each imputed data set and estimate your quantity of interest. Then combine the estimated quantities of interest using either the so-called Rubin's rules, which average the estimates, or by combining $1/B$ simulations from each of the B models. In situations where Rubin's rules may be suspected of having poor properties (Nielsen 2007; Robins and Wang 2000), our approach makes bootstrap-based inference easy: Simply create a large number of imputations (say, 100) and use the empirical distribution of estimates over these imputations for confidence intervals and statistical tests. Standard analysis software (such as Clarify in Stata or Zelig in R) makes it as easy to run an analysis on B data sets and to combine them as it is to run them on one.
7. *What can go wrong?* With gold standard data, MO's two-step estimation procedure makes it, like MI, highly robust to misspecification, especially compared to structural equation-like approaches. The reason is that imputations only affect the missing or mismeasured values in the data set and leave the observed data untouched. Nevertheless, potential pitfalls include the following. First, using MO, or any measurement error procedure, to deal with very small degrees of measurement error may reduce bias at the expense of a larger increase in variance, although in this situation, the procedure will usually make little difference. Second, when entire variables are measured with large amounts of error, the results will be more model dependent: Just as in MI applied to data with high degrees of missingness, less information in the data has consequences that can only be made up by more data or more assumptions. Of course, making believe there is no measurement error in these situations will normally be considerably worse than using MO. MO inferences will still normally remain within the minimal-assumption bounds we offer, and so users should

be sure to consult the bounds as a check. One can also revert back to MI, and coding all cell values with error as missing, if you believe the information in the proxy is so misleading that it is better to ignore. Third, violations of the key assumptions about measurement error, especially strong correlations between the measurement error and the observed or latent data, can create problems for MO. In fact, if there is gold standard data, MAR holds, and if the measurement error strongly violates our assumptions, it might be better to impute the missing data ignoring the mismeasurements (see BHK2, Subsection- Robustness to Violating Assumptions, for evidence on this point). Fourth, over-imputation is the wrong approach for a certain class of measurement error called Berkson error, where the error is added to the proxy to create the truth rather than to the truth to create the proxy. Finally, theoretical conditions exist under which simple techniques like list-wise deletion or ignoring the problem altogether will be preferred over MO, but these conditions normally make it highly unlikely that one would continue to trust the data for any analyses at that point (King et al. 2001).

Empirical Applications

We now offer three different types of illustrations of the use of MO, applying the checklist in the fourth section to different data sets. First, we study measurement error in surveys. Second, we offer a natural setting where the true value is known but recovered using variables with increasing naturally observed measurement error. And finally, we provide a replication where the level of error, caused by aggregating small samples, can be analytically determined.

The Effect of Political Preferences on Vote Choice

In this section, we apply MO to an extremely common source of measurement error, the responses to public opinion surveys. We follow Ansolabehere, Rodden, and Snyder (2008) and study the causal effect of opinions about economic policy on vote choice. These authors argue that measurement error underestimates the effect of policy preferences on vote choice and use a simple alternative method of removing measurement error, averaging many multiple measures of the same concept.

Although the data requirements mean this approach is not always applicable, it is powerful and requires few assumptions, when feasible. They consider $K = 34$ survey items $\{w_1, w_2, \ldots, w_K\}$, all taken to be imperfect

indicators of an unobserved variable, x, and assume common measurement error variance σ_k^2. That is, $w_{ik} = x_i + u_{ik}$ for each i, where $E[u_{ik}] = 0$ and $E\left[u_{ik}^2\right] = \sigma_k^2$. While any individual measure has variance $\sigma_x^2 + \sigma_k^2$, the average of the measures, $\bar{w}_i = \frac{1}{K}\sum_{k=1}^{K} w_{ik}$ has variance $\sigma_x^2 + \bar{\sigma}^2/K$, where $\bar{\sigma}^2 = \frac{1}{K}\sum_{k=1}^{K} \sigma_k^2$ is the average measurement error variance among the items. If all measures have similar amounts of measurement error, then the average of the items will have far lower levels of measurement error than any single item.

We now show that in the more usual situation where researchers have access to one or only a few measures of their key concepts, MO can still recover reliable estimates because it makes more efficient use of the data and available assumptions. It also avoids the assumption that all available measures are indicators of the same underlying concept. To illustrate these features, we reanalyze Ansolabehere et al. (2008) with their data from the American National Election Survey in 1996. Using their general approach, we find that a one standard deviation increase in economic conservatism leads to a 0.24 increase in the probability of voting for Bob Dole.

To use MO, we start by collecting data in Ansolabehere et al. (2008) from the American National Election Survey in 1996 (Checklist item #1). The key component of the data is the large number of measurements of the same underlying concept—economic policy preferences. This is useful information because it will allow us to calculate the measurement error variance necessary for MO. We then perform MO using only 2 of the 34 variables. To avoid cherry-picking results, we reran the analysis using all possible subsets of two variables chosen from the available 34. For each of these pairs, we overimputed the first variable, using the second as a proxy, along with party identification, ideology, and vote choice to minimize the potential for violating IMMA (Checklist item #2). The second proxy allows us to estimate the amount of measurement error in the first mismeasured variable using the techniques in BHK2, Subsection Multiple Proxies (Checklist item #4). Given the distribution of the data, there was no need to transform the variables (Checklist item #3) and we used 20 imputations (Checklist item #5).

