A Unified Approach to Measurement Error and Missing Data: Details and Extensions[*]

Matthew Blackwell[†] James Honaker[‡] Gary King[§]

March 24, 2015

Abstract

We extend a unified and easy-to-use approach to measurement error and missing data. Blackwell, Honaker and King (2015*b*) gives an intuitive overview of the new technique, along with practical suggestions and empirical applications. Here, we offer more precise technical details; more sophisticated measurement error model specifications and estimation procedures; and analyses to assess the approach's robustness to correlated measurement errors and to errors in categorical variables. These results support using the technique to reduce bias and increase efficiency in a wide variety of empirical research.

[*] 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 code and information needed to replicate this work are available in a Dataverse replication file at Blackwell, Honaker and King (2015*a*).

[†] Assistant Professor, Department of Government, Harvard University, Institute for Quantitative Social Science, 1737 Cambridge Street, Cambridge, MA 02138 (mblackwell@gov.harvard.edu, MattBlackwell.org)

[‡] Senior Research Scientist, Institute for Quantitative Social Science, 1737 Cambridge Street, Cambridge, MA 02138 (jhonaker@iq.harvard.edu, hona.kr)

[§] Albert J. Weatherhead III University Professor, Harvard University, Institute for Quantitative Social Science, 1737 Cambridge Street, Cambridge, MA 02138 (king@harvard.edu, GaryKing.org)

1 Introduction

In this paper, we extend and technically undergird a new unified approach to analyses of data sets with both measurement error and missing data, extremely common coexisting conditions throughout applied social science research. The technique is called multiple overimputation (MO) and it generalizes the popular approach of multiple imputation (MI) for missing data as well as a methodology for measurement error. The technique benefits by formalizing the idea of regarding missing data as an extreme form of measurement error, or measurement error as a form of partially missing data. Applied researchers can easily use the technique to preprocess their data to deal with both methodological problems simultaneously and then use whatever technique they might have applied to their data if it had neither problem.

Blackwell, Honaker and King (2015*b*) (BHK, hereinafter) offers an overview of MO, along with empirical applications and suggestions for practice. In this paper, we first offer a more precise formulation of the methodology, its assumptions, and its implementation (Section 2). Next, we show how to estimate rather than assume the variances in more sophisticated measurement error models (Section 3). We then show how MO remains robust for common categorical variables even when specifying a model for continuous measurement error (Section 4) and for different types of correlated measurement error when the measurement error model assumes independence (Section 5). Finally, we offer detailed simulations to provide guidance for the number of data sets to overimpute (Section 6). Section 7 concludes.

2 Model and Estimation

Here we introduce the general MO model and a specific EM algorithm implementation. We also show that it is equivalent to MI with observation-level priors as introduced by Honaker and King (2010). We also offer more general notation than that in the text.

2.1 Notation

Consider data x_{ij} with observations $i = 1, \dots, N$ and variables $j = 1, \dots, p$. Although numerical values for all elements of $x \equiv \{x_{ij} | \forall i, j\}$ exist, we have knowledge about each through a three-part measurement mechanism:

$$m_{ij} = \begin{cases} 0 & \text{if } x_{ij} \text{ is observed} \\ 1 & \text{if } x_{ij} \text{ is missing but an unbiased proxy } w_{ij} \text{ is observed} \\ 2 & \text{if } x_{ij} \text{ is missing} \end{cases} \tag{1}$$

The three cases correspond respectively to (0) direct observation of the true value; (1) unbiased estimation of the true value, $E(w_{ij}) = x_{ij}$; and (2) classical missingness, for which the true value and proxies of it are unavailable. We partition x into sets where the true values are observed or missing: $x = \{x_{\text{obs}}, x_{\text{mis}}\}$ where

$$x_{\text{obs}} \equiv \{x_{ij} \mid (\forall i, j) \wedge (m_{ij} = 0)\} \quad \text{and} \tag{2}$$

$$x_{\text{mis}} \equiv \{x_{ij} \mid (\forall i, j) \wedge (m_{ij} > 0)\}. \tag{3}$$

Thus, we observe portions of the covariates x_{obs}, the measurement mechanism $m \equiv \{m_{ij} \mid \forall i, j\}$, and the proxies $w \equiv \{w_{ij} \mid (\forall i, j) \wedge (m_{ij} = 1)\}$.

Our goal is to infer the distribution of x, $p(x|\theta) \equiv p(x_{\text{obs}}, x_{\text{mis}}|\theta)$, where θ is an unknown parameter vector. We begin with the observed-data probability density function

$$p(m, w, x_{\text{obs}}|\theta, \gamma, \phi) = \int p(m|x_{\text{obs}}, x_{\text{mis}}, w, \phi)p(w|x_{\text{mis}}, x_{\text{obs}}, \gamma)p(x_{\text{obs}}, x_{\text{mis}}|\theta)dx_{\text{mis}}, \tag{4}$$

where ϕ is the parameter of the distribution of m, and γ is the parameter of the distribution of the measurement mechanism.

2.2 Assumptions

We require two key assumptions. First, we enable researchers to avoid specifying a full probability density for the measurement assignment, $p(m|x_{\text{obs}}, x_{\text{mis}}, w, \phi)$. To do this, we generalize MI's (realized) missing at random assumption:

Assumption 1 (Ignorable Measurement Mechanism Assignment (IMMA))**.** *(a) For any values of m, x_{obs}, w, x_{mis}, and, x'_{mis},*

$$p(m|x_{obs}, x_{mis}, w, \phi) = p(m|x_{obs}, x'_{mis}, w, \phi). \tag{5}$$

and (b) the parameters governing m, ϕ, are distinct from those governing w, γ, and x, θ, in Equation 4.

Here, x_{mis} and x'_{mis} are two possible realizations of the missing data, so that 5 says that the distribution of the mismeasurement indicator, m, is the same no matter the value of the missing data.

We also assume knowledge of the measurement error data generation process, allowing its indexed parameter vector to be known or estimated:

Assumption 2 (Measurement error distribution)**.** *The distribution $p(w|x_{mis}, x_{obs}, \gamma)$ is known up to its parameters γ. The parameters γ are either known or a consistent estimator ($\hat{\gamma}$ s.t. $\hat{\gamma} \xrightarrow{p} \gamma$) is available.*

2.3 The Observed-data likelihood

We begin with three definitions. First, we specify the *joint likelihood* that explicitly handles the presence of missingness or measurement error:

$$L_j(\theta, \gamma, \phi) = p(m, w, x_{\text{obs}}|\theta, \gamma, \phi). \tag{6}$$

Second, we write the *profile likelihood* of (θ, γ) with respect to the parameters of m:

$$L_p(\theta, \gamma) = \max_\phi \left[L_j(\theta, \gamma, \phi)\right]. \tag{7}$$

The profile likelihood is useful because the maximum likelihood estimates from it will turn out to be the maximum likelihood estimates based on the joint likelihood. Finally, we define the *ignorance likelihood*, ignoring the measurement indicator m:

$$L(\theta, \gamma) = p(w, x_{\text{obs}}|\theta, \gamma). \tag{8}$$

We now use these definitions to establish the relationship between these quantities, in a manner similar to results for the missingness at random assumption in classical analyses of multiple imputation for missing data.