With the overimputations in hand, we then estimate the effect of that overimputed variable on voting for Bob Dole using a probit model (Checklist item #7). We compare this method with simply taking the pairwise averages and using them as the measure of economic policy preferences. These approaches mimic a common situation when social scientists have access to relatively few variables.

Figure 4 shows the relationship between the two estimates. Each column represents the average of the estimated effects for one measure, averaged

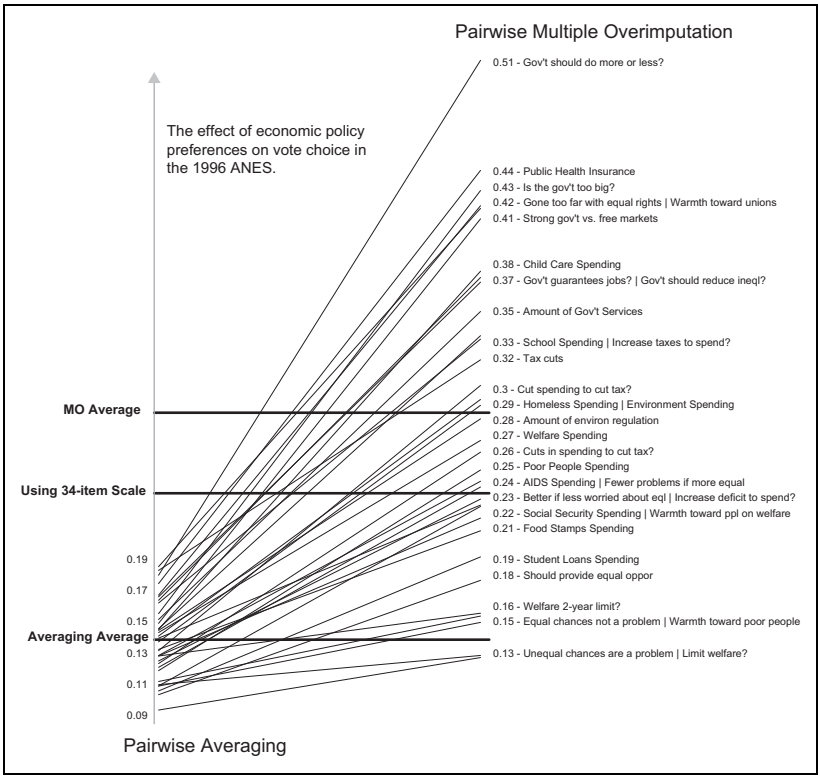

Figure 4. The lines connect estimates from averaging across all pairwise estimates containing the specified variable (left) and estimates from multiple overimputation (MO; right). MO estimates a higher average effect, and one that is closer to the "gold-standard" 34-item scale in each case. Furthermore, MO finds higher estimated effects for classic economic ideology questions and lower effects for questions on welfare and economic opportunity.

across all its pairs. Note that for every variable, MO estimates a larger effect than does averaging, as can be seen by the positive slope of every line. The "gold standard" estimate suggested by Ansolabehere et al. (2008) is well above any of the pairwise averaging estimates, but it lies firmly in the middle of the pairwise MO estimates. This striking result shows that MO makes more efficient use of the available data to correct for measurement error.

While the average results of the pairwise MO align with the 34 measure gold standard, there is considerable variance among the individual measures.

This is in part due to a fundamental difference between MO and averaging (or more general scale construction techniques like factor analysis). MO corrects measurement error on a given variable instead of constructing a new measure of an underlying concept. This often valuable result allows us to investigate how the estimated effect of economic preferences varies across the choice of measure. With pairwise MO, we find that classic economic ideology items regarding the size of government and its role in the economy have a much larger estimated effect on vote choice than questions on welfare policy, equal opportunity, and poor people—all of which were treated the same under averaging. Furthermore, the lowest estimated effects come from variables that relate to views of the poor and their benefits from the government, which in part may be stronger proxies for other issues such as racial politics.

As Ansolabehere et al. (2008) point out, averaging is a "tried and true" method for alleviating measurement error and it works well when many questions exist for a given concept. When, as usual, less information is available, MO may be able to extract more information from the available data.

Unemployment and Presidential Approval

To show a practical example of our MO solution with increasing levels of measurement error, we next construct a measurement error process from a natural source of existing data.

It is often the case, particularly in yearly aggregated cross-national data, that key independent variables are not measured or available at the correct point in time the model requires. Some economic and demographic statistics are only collected at intervals, sometimes as rarely as once every 5 or 10 years. The available data, measured at the wrong period in time, are often used as a reasonable proxy for the variable's value in the desired point in time, with the understanding that there is measurement error which increases the more distant the available data are from the analyst's desired time period.

We mimic this process in actual data by intentionally selecting a covariate at increasing distance in time from the correct location, as a natural demonstration of our method in real data. In our example, we are interested in the relationship between the level of unemployment and the level of Presidential approval in the United States, for which there are rich data of both series over time.[14]

We assume that the correct relationship is approximately contemporaneous. That is, the current level of unemployment is directly related to the President's approval rating. Unemployment moves over time, so the further in time our measure of unemployment is from the present moment, the weaker the proxy for the present level of unemployment, and the more the

measurement error in the available data. We iteratively consider repeated models where the measurement of unemployment we use grows one additional month further from the present time.

We compare this to the most common measurement error model employed in the social sciences, the *errors-in-variables* model (EIV). The EIV approach, reviewed in BHK2, Section- Robustness to Correlated Measurement Errors, relies on the existence of multiple proxies. To naturally create two proxies with increasing levels of measurement error, we use a measure of unemployment k months before the dependent variable and k months after.[15]

We estimate the relationship between unemployment and presidential approval using our MO framework, and the common EIV approach, while using pairs of proxies that are from 1 to 12 months away from the present. We also estimate the relationship between approval and all individual lags and leads of unemployment; these give us all the possible denial estimators, with all the available proxies. In Figure 5, these coefficients from the denial estimators, are shown in red, where the red bar represents the 95 percent confidence interval for the coefficient and the center point the estimated value. The x-axis measures how many months in time the covariate used in the model is from the month of the dependent variable. Positive values of x use proxies that are measured later than desired, negative values are measured too far in the past. The correct, contemporaneous relationship between unemployment and approval is in the center of this series (when x is 0) marked in black.

The EIV estimates are shown in blue. We see that with increased measurement error in the available proxies, the EIV estimates rapidly deteriorate. When the proxies for current unemployment are four months from the value of the dependent variable, the EIV estimates of the relationship are 1.40 times the true value, that is, they are biased by 40 percent. At six months, the confidence interval no longer contains the true value and the bias is 98 percent. With unemployment measured at a one-year gap, EIV returns an estimate 6.5 times the correct value. The MO estimates, however, are comparatively robust across these proxies. The confidence intervals expand gradually as the proxies contain less information and more measurement error. The bias is always moderate, between $+16$ percent and -12 percent and always clearly superior to the denial estimator, until the proxies are fully 12 months distant from the dependent variable. Finally, at one year's distance, the MO estimates are biased by 46 percent, while the denial estimator is biased at -48 percent.[16]

We could do better than shown; we do not propose that this is the best possible model for covariates that are mismeasured in time.[17] Rather, what we have shown in this example is that in naturally occurring data, in a simple research question, where we can witness and control a measurement error

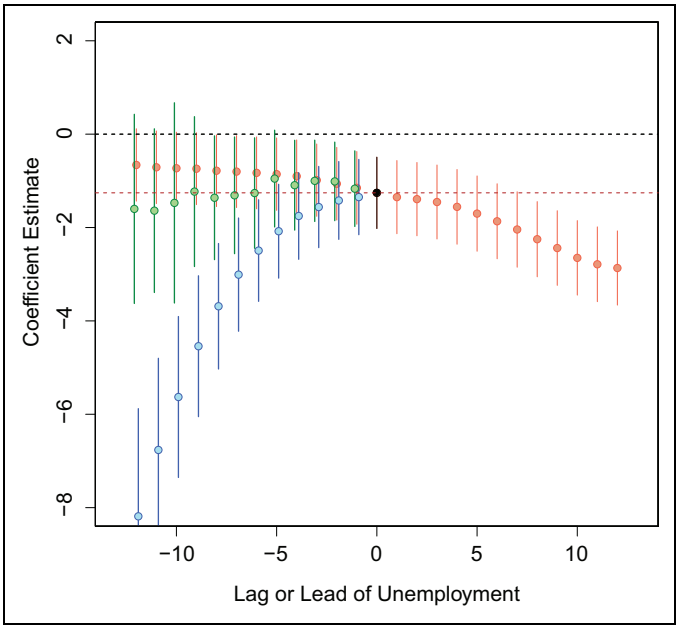

Figure 5. An experiment in measurement error, in the estimation of the relationship between unemployment and Presidential approval, whose true, contemporaneous value is shown in black. The blue confidence intervals represent errors-in-variables (EIV) estimates of this relationship using proxies of unemployment measured increasingly distant in time. The EIV estimates fail quickly, as the proxies move away from month zero. The green estimates show the robust multiple overimputation (MO) estimates of the relationship. These are consistently superior to the red estimates which show the denial estimators using the unemployment rates mismeasured in time, ignoring the measurement error.

process, the most commonly used model for measurement error fails catastrophically, and our framework is highly robust to even a difficult situation with proxies with negatively correlated errors.

Social Ties and Opinion Formation

Having looked at examples where other measurement error methodologies are available, we turn to a conceptually simple example that poses a number of difficult methodological hazards. We examine here the small area estimation challenges faced in the work of Huckfeldt, Plutzer, and Sprague (1993).