Theorem 1. *Under IMMA (Assumption 1), the profile likelihood for (θ, γ),is proportional to the ignorance likelihood: $L_p(\theta, \gamma) \propto L(\theta, \gamma)$.*

Proof. Under IMMA, we can factor the joint likelihood as

$$
\begin{aligned}
L_j(\theta, \gamma, \phi) &= p(m, w, x_{\text{obs}} | \theta, \gamma, \phi) && (9)\\
&= \int p(x_{\text{obs}}, x_{\text{mis}} | \theta) p(w | x_{\text{mis}}, x_{\text{obs}}, \gamma) p(m | x_{\text{obs}}, x_{\text{mis}}, w, \phi) dx_{\text{mis}}, && (10)\\
&= \int p(x_{\text{obs}}, x_{\text{mis}} | \theta) p(w | x_{\text{mis}}, x_{\text{obs}}, \gamma) p(m | x_{\text{obs}}, w, \phi) dx_{\text{mis}}, && (11)\\
&= p(m | x_{\text{obs}}, w, \phi) \int p(x_{\text{obs}}, x_{\text{mis}} | \theta) p(w | x_{\text{mis}}, x_{\text{obs}}, \gamma) dx_{\text{mis}}, && (12)\\
&= p(m | x_{\text{obs}}, w, \phi) L(\theta, \gamma). && (13)
\end{aligned}
$$

The second equality is simply the definition of the joint likelihood, the third follows from IMMA, the fourth and fifth from the properties of expectations. Substituting this expression into the definition of the profile likelihood (Equation 7) gives:

$$
\begin{aligned}
L_p(\theta, \gamma) &= \max_{\phi} \left[p(m | x_{\text{obs}}, w, \phi) L(\theta, \gamma) \right] && (14)\\
&= L(\theta, \gamma) \max_{\phi} \left[p(m | x_{\text{obs}}, w, \phi) \right] && (15)\\
&\propto L(\theta, \gamma), && (16)
\end{aligned}
$$

where the last equality holds because the second factor of the second line is constant with respect to (θ, γ). □

What Theorem 1 tells us is that inferences based on $L(\theta, \gamma)$ will be the same as inferences about θ and γ from the profile likelihood and, thus, the joint likelihood. Because we state IMMA in terms of all values of m, x_{obs}, and w, then this result will hold both in our realized sample and in repeated samples. Thus, both Bayesian inference and general frequentist inference are justified under this result (Seaman et al., 2013). From here on, we thus refer to $L(\theta, \gamma)$ as the *observed-data likelihood*, where its dependence on IMMA is implicit.

2.4 An EM algorithm for missing data and measurement error

The EM algorithm is commonly used to maximize the likelihoods intractable due to missingness. In this section, we show how we can incorporate measurement error into the standard EM algorithm for missing data. For expository purposes, we assume that the true values of the measurement error distribution parameters, γ, are known and, thus, fixed. In Section 3.2, we consider extensions for when γ is estimated.

The object in the EM-algorithm is the expectation of the log-likelihood of the complete data, averaged across the missing data given a current guess of the parameters:

$$Q(\theta|\theta^{(t)}) = \int \log\left[p(x_{\text{mis}}, x_{\text{obs}}, w|\theta, \gamma)\right] p(x_{\text{mis}}|x_{\text{obs}}, w, \theta^{(t)}, \gamma) dx_{\text{mis}}. \tag{17}$$

Note that the distribution of the missing data which we are averaging across fixes the value of θ, while the value in the complete-data log-likelihood is allowed to vary. The EM algorithm proceeds in two steps. First, in the E-step, we compute the function $Q(\theta|\theta^{(t)})$, which involves calculating the posterior distribution of the missing data, conditional on the observed. The second step, called the M-step, maximizes $Q(\theta|\theta^{(t)})$ with respect to θ. This becomes the new candidate value, $\theta^{(t+1)}$, and the algorithm iterates until convergence. Under suitable conditions (Wu, 1983), the EM algorithm converges to the maximum likelihood estimate of θ under the observed-data likelihood, $L(\theta, \gamma)$.

The complete-data log-likelihood is straightforward in our general setup:

$$\log\left[p(x_{\text{mis}}, x_{\text{obs}}, w|\theta, \gamma)\right] = \log\left[p(x_{\text{mis}}, x_{\text{obs}}|\theta)\right] + \log\left[p(w|x_{\text{mis}}, x_{\text{obs}}, \gamma)\right]. \tag{18}$$

Because the last term only depends on the known (mis)measurement process, it will not factor into the maximization of $Q(\theta|\theta^{(t)})$. This follows from our (implicit) assumption that the parameters of the measurement process are distinct from the parameters of the target data x. The measurement information will enter the EM algorithm through its influence on the posterior distribution of x_{mis}:

$$p(x_{\text{mis}}|x_{\text{obs}}, w, \theta^{(t)}, \gamma) \propto p(x_{\text{mis}}|x_{\text{obs}}, \theta^{(t)})p(w|x_{\text{mis}}, x_{\text{obs}}, \gamma). \tag{19}$$

Thus, we can write the objective function in this case as:

$$Q(\theta|\theta^{(t)}) = \int \log\left[p(x_{\text{mis}}, x_{\text{obs}}|\theta)\right] p(x_{\text{mis}}|x_{\text{obs}}, \theta^{(t)})p(w|x_{\text{mis}}, x_{\text{obs}}, \gamma) dx_{\text{mis}}. \tag{20}$$

Because (19) is a posterior for the missing data and (18) is a complete-data log-likelihood, all of the properties of the general EM algorithm will apply to it here with the form (20). Namely, this algorithm will converge to the maximum likelihood estimate of θ under the observed-data likelihood. If we replace the true parameters, γ, with a consistent estimate, $\hat{\gamma}$, the EM algorithm will still converge to a consistent estimator of the complete-data parameters by standard asymptotic arguments (Newey and McFadden, 1994). While estimating γ does not affect the consistency of our EM estimator, it does affect the variance of the asymptotic distribution. In order to achieve consistent variance estimates, we can rely on the non-parametric bootstrap.

2.5 Multiple Overimputation

Analyzing the true values of the data x would be much easier than the observed data since the mismeasured and missing components contribute to the likelihood (8) in complicated ways. Thus, MO seeks to form a series of complete, ideal datasets: $x^{(1)}, x^{(2)}, \ldots, x^{(B)}$. Each of these overimputed datasets is of the form $x^{(b)} = (x_{\text{obs}}, x_{\text{mis}}^{(b)})$, so that the perfectly measured data is constant across the overimputations. We refer to this as overimputation because we replace observed data w with draws from an imputation model for x_{mis}. To form these overimputations, we take draws from the posterior predictive distribution of the unobserved data:

$$(x_{\text{mis}}^{(b)}) \sim p(x_{\text{mis}}^{(b)} | x_{\text{obs}}, w) = \int p(x_{\text{mis}}^{(b)} | x_{\text{obs}}, w, \theta, \gamma) p(\theta, \gamma | x_{\text{obs}}, w) d\theta d\gamma. \tag{21}$$

Once we have these B overimputations, we can simply run B separate analyses on each dataset and combine them using straightforward rules in one of the two standard ways in MI. See BHK, Section 2.1.

Equation (21) suggests one way to create multiple imputations: (1) draw $\theta^{(b)}$ from its posterior $p(\theta | x_{\text{obs}}, w, \gamma)$, then (2) draw $(x_{\text{mis}}^{(b)})$ from $p(x_{\text{mis}}^{(b)} | x_{\text{obs}}, w, \theta^{(b)}, \gamma)$. Usually these procedures are implemented with either data augmentation (that is, Gibbs sampling) or the expectation-maximization (EM) algorithm combined with an additional sampling step. We focus here on how our method works in the EM algorithm, since these two approaches are closely linked and often lead to similar inferences (Schafer, 1997; King et al., 2001; Honaker and King, 2010). The EM algorithm from Section 2.4 will give us either a maximum likelihood (MLE) or maximum a posteriori (MAP) estimate of the parameters of the complete data, $\hat{\theta}$. With this estimate, there are various ways to draw from the predictive posterior distribution. Here, we apply the EM algorithm to a series of data sets drawn using the nonparametric bootstrap and use the estimated value of $\hat{\theta}_b$ in that bootstrapped sample as a draw from the posterior of θ. Of course, MO is not limited to the EM algorithm; we could replace it with a suitable MCMC approach as well.

2.6 A Specific Model within the Class

In the above description of the model, we have left the distributions unspecified. To implement a specific model, we must of course specify this additional information. To make our estimation approach practical, we narrow the class of allowable data generation processes to any distribution that possesses the property of *statistical duality.* This is a simple property (related to self-conjugacy in Bayesian analysis) possessed by a variety of distributions, such as normal, Laplace, Gamma, Inverse Gamma, Pareto, and others (Bityukov et al., 2006).[1] Here, we offer the most common choice in practice, which is that the complete (but partially unobserved) data (x) is multivariate normal with mean μ and covariance Σ. This implies that any conditional distribution of the ideal is also normal.