The authors are interested in the social ties that shape attitudes on abortion. In particular, they are interested in contrasting how differing networks and contexts, such as the neighborhood one lives in, and the church you participate in, shape political attitudes.

Seventeen neighborhoods were chosen in South Bend, Indiana, and 1,500 individuals randomly sampled across these neighborhoods. This particular analysis is restricted to the set of people who stated they belonged to a church and could name it. The question of interest is what shapes abortion opinions, the individual-level variables common in random survey designs (income, education, party identification), or the social experiences and opinions of the groups and contexts the respondent participates in. Abortion attitudes are measured by a six-point scale summing how many times you respond that abortion should be legal in a set of six scenarios.

The key variable explaining abortion opinion is how liberal or conservative are the attitudes toward abortion at the church or parish to which you belong. This is measured by averaging over the abortion attitudes of *all the other people in the survey* who state they go to the same named church or parish as you mention. Obviously, in a random sample, even geographically localized, this is going to be an average over a small number of respondents. The median number is 6.[18] The number tends to be smaller among Protestants who have typically smaller congregations than Catholics who participate in generally larger parishes. In either case, the church positions are measured with a high degree of measurement error because the sample size within any church is small. This is a classic "small area estimation" problem. Here we know the sample size, mean, and standard deviation of the sampled opinions from within any parish that lead to the construction of each observation of this variable.

This is an example of a variable with measurement error, where there are no other proxies available, but we can analytically calculate the observation-level priors. For any individual, i, if c_i is the set of n_i respondents who belong to i's church (not including i), the priors are given by:

$$p\left(w_i|x_i^*\right) = \mathcal{N}(\bar{c}_i, sd(c_i)/\sqrt{n_i}), \tag{7}$$

where the $sd(c_i)$ can be calculated directly as the standard deviation within a group if n_i is generally large, or we can estimate this with the within-group variance, across all groups, as $\sqrt{1/n\sum_i\left(w_{ij} - \bar{w}_j\right)^2}$.

This is clearly a case where the measurement error is heteroskedastic; different respondents will have different numbers of fellow parishioners included in the survey. Moreover, this degree of measurement error is not

itself random, as Catholics—who tend to have more conservative attitudes toward abortion—are from generally larger parishes, thus their church attitude will be measured with less error than Protestants who will have greater measurement error in their church attitude while being more liberal. The direction of the measurement error is still random, but the variance in the measurement error is correlated with the dependent variable. Furthermore, while we have focused on the measurement error in the church attitude variable, the authors are interested in distinguishing the socializing forces of church and community, and the same small area estimation problem applies to measuring the average abortion position of the community a respondent lives in. Obviously though, the sample size within any of the 17 neighborhoods is much larger than for the parishes and thus the degree of measurement error is smaller in this variable.[19] Finally, as it is survey data, there is a variety of missing data across the variables due to nonresponse. Despite all these complicating factors, this is a set up well suited to our method. The priors are analytically tractable, the heterogeneous nature of the measurement error poses no problems because we set priors individually for every observation, and measurement error across different variables poses no problems because the strength of the MI framework is handling different patterns of missingness.[20]

We replicate the final model in table 2 of Huckfeldt et al. (1993). Our Table 1 shows the results of the naive regression subject to measurement error in the first column. Parish attitudes have no effect on the abortion opinions of churchgoers, but individual-level variables, such as education and party identification and the frequency with which the respondent attends church, predict abortion attitudes. The act of going to church seems to decrease the degree of support for legalized abortion, but the beliefs of the fellow congregants in that church have no social effect or pressure. Interestingly, Catholics appear to be different from non-Catholics, with around a half point less support for abortion on a six-point scale.

The second column applies our model for measurement error, determining the observation-level priors for neighborhood and parish attitudes analytically as a function of the sample of respondents in that neighborhood and parish. Only the complete observations are used in column 2, so differences with the original model are due to corrections of the measurement error in the small area estimates. We see now the effect of social ties. Respondents who go to churches where the support for legal abortion is higher, themselves have greater support for legal abortion. This may be because abortion is a moral issue that can be shaped in the church context and influenced by coreligionists, or this maybe a form of self-selection of

Table 1. Determinants of Abortion Attitudes.

	Naive Regression Model	MO Measurement Only	MO Measurement and Missingness
Constant	3.38** (1.12)	0.34 (1.79)	−0.97 (1.56)
Education	0.17** (0.04)	0.15** (0.04)	0.14** (0.03)
Income	−0.05 (0.05)	−0.05 (0.05)	−0.01 (0.05)
Party ID	−0.10* (0.04)	−0.11* (0.04)	−0.08* (0.04)
Church attendance	−0.57** (0.07)	−0.55** (0.07)	−0.51** (0.06)
Mean neighborhood Attitude	0.11 (0.21)	0.68 (0.44)	0.85* (0.39)
Mean parish Attitude	0.13° (0.07)	0.38* (0.17)	0.42** (0.14)
Catholic	−0.48* (0.27)	−0.26 (0.23)	−0.05 (0.18)
n	521	521	772

Note: Mean parish attitudes are estimated by the average across those other respondents in the survey who attend the same church. These "small area estimates" with small sample size and large standard errors have an analytically calculable measurement error. Without accounting for measurement error there is no discernable effect (column 1) but after applying MO (column 2) to correct for measurement error, we see that the average opinion in a respondent's congregation predicts their own attitude toward abortion. MO = multiple overimputation.
**$p < .01$.
*$p < .05$.
°$p < .10$.

church attendance to churches that agree on the abortion issue. With either interpretation, this tie between the attitudes in the network of the respondent's church and the respondent's own personal attitude disappears due to measurement error caused by the inevitable small samples of parishioners in any individual church.

Of course, our MO approach can simultaneously correct for missing data also, and MI of nonresponse increases by one half the number of observations available in this regression.[21] Most of the same results remain, while the standard errors shrink due to the increase in sample size. Similar to the parish variable, local neighborhood attitudes are now statistically significant at the 95 percent level. The one variable that changes noticeably is the dummy variable for Catholics which is halved in effect and no longer statistically significant once we correct for measurement error, and the rest of the effect disappears when we impute missing data.[22] In all, MO strengthens the author's findings, finds support for their theories in this particular model where previously there was no result, and aligns this regression with the other models presented in their work.

Conclusion

Measurement error is a prevalent, but commonly ignored, problem in the social sciences. The difficulties of use and assumptions have allowed few methods proposed for it to be widely used in practice. We generalize the MI framework to handle observed data measured with error. Our generalization overwrites observed but mismeasured observations with a distribution of values reflecting the best guess and uncertainty in the latent variable. We view missing values as an extreme form of measurement error. However, correctly implementing the MI framework to also handle "partially missing" data, via informative observation-level priors derived from the mismeasured data, allows us to unify the treatment of all levels of measurement error including the case of completely missing values.

This approach enables the rigorous treatment of measurement error across multiple covariates, with heteroskedastic errors, and in the presence of violations of assumptions necessary for common measurement treatments. The model works in survey data and time-series, cross-sectional data, and with priors on individual missing values or those measured with error. With MO, scholars can preprocess their data to account for measurement error and missing data and then use the overimputed data sets our software produces with whatever model they would have used without it, ignoring the measurement issues. Along with the more application-specific techniques of, for example, Imai and Yamamoto (2010) and Katz and Katz (2010), this represents a way to take measurement error more seriously.

The advances described here can be implemented when the degree of measurement error can be analytically determined from known sample properties, estimated with additional proxies, or even when it can only be bounded by the analyst. However, often the original creators of new measures are in the best position to know the degree of measurement error present and we would encourage those who create data to include their estimates of variable or cell-level measurement error as important auxiliary information, much as sampling frame weights are considered essential in the survey literatures. Now that easy-to-use procedures exist for analyzing these data, we hope this information will be made more widely available and used.

Authors' Note

For helpful comments, discussions, and data, we thank Walter Mebane, Gretchen Casper, Simone Dietrich, Justin Grimmer, Sunshine Hillygus, Burt Monroe, Adam Nye, Michael Peress, Eric Plutzer, Mo Tavano, Shawn Treier, Joseph Wright, and

Chris Zorn. All data, code, and information needed to replicate this work are available in a Dataverse replication file at Blackwell, Honaker, and King (2015a).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the NSF (SES-1059723).

Notes

1. A *cell* here is one observation of one variable, which would be one element in a data matrix. We call these priors because they represent prior information with respect to the imputation model. They are posterior distributions with respect to the measurement error model since they give us information on the location of the true value, conditional on the mismeasured value.
2. One difference between missing data and data measured with error is that with missing data, there are no additional parameters governing "how missing" the observation is. With mismeasured data, we have to estimate its error variance. But this is possible with a variety of data types, including a validation sample, gold standard data, or multiple proxies, all of which we discuss subsequently.
3. This observed-data likelihood is difficult to maximize directly in real data sets with arbitrary patterns of missingness. Fast algorithms to maximize it have been developed that use the relationship between equations (2), (5), and the implied regressions, using iterative techniques, including variants of Markov chain Monte Carlo, EM, or EM with bootstrapping.
4. In this setting, the prior means are *empirical priors*, set by the (mismeasured) data rather than auxiliary information or prior knowledge (Carlin and Louis 2000; Efron 2013; Maritz and Lwin 1989; Robbins 1956).
5. This measurement error model implicitly assumes that the error is non-differential or uncorrelated with the outcome variable. However, in BHK2, Subsection- Robustness to Violating Assumptions, we show that multiple overimputation (MO) is relatively robust to at least small violations of this assumption, at least when compared to a more traditional errors-in-variables approach.
6. We let y_i, the dependent variable of the analysis model, follow a Bernoulli distribution with probability $\pi_i = 1/(1 + \exp(-X_i\beta))$, where $X_i = \left(x_i^*, z_i^*, s_i\right)'$ and $\beta = (-7, 1, 1, -1)$. The observed variables, x and z, are generated from a normal

distribution with means x^* and z^* and measurement error variances 0.5. We allow scattered missingness of a random 10 percent of all the cell values of x, and z when s (a perfectly measured covariate) is greater than its mean. We created the true, latent data (x^*, z^*, s) by drawing from a multivariate normal with mean vector $(5, 3, 1)$ and covariance matrix (1.5 0.5 −0.2, 0.5 1.5 −0.2, −0.2 −0.2 0.5). Sample sizes are 1,000.