The above measurement error distribution is in its most general form, a function of the entire ideal data vector, x, and some parameters, γ. As noted by Stefanski (2000), all approaches to correcting measurement error must include additional information about this distribution. In the simple but powerful special case, we assume that $w_{ij} \overset{\text{ind}}{\sim} \mathcal{N}(x_{ij}, \lambda_{ij}^2)$ where the measurement error variance λ_{ij}^2 is known or estimable using techniques from Section 3. Our assumption corresponds to that of classical measurement error, yet our modified EM algorithm can handle more general cases than this. If the measurement error is known to be biased or dependent upon another variable, we can simply adjust the cell-level means above and proceed as usual. Essentially, one must have knowledge of *how* the variable was mismeasured. The simulation results in Section 5.2 further indicate that MO is robust to these assumptions in certain situations.

With the measurement error model above, the normality of the data makes the calculation of the sufficient statistics straightforward. In a slight abuse of notation, we can gather the independent measurement error distributions on w_{ij} into a multivariate normal with mean $x_{\text{mis},i}$ and covariance matrix $\Lambda_i = \lambda_i^2 I$, where $\lambda_i^2 = \{\lambda_{ij}^2;\ m_{ij} = 1\}$ and I is the identity matrix with dimension equal to $\sum_j \mathbb{I}\{m_{ij} = 1\}$. To ease exposition, let $\theta = (\mu, \Sigma)$ and consider the case without fully missing values. When x comes from an exponential family, the E-step involves calculating the expected value of the complete-data sufficient statistics, marginalized across the missing data:

$$E[T(x)|x_{\text{obs}}, w, \theta^{(t)}] = \int T(x)\ p(x_{\text{mis}}|x_{\text{obs}}, \theta^{(t)}) p(w|x_{\text{mis}}, \gamma) dx_{\text{mis}} \tag{22}$$

$$= \int T(x)\ \prod_i p(x_{i,\text{mis}}|x_{i,\text{obs}}, \theta^{(t)}) p(w_i|x_{i,\text{mis}}, \Lambda_i) dx_{i,\text{mis}} \tag{23}$$

[1] If a function $f(a, b)$ can be expressed as a family of probability densities for variable a given parameter b, $p(a|b)$, and a family of densities for variable b given parameter a, $p(b|a)$, so that $f(a, b) = p(a|b) = p(b|a)$, then $p(a|b)$ and $p(b|a)$ are said to be statistically dual.

Here $T(x)$ is the set of sufficient statistics for the multivariate normal. Calculating this expression is equivalent to evaluating $Q(\theta|\theta^{(t)})$ in (20).

In order to calculate the expectation in (23), we must know the full conditional distribution, which is $p(x_{i,\text{mis}}|x_{i,\text{obs}}, w_i, \theta, \Lambda_i) \propto p(x_{i,\text{mis}}|x_{i,\text{obs}}, \theta)p(w_i \mid x_{i,\text{mis}}, \Lambda_i)$. Note that each of the distributions is (possibly multivariate) normal, with $[x_{i,\text{mis}}|x_{i,\text{obs}}, \theta] \sim \mathcal{N}(\mu_{\text{e}|\text{o}}, \Sigma_{\text{e}|\text{o}})$ and $[w_i|x_{i,\text{mis}}, \Lambda_i] \sim \mathcal{N}(x_{i,\text{mis}}, \Lambda_i)$, where $(\mu_{\text{e}|\text{o}}, \Sigma_{\text{e}|\text{o}})$ are deterministic functions of θ and $x_{i,\text{obs}}$. This distribution amounts to the regression of $x_{i,\text{mis}}$ on $x_{i,\text{obs}}$. If the values were simply missing, rather than measured with error, then the E-step would simply take the expectations with respect to this conditional expectation. With measurement error, we must combine these two sources of information. Using standard results on the normal distribution, we write the full conditional as

$$(x_{i,\text{mis}}|x_{i,\text{obs}}, w_i, \theta^{(t)}, \Lambda_i) \sim \mathcal{N}(\mu^*, \Sigma^*), \quad \Sigma^* = (\Lambda_i^{-1} + \Sigma_{\text{e}|\text{o}}^{-1})^{-1}, \quad \mu^* = \Sigma^*(\Lambda_i^{-1} w_i + \Sigma_{\text{e}|\text{o}}^{-1} \mu_{\text{e}|\text{o}}). \tag{24}$$

We simply change our E-step to calculate this expectation for each cell measured with error and proceed with the M-step as usual.[2] That is, we would find the values of $\mu^{(t+1)}$ and $\Sigma^{(t+1)}$ that maximize the likelihood with these updated sufficient statistics in (23). Note that while we assume that the measurement errors on different variables are independent, one could incorporate dependence into Λ_i. The result in (24) is identical to the results in Honaker and King (2010), when we set a prior distribution for $x_{i,\text{mis}}$ that is normal with mean w_i and variance Λ_i. See their paper for additional implementation details.

Finally, we note that the imputation of purely missing values for a given observation will not directly depend on imputations of the mismeasured value for that observation. This is because imputations are always computed from a regression of a missing value on the observed data for that observation. Thus, since both missing and mismeasured variables are missing for a given observation, the imputation of one will not affect the other. If there is gold-standard data for a mismeasured value, then the mismeasured variable is observed and, thus, the imputations of the missing values can depend on that variable. In this case, the priors can influence the imputed values for the missing data through their influence on the complete-data parameters. This is appropriate, though, because the priors help to push the complete-data parameters to reflect the

[2] If there are missing values in unit i, we need to alter the definitions of the distributions to account for values in $x_{i,\text{mis}}$ that have no corresponding w_i.

true, underlying data, which is exactly the information being used to impute missing values with gold-standard measurements in the mismeasured variable.

3 Directly Estimating Measurement Error Variances

Above we assume that the amount of measurement error is known or estimable from the data. In Section 3 of BHK, we show how to analyze variables with a known proportion of measurement error or how to bound estimates when it is unknown. However, auxiliary variables often exist which can provide feasible estimates of level of measurement error.[3] The other models in the literature reviewed in Section 2.5 all rely on the existance of such auxiliary information. First, when replicated correlated proxies are available, we show how to estimate λ^2 directly (Section 3.1). Second we show how to proceed when λ_i^2 varies over the observations i or when gold standard observations are available (Section 3.2).

3.1 Multiple Proxies

When multiple proxies (or "repeated measures") of the same true variable are available, we can use relationships among them to provide point estimates of the required variances, and to set the priors in MO. For example, suppose for the same true variable x_{ij} we have two unbiased proxies with normal errors that are independent after conditioning on x_{ij}:

$$w_{i1} = x_{ij} + u_i : \;\; u_i \sim N\big(0, \lambda_u^2\big), \qquad w_{2i} = a x_{ij} + b + v_i : \;\; v_i \sim N\big(0, (c\,\lambda_u)^2\big) \tag{25}$$

where a, b, c are unknown parameters, that rescale the additional proxy measure to a different range, mean, and different degree of measurement error. The covariances and correlations between these proxies can be solved as $\mathrm{E}\big[\mathrm{cov}(w_1, w_2)\big] = a \; \mathrm{var}(x_j)$ and $\mathrm{E}\big[\mathrm{cor}(w_1, w_2)\big] = \gamma \; \mathrm{var}(x_j)/\mathrm{var}(w_1)$, where a is one of the scale parameters above, and γ is a ratio:

$$\gamma^2 = a^2 \frac{\mathrm{var}(w_1)}{\mathrm{var}(w_2)} = \frac{\mathrm{var}(x_j) + \mathrm{var}(u)}{\mathrm{var}(x_j) + (c^2/a^2)\mathrm{var}(u)}, \tag{26}$$

where all of these expectations are over units. If the measurement error is uncorrelated with x_j the variances decompose as $\lambda_u^2 = \mathrm{var}(w_1) - \mathrm{var}(x_j)$. This leads to two feasible estimates of the error

[3] If it were possible for a measurement error model to work in all contexts, without any auxiliary information to describe the measurement error, the same data set could answer any question!

variances for setting priors. First:

$$s^2(u) = \text{var}(w_1) - \text{cov}(w_1, w_2) = \text{var}(w_1) - \text{var}(x_j)\ a \tag{27}$$

which is exactly correct when $a = 1$, that is, when w_2 is on the same scale (with possibly differing intercept) as w_1. Similarly,

$$s^2(u) = \text{var}(w_1)\big(1 - \text{cor}(w_1, w_2)\big) = \text{var}(w_1) - \text{var}(x_j)\ \gamma \tag{28}$$

which is exactly correct when $c = a \Leftrightarrow \gamma = 1$, that is, the second proxy has the same relative proportion of error as the original proxy.