7. In this simulation, the variance of the estimates is swamped by the squared bias of the estimates, so that any difference in the mean square error (MSE) is almost entirely due to bias, rather than efficiency. More succinctly, these plots are substantively similar if we replace MSE with bias.
8. For a related problem of "editing" data with suspicious measurements, Ghosh-Dastidar and Schafer (2003) develop an innovative multiple imputation framework similar in spirit to ours, albeit with an implementation specific to their application.
9. S. Y. Lee (2007) covers a number of Bayesian approaches to structural equation modeling, including some that take into consideration missing data.
10. For these simulations, we have $y_i = \beta x_i + \epsilon_i$ with $\beta = 1$, $\epsilon_i \sim \mathcal{N}(0, 1.5^2)$, $x_i^* \sim \mathcal{N}(5, 1)$, and $\sigma_u^2 = 1$. Thus, we have $\rho = 0.5$. We used sample sizes of 1,000 and 10,000 simulations.
11. More generally, simulations run at various values of the true ρ lead to the same qualitative results as presented here. Underestimates of ρ lead to underestimates of the true slope and overestimate of ρ lead to overestimates of the true slope.
12. If we use MO at all levels of ρ to generate the most assumption-free MO-based bounds possible, the bounds largely agree with the minimal-assumptions bounds.
13. These simulations also point to the use of MO as a tool for sensitivity analysis. MO not only provides bounds on the quantities of interest but can provide what the estimated quantity of interest would be under various assumptions about the amount of measurement error.
14. Monthly national unemployment is taken from the Bureau of Labor Statistics, labor force series. Presidential approval is from the Gallup historical series, aggregated to the monthly level. We use data from 1971 to 2011. We use the last three years of each four-year Presidential term of office, to avoid approval levels within the "honeymoon" period, without adding controls into the model. We added a monthly indicator for cumulative time in office, but this only slightly strengthened these results, and so we leave the presentation as the simplest, bivariate relationship.
15. That is, if we are attempting to explain current approval, we assume that the unemployment k months in the past (the k lag) and k months in the future (the k lead) are proxies for the current level of unemployment, which we assume is

unavailable to our analyst. As k increases, the measures of unemployment may have drifted increasingly far from the present unemployment level, so both proxies employed have increased measurement error. We use these same two proxies in each of our MO models (as previously described in the Monte Carlo Evidence subsection and BHK2 Subsection- Multiple Proxies).

16. A partial explanation can be understood from robustness results we show in BHK2, Section Robustness to Correlated Measurement Errors. In periods where unemployment trends upward (or downward), the k-month lag and the k-month lead of unemployment will generally have opposite signed measurement error. So the measurement errors in the proxies will be negatively correlated. In figure 8 of BHK2, we demonstrate that this is a problem for both models, but that MO is much more robust to this violation than the errors-in-variables model.
17. Adding other covariates into the imputation model could increase the efficiency of the overimputations. Averaging the two proxies would give an interpolation that might be a superior proxy to those used, and we demonstrate an application of averaging across proxies in MO in the The Effect of Political Preferences on Vote Choice subsection. Moreover, in many applications, if there is periodic missingness over time in a variable, the best approach might be to impute all the missing values in the series with an imputation model built for time-series cross-sectional data, such as developed in Honaker and King (2010); this reinforces the main thesis of our argument, that measurement error and missing data are fundamentally the same problem.
18. The mean is 10.2 with an interquartile range of 3 to 20.
19. Within parishes, the median sample size is 6, and only 6 percent of observations have at least 30 observed responses to the abortion scale among fellow congregants in their parish. We use the small sample, within-group estimate for the standard deviations, pooling variance across parishes, when the resulting standard error is 0.2 greater than the standard error estimate using only observations within a particular parish. This changes a few small parishes. Within neighborhoods, however, the median sample size is 47, fully 95 percent of observations have 30 or more respondents in their neighborhood, and so we always estimate the standard deviation in each neighborhood directly from only the observations in that neighborhood.
20. For additional work on small area estimation from an MO framework, see Honaker and Plutzer (2011). In particular, there are additional possible efficiency gains from treating the errors within individuals in the same church or community as correlated, as well as bringing in auxiliary Census data, and this work shows how to approach this with two levels of imputations at both the individual and aggregated level.
21. Forty-seven percent of this missingness is due to respondents who answer some, but not all, of the abortion scenarios that constitute the abortion scale.

Knowing the pattern of answers to the other completed abortion questions, as well as the other control variables in the model, help predict these missing responses.

22. Catholics are still less likely to support abortion (a mean support of 3.1 compared to 3.7 for non-Catholics), but this difference is explained by variables controlled for in the model such as individual demographics and the social ties of Catholic churches which have lower mean parish attitudes than non-Catholic churches.

References

Ansolabehere, Stephen, Jonathan Rodden, and James M. Snyder. 2008. "The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting." *American Political Science Review* 102: 215-32.

Berger, James. 1994. "An Overview of Robust Bayesian Analysis (with Discussion)." *Test* 3:5-124.

Bernaards, Coen A., Thomas R. Belin, and Joseph L. Schafer. . 2007. "Robustness of a Multivariate Normal Approximation for Imputation of Incomplete Binary Data." *Statistics in Medicine* 26:1368-82. Retrieved April 1, 2015. (http://dx.doi.org/10.1002/sim.2619).

Black, Dan A., Mark C. Berger, and Frank A. Scott. 2000. "Bounding Parameter Estimates with Nonclassical Measurement Error." *Journal of the American Statistical Association* 95:739-48. Retrieved April 1, 2015. (http://www.jstor.org/stable/2669454).

Blackwell, Matthew, James Honaker, and Gary King. 2015a. "Replication Data for: A Unified Approach to Measurement Error and Missing Data: Overview." UNF: 5: n/rveBXUX+nOxE6Z5xsWNg==. Retrieved April 1, 2015. (http://dx.doi.org/10.7910/DVN/29606 IQSS Data-verse Network [Distributor]).

Blackwell, Matthew, James Honaker, and Gary King. 2017. "A Unified Approach to Measurement Error and Missing Data: Details and Extensions." *Sociological Methods and Research* 46:342-69.

Brownstone, David and Robert G. Valletta. 1996. "Modeling Earnings Measurement Error: A Multiple Imputation Approach." *Review of Economics and Statistics* 78: 705-17.

Carlin, Bardley P. and Thomas A. Louis. 2000. *Bayes and Empirical Bayes Methods for Data Analysis*. 2nd ed. Boca Raton, FL: CRC Press.

Carroll, R. J., D. Ruppert, and L. A. Stefanski. 1995. *Measurement Error in Nonlinear Models*. Vol. 63. Boca Raton, FL: CRC.

Carroll, Raymond J. and Leonard A. Stefanski. 1990. "Approximate Quasi-likelihood Estimation in Models with Surrogate Predictors." *Journal of the American*

Statistical Association 85:652-63. Retrieved April 1, 2015. (http://www.jstor.org/stable/2290000).

Cole, Stephen R., Haitao Chu, and Sander Greenland. 2006. "Multiple-imputation for Measurement-error Correction." *International Journal of Epidemiology* 35: 1074-81.

Cook, J. and L. Stefanski. 1994. "Simulation-extrapolation Estimation in Parametric Measurement Error Models." *Journal of the American Statistical Asociation* 89: 1314-28.

Efron, Bradley. 2013. *Large-scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge, UK: Cambridge University Press.

Freedman, Laurence S., Douglas Midthune, Raymond J. Carroll, and Victor Kipnis. 2008. "A Comparison of Regression Calibration, Moment Reconstruction and Imputation for Adjusting for Covariate Measurement Error in Regression." *Statistics in Medicine* 27:5195-216.

Fuller, Wayne A. 1987. *Measurement Error Models*. New York: Wiley.

Ghosh-Dastidar, B. and J. L. Schafer. 2003. "Multiple Edit/Multiple Imputation for Multivariate Continuous Data." *Journal of the American Statistical Association* 98:807-17.

Graham, John W., Allison E. Olchowski, and Tamika D. Gilreath. 2007. "How Many Imputations Are Really Needed? Some Practical Clarifications of Multiple Imputation Theory." *Prevention Science* 8:206-13.

Guo, Ying and Roderick J. Little. 2011. "Regression Analysis with Covariates that Have Heteroscedastic Measurement Error." *Statistics in Medicine* 30:2278-94. Retrieved April 1, 2015. (http://dx.doi.org/10.1002/sim.4261).

Guo, Ying, Roderick J. Little, and Daniel S. McConnell. 2012. "On Using Summary Statistics from an External Calibration Sample to Correct for Covariate Measurement Error." *Epidemiology* 23:165-74.

Guolo, Annamaria. 2008. "Robust Techniques for Measurement Error Correction: A Review." *Statistical Methods in Medical Research* 17:555-80.

He, Yulei and Alan M. Zaslavsky. 2009. "Combining Information from Cancer Registry and Medical Records Data to Improve Analyses of Adjuvant Cancer Therapies." *Biometrics* 65:946-52.

Honaker, James and Gary King. 2010. "What to Do About Missing Values in Time Series Cross-section Data." *American Journal of Political Science* 54: 561-81. Retrieved April 1, 2015. (http://gking.harvard.edu/files/abs/pr-abs.shtml).

Honaker, James, Gary King, and Matthew Blackwell. 2010. "Amelia II: A Program for Missing Data." Retrieved April 1, 2015. (http://gking.harvard.edu/amelia).

Honaker, James and Eric Plutzer. 2011. "Small Area Estimation with Multiple Over-imputation." Paper presented at the Midwest Political Science Association, Chicago, IL, March 2011.