3.2 Heteroskedastic Measurement Error

In some applications, the amount of measurement error may vary across observations. Although most corrections in the literature ignore this possibility, it is easy to include in the MO framework, and doing so often makes estimation easier. To include this information, we add a subscript i to the variance of the measurement error: $p(w_i|x_{ij}, \lambda^2_{ui})$. We then consider two examples.

First, suppose the data include some observations measured with error and some without error. For the gold-standard observations we have $w_i = x_{ij}$, or equivalently $\lambda^2_{ui} = 0$. The imputation model would only overimpute cell values measured with error, leaving the "gold-standard" observations as is. If the other observations have a common error variance, λ^2_u, then we can easily estimate this quantity, since the variance of the gold-standard observations is σ^2_x and the mismeasured observations have variance $\sigma^2_x + \lambda^2_u$. This leads to the feasible estimator,

$$\hat{\lambda}^2_u = \hat{\sigma}^2_{mm} - \hat{\sigma}^2_{gs}, \tag{29}$$

where $\hat{\sigma}^2_{mm}$ is the estimated variance of the mismeasured observations and $\hat{\sigma}^2_{gs}$ is the estimated variance of the gold-standard observations.

As a second special case of heteroskedastic measurement error, MO can handle situations where the variance is a linear function of another variable. That is, when $\lambda^2_{ui} = rz_i$, where z_i is variable and r is the proportional constant relating the variable to the error variance. If we know r (or we can estimate it through variance function approaches), then we can easily incorporate this into the prior above using $w_i \sim \mathcal{N}(w_i|x_{ij}, rZ_i)$.

4 Robustness to Categorical Variables Measured with Error

While our imputation model assumes the data is drawn from a multivariate normal distribution, non-normal variables, such as categorical variables, can be included in the imputation and can even be overimputed for measurement error. It is well known in the MI literature that imputation via a normal model works well for categorical variables, and indeed as well as models designed especially for categorical variables and even when the analysis model is nonlinear (Schafer, 1997; Schafer and Olsen, 1998).

To illustrate how these findings extend to the MO context, we construct a simulation study with measurement error on a binary variable. To do so, we use a similar setup to BHK, except that both the underlying latent variable (which we call x_i) and the mismeasured proxy (w_i) are each a five-category ordinal variable. That is, x_i is drawn from a Binomial distribution with five trials and a 0.2 probability of success. Note that this forces x_i to exhibit skew. To create measurement error, we add $u_i \sim \mathcal{N}(0, 0.2)$ to the categorical variable and then use cutoffs at 0.5, 1.5, 2.5, 3.5, and 4.5 to force the noisy variable to also be categorical. This leads to roughly 20% of the observations to be miscoded and leads to a proportion of variance due to measurement error, ρ, of 0.2. The goal, as before, is to correctly recover the effect of x_i on y_i, where we have $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$.[4]

We follow the same procedure as in Section 3 of BHK and run MO with a range of assumptions about ρ, where ρ is the proportion of variance due to measurement error. In this simulations we use the MO in its default form and leave the values overimputed from the normal model as in, without rounding them to their nearest category. Figure 1 shows the results of these simulations, with a vertical dashed line that represents the amount of variance due to the continuous measurement error, u_i. Note that when we discretize the mismeasured variable, it reduces the variance slightly, the true ρ will be slightly less than the one denoted on the plot. The results from this more challenging test are largely similar to our earlier simulations. Even in the face of skewed, categorical latent and observed variables, MO is able to recover good estimates of the slope, β_1, around the true amount of measurement error. Further, MO minimizes the RMSE at the point as well. While this is a promising approach for categorical variables, tailored methods for misclassification, such as Katz

[4] In this simulation, $\beta = (0, 1, 0.5)$ and $\varepsilon_i \sim \mathcal{N}(0, 1.5)$. The additional covariate, z_i, is distributed i.i.d. $z_i \sim \mathcal{N}(1, 1)$. The same size is 1,000 and there are 1,000 simulations.

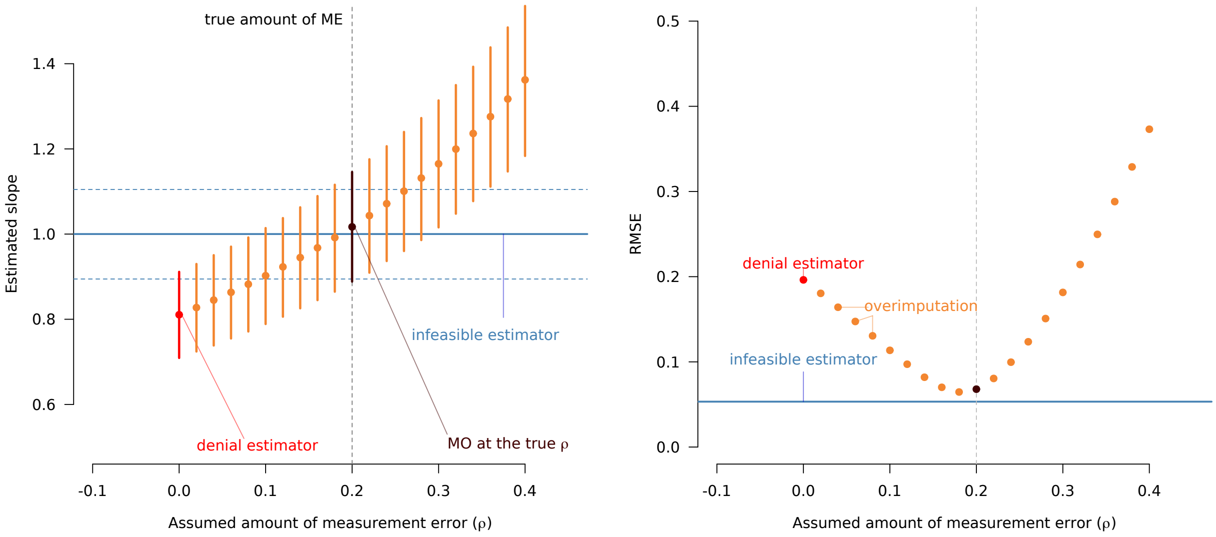


Figure 1: *Simulation results with categorical latent and observed variables.*

and Katz (2010), may outperform MO in applications that primarily focus on measurement errors on binary variables.

5 Robustness to Correlated Measurement Errors

In this Section and the next, we show how MO is robust to data problems that may occur in a large number of settings and applications. We show here how MO is robust to theoretical measurement dilemmas that occur regularly in social science data.

Until now we have assumed that measurement error is independent of all other variables. We now show how to relax this assumption. Many common techniques for treating measurement error make this strong assumption and are not robust when it is violated. For example, probably the most commonly implemented measurement error model (in the rare cases that a correction is attempted at all) is the classic errors-in-variables (EIV) model. We thus first briefly describe the EIV model to illustrate the strong assumptions required. The EIV model is also a natural point of comparison to MO, since both can be thought of as replacing mismeasured observations with predictions from auxiliary models.

5.1 The Foundation: The Errors-in-variables Model

As before, assume y_i and x_i are jointly normal with parameters as in BHK, Equation 2. Suppose instead of x_i we have a set of proxy variables which are measures of x_i with some additional normally distributed random noise:

$$w_{i1} = x_i + u_i, \qquad u_i \sim \mathcal{N}(0, \sigma_u^2); \tag{30}$$

$$w_{i2} = x_i + v_i, \qquad v_i \sim \mathcal{N}(0, \sigma_v^2); \tag{31}$$

ordered such that $\sigma_u^2 < \sigma_v^2$, making w_1 the superior of the two proxies as it has less noise.