Hopkins, Daniel and Gary King. 2010. "Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Incomparability." *Public Opinion Quarterly* 1-22. Retrieved April 1, 2015. (http://gking.harvard.edu/files/abs/implement-abs.shtml).

Huckfeldt, Robert, Eric Plutzer, and John Sprague. 1993. "Alternative Contexts of Political Behavior: Churches, Neighborhoods, and Individuals." *Journal of Politics* 55:365-81.

Imai, Kosuke, Gary King, and Olivia Lau. 2008. "Toward a Common Framework for Statistical Analysis and Development." *Journal of Computational Graphics and Statistics* 17:1-22. Retrieved April 1, 2015. (http://gking.harvard.edu/files/abs/z-abs.shtml).

Imai, Kosuke and Teppei Yamamoto. 2010. "Causal Inference with Differential Measurement Error: Nonparametric Identification and Sensitivity Analysis." *American Journal of Political Science* 54:543-60.

Katz, Jonathan N. and Gabriel Katz. 2010. "Correcting for Survey Misreports Using Auxiliary Information with an Application to Estimating Turnout." *American Journal of Political Science* 54:815-35.

King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95:49-69. Retrieved April 1, 2015. (http://gking.harvard.edu/files/abs/evil-abs.shtml).

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44:341-55. Retrieved April 1, 2015. (http://gking.harvard.edu/files/abs/making-abs.shtml).

King, Gary and Langche Zeng. 2002. "Estimating Risk and Rate Levels, Ratios, and Differences in Case-control Studies." *Statistics in Medicine* 21: 1409-27. Retrieved April 1, 2015. (http://gking.harvard.edu/files/abs/1s-abs.shtml).

Klepper, Steven and Edward E. Leamer. 1984. "Consistent Sets of Estimates for Regressions with Errors in All Variables." *Econometrica* 52:163-84. Retrieved April 1, 2015. (http://www.jstor.org/stable/1911466).

Leamer, Edward. 1978. *Specification Searches*. New York: Wiley.

Lee, Katherine J. and John B. Carlin. 2010. "Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation." *American Journal of Epidemiology* 171:624-32. Retrieved April 1, 2015. (http://aje.oxfordjournals.org/content/171/5/624.abstract).

Lee, Sik-Yum. 2007. *Structural Equation Modeling: A Bayesian Approach*. Vol. 680. New York: John Wiley.

Maritz, J. S. and T. Lwin. 1989. *Empirical Bayes Methods*. 2nd ed. London, UK: Chapman and Hall.

Meng, Xiao-Li. 1994. "Multiple-imputation Inferences with Uncongenial Sources of Input." *Statistical Science* 9:538-73.

Messer, Karen and Loki Natarajan. 2008. "Maximum Likelihood, Multiple Imputation and Regression Calibration for Measurement Error Adjustment." *Statistics in Medicine* 27:6332-50.

Nielsen, Søren Feodor. 2007. "Proper and Improper Multiple Imputation." *International Statistical Review* 71:593-607.

Robbins, Herbert. 1956. "An Empirical Bayes Approach to Statistics." *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* 1:157-63.

Robins, James and Naisyin Wang. 2000. "Inference for Imputation Estimators." *Biometrika* 87:113-24.

Rubin, Donald B. 1976. "Inference and Missing Data." *Biometrika* 63:581-92.

Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *The Annals of Statistics* 6:34-58.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.

Rubin, Donald and Nathaniel Schenker. 1986. "Multiple Imputation for Interval Estimation for Simple Random Samples with Ignorable Nonresponse." *Journal of the American Statistical Association* 81:366-74.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. London, UK: Chapman & Hall.

Schafer, Joseph L., Trena M. Ezzati-Rice, W. Johnson, Meena Khare, Roderick J. A. Little, and Donald B. Rubin. 1996. "The NHANES III Multiple Imputation Project." In *Proceedings of the Survey Research Methods Section*. Vol. 60, 28-37. Alexandria, VA: The American Statistical Association. Retrieved April 1, 2015. https://www.amstat.org/sections/SRMS/Proceedings/papers/1996_004.pdf.

Stefanski, L. A. 2000. "Measurement Error Models." *Journal of the American Statistical Association* 95:1353-58.

Wang, Naisyin and James Robins. 1998. "Large-sample Theory for Parametric Multiple Imputation Procedures." *Biometrika* 85:935-48.

White, Ian R. 2006. "Commentary: Dealing with Measurement Error: Multiple Imputation or Regression Calibration?" *International Journal of Epidemiology* 35: 1081-82. Retrieved April 1, 2015. (http://ije.oxfordjournals.org/content/35/4/1081.short).

Author Biographies

Matthew Blackwell is an assistant professor in the Department of Government at Harvard University.

James Honaker is a senior research scientist in the Data Science Program at the Institute for Quantitative Social Science at Harvard University.

Gary King is the Albert J. Weatherhead III university professor and director of the Institute for Quantitative Social Science at Harvard University. For more information, see GaryKing.org.