Suppose the true relationship is $y_i = \alpha x_i + \epsilon_{i1}$, and we instead use the best available proxy and estimate $y_i = \beta w_{i1} + \epsilon_{i2} = \beta(x_i) + \beta(u_i) + \epsilon_{i2}$. We then get some degree of attenuation $0 < \beta < \alpha$ since the coefficient on u_i should be zero. This attenuation is shown in one example in the right of Figure 2 where the relationship between y_i and w_1 shown in red is weaker than the true relationship with x_i estimated in the left graph and copied in black on the right.

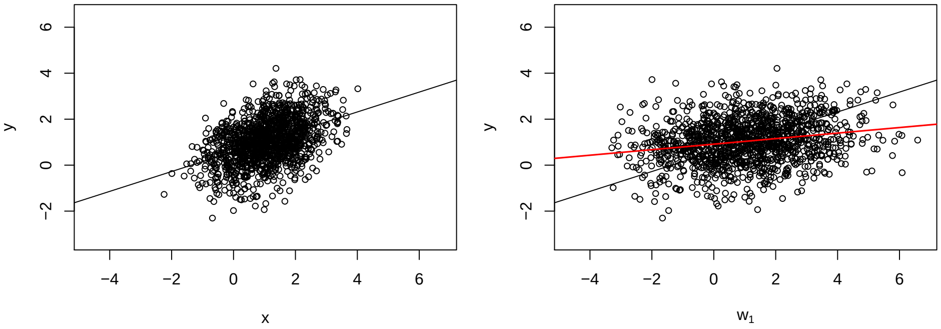


Figure 2: *On the left we see the true relationship between y_i and the latent x_i. When the mismeasured proxy w_{i1} is used instead, the estimated relationship (shown in red) is attenuated compared to the true relationship (shown in black in both graphs).*

In this simple example we can calculate the expectation of this attenuation. The coefficient on w_{i1} will be

$$\mathrm{E}[\hat{\beta}_1] = \mathrm{E}\left[\frac{\sum_i (x_i + u_i - (\overline{x+u}))(y_i - \bar{y})}{\sum_i (x_i + u_i - (\overline{x+u}))^2}\right] = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2 + \sigma_u^2}, \tag{32}$$

where $\overline{x+u}$ and $\bar{x}$ are the sample means of w_{i1} and x_i, respectively. The last term in the denominator, σ_u^2, causes this attenuation. If the variance of the measurement error is zero the term drops out and we get the correct estimate. As the measurement error increases, the ratio tends to zero.

The coefficients in the EIV approach can be estimated either directly or in two stages. A two-stage estimation procedure is the common framework to build intuition about the model and the role of the additional proxy measure. In this approach, we first obtain estimates of x_i from the relationship between the w's since they only share x_i in common, $\hat{w}_{i1} = \hat{\gamma} w_{i2}$, and then use these predictions to estimate $y_i = \delta \hat{w}_{i1} + \epsilon_{i3}$, where now $\hat{\delta}$ is an unbiased estimate of α. The relationship between the two proxy variables is shown in the left of Figure 3, and the relationship between the first stage predicted values of w_{i1} and y_i is shown in green in the right figure. This coincides almost exactly with the true relationship still shown in black in this figure.

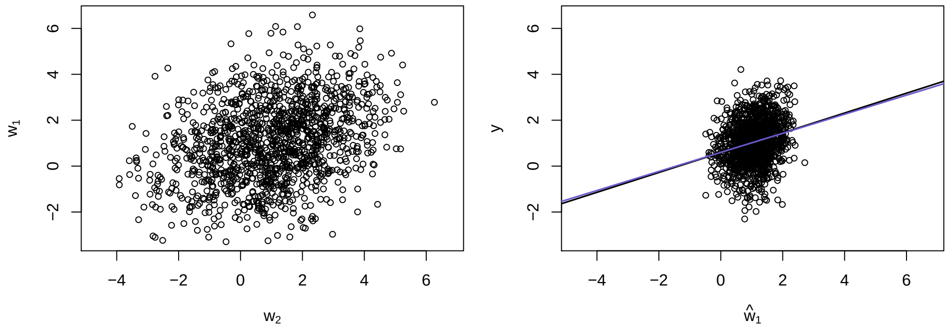


Figure 3: *The relationship between two mismeasured proxy variables (left), and the relationship between the predicted values from this model and y (right). The relationship here, shown in green, recovers the true relationship, shown in black.*

In Figure 4 we illustrate how the EIV model performs in data that meet its assumptions. The black distributions represent the distribution of coefficients estimated when the latent data x^* is available in a simulated dataset of size 500.[5] The naive regressions that do not account for measurement error are shown in red in both graphs. The coefficient on w_{i1} is attenuated to towards zero (bottom panel). The estimated constant term is biased upwards to compensate (top panel). In each simulated dataset, we use the EIV model (in blue), and see that the distribution of estimated parameters using the proxies resembles the distribution using the latent data, although with slightly greater variance. Thus there is some small efficiency loss, but the EIV model clearly recovers unbiased estimates when its assumptions are met.

We also run the MO on the same simulated datasets in which we ran the EIV model. The

[5] In these simulations, $n = 200$, $(x_i, y_i) \sim \mathcal{N}(\mu, \Sigma)$, $\mu = (1, 1)$, $\Sigma = (1\ 0.4, 0.4\ 1)$, $\sigma_u^2 = 0.5$, $\sigma_v^2 = 0.5$.

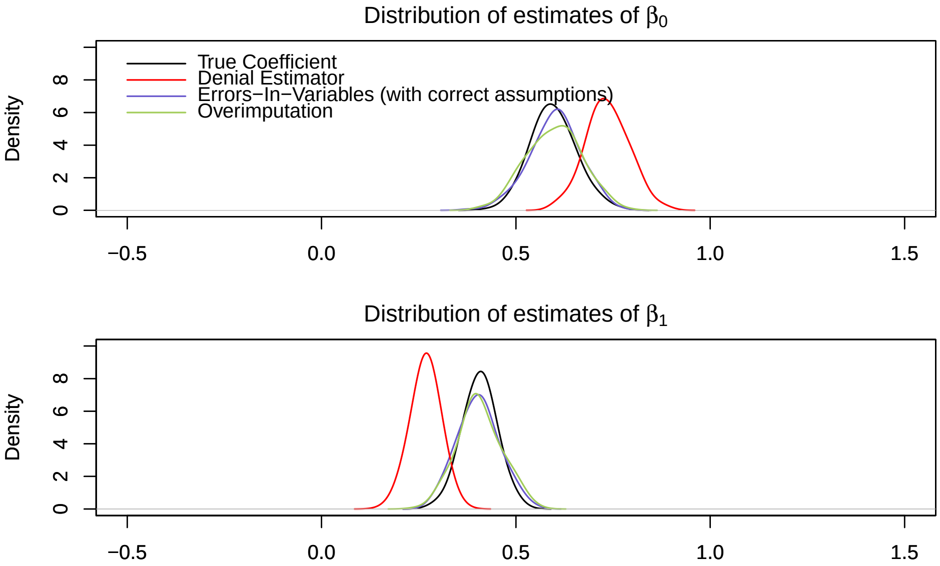


Figure 4: *Coefficients estimated from variables with measurement error (shown in red) attenuate the effect of the independent variable towards zero, and also bias the constant in compensation. The estimates recovered from the* EIV *model (in green) recover the true distribution, but are of course less efficient (slightly higher variance) than the original latent data (in black).*

distribution of coefficients (which we present below) recovers the distribution that would have been estimated if the latent data had been available. Thus, in the simple setting where the assumptions of the EIV model are met, our approach performs equivalently.

5.2 Robustness to Violating Assumptions

If we think of the coefficient on x_i as the ratio of $\text{cov}(x, y)$ to $\text{var}(x)$, then the attenuation in equation (32) is being driven by the fact that $\text{var}(w_1) > \text{var}(x)$ because of the added measurement error. Therefore $\text{var}(w_1)$ is not a good estimate of $\text{var}(x)$, even though $\text{cov}(w_1, y)$ *is* a good measure of $\text{cov}(x, y)$. With this in mind, the numerically simpler—but equivalent—one stage approach to the errors-in-variables model has a useful intuition. We substitute $\text{cov}(w_1, w_2)$ as an estimate of $\text{var}(x)$ because w_1, w_2 only covary through x. Thus we have as our estimate of the relationship:[6]

$$\hat{\delta} = \frac{\sum_i (w_{i1} - \bar{w}_1)(y_i - \bar{y})}{\sum_i (w_{i1} - \bar{w}_1)(w_{i2} - \bar{w}_2)} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y}) + u_i(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2 + u_i(x_i - \bar{x}) + v_i(x_i - \bar{x}) + u_i v_i}. \tag{33}$$

[6] In a multivariate setting this becomes $\hat{\delta} = (W_1' W_2)^{-1} W_1' Y$ where W_j is the set of regressors using the j-th proxy measure for x.

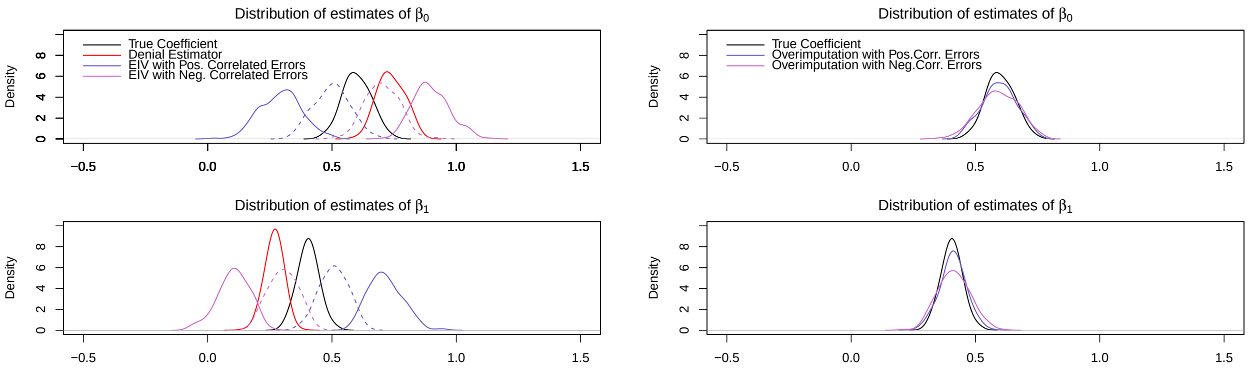
In order to recover the true relationship between x and y we need the last term in the numerator and the last three in the denominator to drop out of equation (33). To obtain a consistent estimate, then, EIV requires: (1) $\mathrm{E}(u_i{\cdot}y_i) = 0$, (2) $\mathrm{E}(u_i{\cdot}x_i) = 0$ and $\mathrm{E}(v_i{\cdot}x_i) = 0$, and (3) $\mathrm{E}(u_i{\cdot}v_i) = 0$. Indeed, when these conditions are not met the resulting bias in the EIV correction can easily be larger than the original bias caused by measurement error. However, as we now show in the following three subsections, MO is robust to violations of all but the last condition.

5.2.1 Measurement error correlated with y

The first of the conditions for EIV to work is that the measurement error is unrelated to the observed dependent variable. As an example of this problem, we might think that infant mortality is related to international aid because donors want to reduce child deaths. If countries receiving aid are intentionally underreporting infant mortality, to try to convince donors the aid is working, then the measurement error in infant mortality is negatively correlated with the dependent variable, foreign aid. If instead countries searching for aid are intentionally overreporting infant mortality as a stimulus for receiving aid, then measurement error is positively correlated with the dependent variable. Both scenarios are conceivable. This problem with the errors-in-variables approach is well known, because the errors-in-variables model has an instrumental variables framework, and this is equivalent to the problem of the instrument being exogenous of y in the more common usage of instrumental variables as a treatment for endogeneity.

In Figure 5(a) we demonstrate this bias with simulated data.[7] The violet densities show the distribution of parameter estimates when there is negative correlation of 0.1 (dashed) and 0.3 (solid) between the measurement error and the dependent variable. In the latter case the bias in the correction has exceeded the original bias from measurement error, still depicted in red. The blue densities show that positive correlation of the errors create bias of similar magnitude in the opposite direction. Again, the size of the bias can be greater than that originally produced by the measurement error we were attempting to correct. Moreover, the common belief with measurement issues is that any resulting bias attenuates the coefficients so that estimates are at least conservative,

[7] In these simulations, similar to previous, $n = 200$, $(x_i, y_i, u_i, v_i) \sim \mathcal{N}(\mu, \Sigma)$, $\mu = (1, 1, 0, 0)$, $\Sigma = (1\ 0.4\ 0\ 0, 0.4\ 1\ 0\ \rho, 0\ 0\ \sigma_u^2\ 0, 0\ \rho\ 0\ \sigma_v^2)$, $\sigma_u^2 = 0.5$, $\sigma_v^2 = 0.5$. Thus, the measurement errors are drawn at the same time as x_i and y_i with mean zero. While ρ allows the error, v_i, to covary with y_i, and across the simulations it is set as one of $\rho \in \{-0.3, -0.1, 0.1, 0.3\}$. The observed mismeasured variables are constructed as $w_{i1} = x_i + u_i,\ w_{i2} = x_i + v_i$.



(a) Estimates from Errors-in-Variables

(b) Estimates from Multiple Overimputation

Figure 5: *With data generated so that proxy variables are correlated with the dependent variable,* EIV *(left graphs) gives biased estimates whereas* MO *(right graphs) gives robust, unbiased estimates.*

however, here we see that the bias in the error-in-variables approach can actually exaggerate the magnitude of the effect.

We now analyze the same simulated datasets with MO. To apply the MO model, we estimate the measurement error variance from the correlation between the two proxies and leave the mean set to the better proxy. As Figure 5(b) indicates, MO recovers the distribution of coefficients for each of the data generation processes: The violet line represents the distribution when there is positive correlation while the blue line (barely visible under the other two) represents the distribution with negative correlation. All three distributions are close to each other and close to the true distribution in black using the latent data.

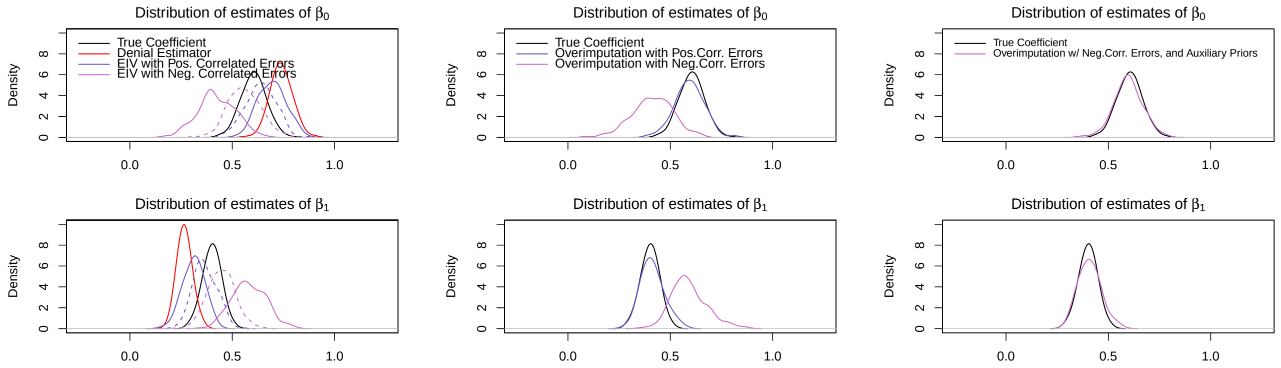
5.2.2 Measurement error correlated with the latent true value

The second requirement of the EIV model is that the measurement error is independent of the latent variable. If, for example, we believe that income is poorly measured, and wealthier respondents feel pressure to underreport their income while poorer respondents feel pressure to overreport, then the measurement error can be correlated with the latent variable.

In Figure 6(a) we demonstrate the bias this produces in EIV. Here, the error in w_{i2} is correlated with the latent x_i.[8] The biases are in the opposite directions as when the correlation is with y_i,

[8] Similar to the construction of the last simulations, we set $n = 200$, $(x_i, y_i, u_i, v_i) \sim \mathcal{N}(\mu, \Sigma)$, $\mu = (1, 1, 0, 0)$, $\Sigma = (1\ 0.4\ 0\ \rho, 0.4\ 1\ 0\ 0, 0\ 0\ \sigma_u^2\ 0, \rho\ 0\ 0\sigma_v^2)$, $\sigma_u^2 = 0.5$, $\sigma_v^2 = 0.5$ and sequencing $\rho \in \{-0.3, -0.1, 0.1, 0.3\}$ across sets of simulations.

although lesser in magnitude. Errors positively correlated with x_i lead to attenuated coefficients, and negatively correlated errors lead to overstated coefficients, as shown by the blue and violet distributions in Figure 6(a), respectively. Dashed lines are the result of small levels of correlations (± 0.1) and the solid lines a greater degree (± 0.3).



(a) Estimates from Errors-in-Variables (b) Multiple Overimputation (MO) (c) MO with Auxiliary Priors

Figure 6: *Here we show the estimates when the error in the instrument w_{i2} is correlated with the latent variable x_i. Positive (blue distributions) correlation leads to attentuated estimated effects in the errors-in-variables framework, and negative (violet) correlation exagerates the effect, as shown in the left. The* MO *estimates show no bias.*

The coefficients resulting from MO, with measurement error variance estimated from the correlation between the proxies, are contrasted in Figure 6(b). In the positive correlation case the distributions recover the same parameters. Because they sit on top of each other, only the simulations with the greatest correlation (± 0.3) are shown. For both parameters, with positive correlation, the MO estimates reveal no bias. With negative correlation, neither equation 27 nor 28 give good estimates of the measurement error. Here we see bias in the coefficients, similar to the EIV model. The key cause of the bias is the inability to judge the measurement error when the proxy is negatively correlated with the true latent x. If we had additional auxiliary information to set the prior variance for w_{i1} correctly, and used w_{i2} simply as a predictor, we would recover unbiased estimates with the MO model, as is shown in the simulations in 7(c). This follows the results seen previously in 1 and in figure 3 of Blackwell, Honaker and King (2015*b*).

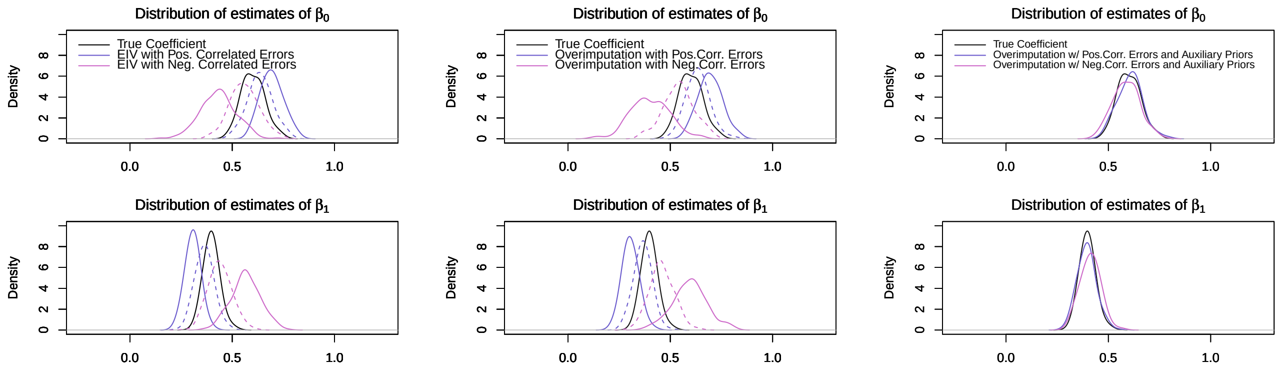
5.2.3 Measurement errors that covary across proxies

The final condition requires the errors in the proxies be uncorrelated. If all the alternate measures of the latent variable have the same error process then the additional measures provide no additional information. For example, if we believe GDP is poorly measured, it is not enough to find two alternate measures of GDP; we also need to know that those sources are not making the same errors in their assumptions, propagating the same errors from the same raw sources, or contaminating each other's measure by each making sure their estimates are in line with other published estimates. To the extent the errors in the alternate measures are correlated, then σ_{uv} will attenuate the estimate in the same fashion as σ_u^2 did originally.

Thus, we now simulate data where the measurement errors across alternate proxies are correlated.[9] Figure 7(a) shows positively (negatively) correlated errors lead to bias in the EIV estimates that are in the same (opposite) direction as the original measurement error. Intuitively, if the errors are perfectly correlated, both the original proxy, and the alternate proxy would be the exact same variable, and thus all of the original measurement error would return. Importantly, what we see is that this is a limitation of the data that MO cannot overcome when cell level priors are directly created from the observed data. As alternate proxies contain correlated errors, identifying the amount of the variance in the proxies by the correlation of the measures is misleading. Positive or negative correlation in the measurement errors leads respectively to under or over estimation of the amount of measurement error in the data, directly biasing results as in EIV. When cell priors are set by the use of auxiliary proxies, our method continues to require the measurement errors (although not the indicates themselves of course) be uncorrelated across alternate measures, so that it is possible to consistently estimate the degree of measurement error present in the data. Again, if there exists additional auxiliary information to set the prior variances, other than the proxies, the MO model has very small bias, as shown in simulations in figure 7(c), where we assume the true measurement error variance is known from auxiliary information about the measurement procedure.

Even in this most difficult of settings, MO remains robust. In another set of simulations, we compare how various estimators perform when both proxies are correlated with y_i. Allowing these

[9] Here we set $n = 200$, $(x_i, y_i, u_i, v_i) \sim \mathcal{N}(\mu, \Sigma)$, $\mu = (1, 1, 0, 0)$, $\Sigma = (1\ 0.4\ 0\ 0, 0.4\ 1\ 0\ 0, 0\ 0\ \sigma_u^2\ \rho, 0\ 0\ \rho\ \sigma_v^2)$, $\sigma_u^2 = 0.5$, $\sigma_v^2 = 0.5$ and sequencing $\rho \in \{-0.3, -0.1, 0.1, 0.3\}$ across sets of simulations.



(a) Estimates from Errors-in-Variables (b) Multiple Overimputation (MO) (c) MO with Auxiliary Priors

Figure 7: *With data generated so that proxy variables have measurement error correlated with each other (so that new information is not availble with measures) both* EIV *(left graphs) and* MO *(right graphs) gives biased estimates.*

simulations to vary the amount of correlation gives an indication of how various estimators perform in this difficult situation.[10] Figure 8 shows that MO outperforms EIV at every level of this correlation. When the dependence between the error and y_i is weak, MO almost matches its zero-correlation minimum. Thus, MO appears to be robust to even moderate violations of the these assumptions, especially when compared with other measurement error approaches. Interestingly, the denial estimator can perform better than all estimators under certain conditions, yet these conditions depend heavily on the parameters of the data. If we change the effect of x_i on y_i from negative to positive, the performance of the denial estimator reverses itself. Since we obviously have little knowledge about all of these parameters *a priori*, the denial estimator is of little use.

Since there are gold-standard data in these simulations, we can also investigate the performance of simply discarding the mismeasured data and running MI. As expected, MI is unaffected by the degree of correlation since it disregards the correlated proxies. Yet these proxies have *some* information when the correlation is around zero and, due to this, MO outperforms MI in this region. As the correlation increases, though, it becomes clear that simply imputing the mismeasured cells

[10] These simulations follow the pattern above except they include a perfectly measured covariate, z_i, which determines which observations are selected for mismeasurement. Thus, we have $(x_i, y_i, z_i, u_i, v_i,) \sim N(\mu, \Sigma)$, with $\mu = (1, 1, -1, 0, 0)$ and $\Sigma = (1\ \sigma_{xy}\ \ -0.4\ 0\ 0, \sigma_{xy}\ 1\ \ -0.2\ \rho\sigma_u\ \rho\sigma_v, -0.4\ \ -0.2\ 1\ 0\ 0\ 0\ \rho\sigma_u\ 0\ \sigma_u^2\ 0, 0\ \rho\sigma_v\ \ 0\ 0\ \sigma_v^2)$ with $\sigma_u^2 = 0.5$ and $\sigma_v^2 = 0.75$. We ran simulations at both $\sigma_{xy} = 0.4$ and $\sigma_{xy} = -0.4$. Each observation had probability $\pi_i = (1+e^{3.5+2z})^{-1}$, which has a mean of 0.25. We used the multiple proxies approach to estimating the measurement error. For EIV, we use applied the model as if the entire variable were mismeasured.

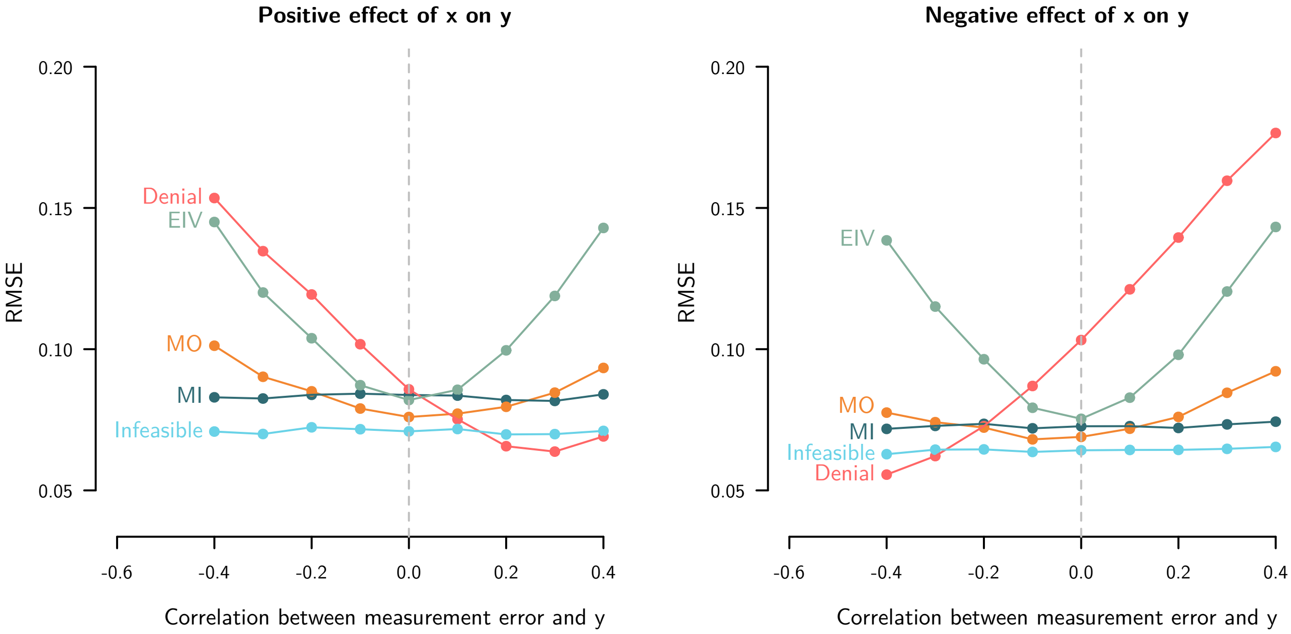


Figure 8: *Root mean squared error for various estimators with data generated so that each proxy variable has measurement error correlated with the dependent variable. On the left, x_i has a positive relationship with y_i and on the right, it has a negative effect. Note that both* EIV *(green) and* MO *(orange) perform worse as the correlation moves away from zero, but* MO *always performs better. The denial estimator can actually perform well in certain situations, yet this depends heavily on the direction of the relationship. Both the infeasible estimator and* MI *are unaffected by the amount of correlation.*

has more desirable properties. Of course, with such high correlation, we might wonder if these are actually proxies in our data or simply new variables.

These simulations give key insights into how we should handle data measured with error. MO is appropriate when we have a variable that we can reasonably describe as a proxy—that is, having roughly uncorrelated, mean-zero error. Even if these assumptions fail to hold exactly, MO retains its desirable properties. In situations where we suspect that the measurement error on all of our proxies has moderate correlation with other variables in the data, it may be wiser to treat the mismeasurement as missingness and use multiple imputation. Of course, this approach assumes there exist gold-standard data, which may be scarce.

6 The Number of Data sets to Overimpute

Multiple imputation relies on simulation to reflect the uncertainty due to missing data and requires a sufficient number of in order to do so accurately. The simulations in this case are the number of imputed datasets. There is a large body of research analyzing the performance of multiple imputation as the number of imputed datasets grows, with a general consensus that somewhere between 5 and 50 imputations is sufficient to obtain approximately valid confidence intervals. With multiple overimputation, we might worry that the increased demands on the imputation model might require additional imputations. In this section we perform a simulation experiment to see how our uncertainty estimates perform as a function of the number of imputations.

To do so, we use the basic design from BHK Section 2.4, vary the number of imputed datasets, and assess the confidence interval coverage for the estimated coefficients in the outcome model.[11] We present the results in Table 6, which shows that the qualitative results from MI carry over to MO. In terms of both point estimates and confidence interval coverage, there a only minor improvements beyond 10-25 imputed datasets. The lowest number of imputations can lead to confidence intervals that are both too narrow and too wide.

B	x_i Coverage	x_i MSE	z_i Coverage	z_i MSE
3	0.931	0.0162	0.936	0.0147
5	0.945	0.0157	0.959	0.0145
10	0.944	0.0152	0.954	0.0135
25	0.945	0.0147	0.957	0.0132
50	0.944	0.0147	0.953	0.0131
100	0.948	0.0146	0.951	0.0131

Table 1: This table shows the coverage and mean square error for the MO in the simulations from BHK Section 2.4 with 1,000 simulations as we vary the number of imputed datasets. The gains to both coverage and MSE are minimal after 10-25 imputations.

The amount of measurement error in a variable increases the amount of missing information, so we would expect the performance of MO to vary as a function of the amount of measurement error in the data. It is instructive, then, to investigate the coverage of the MO approach In Table 6, we display the coverage of nominal 95% confidence intervals as a function of both B and ρ. Here we are using the same simulation design as in BHK Section 3 using the correct value of ρ in the imputation.

[11] Here we set the assumed amount of measurement error to the correct level.

As is clear from the Table, our intuition is correct—the coverage performance of the MO weakens as the amount of measurement error increases. In fact, when the $\rho > 0.5$, the confidence intervals become strikingly uninformative. This makes sense considering the fact that less than half of our mismeasured variable's variance is due to the true value. When ρ is at more reasonable values, we see similar results to Table 6 in that the marginal value of increasing B is fairly low after 10-25 imputations.

	ρ								
B	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3	0.953	0.936	0.936	0.908	0.920	0.919	0.915	0.912	0.907
5	0.953	0.942	0.951	0.955	0.953	0.961	0.965	0.965	0.977
10	0.945	0.958	0.952	0.962	0.975	0.991	0.987	0.994	0.997
25	0.946	0.967	0.968	0.976	0.986	0.995	0.997	0.999	1.000
50	0.952	0.958	0.970	0.978	0.987	0.999	0.999	0.999	1.000
100	0.945	0.965	0.968	0.975	0.994	0.997	1.000	1.000	1.000

Table 2: This table shows the coverage for a nominal 95% confidence interval for the slope in simulation with the setup in BHK Section 3, varying the amount of error in the proxy (ρ) and the number of imputations used (B). In these simulations there is very little information for the imputations, so as the error dominates the model, the confidence intervals become non-informative. But with limited amounts of measurement error, even as small as 5 imputed datasets gives good coverage.

7 Concluding Remarks and Further Extensions

This paper, taken together with its companion Blackwell, Honaker and King (2015*b*), offer a new approach to preprocessing data so that applied researchers can attend to the serious problems of missing data and measurement error scattered throughout most data sets used in the social sciences. It should be attractive to applied researchers especially since, after attending to these issues, they can apply whatever method they would have in the absence of these problems. This enables social scientists and others to get back to the substantive problems that drove them to analyze their data to begin with.

The methods offered here are implemented in Amelia (Honaker, King and Blackwell, 2010), and can be easily combined with other techniques as part of the same algorithm, most of which we have not touched on in these two papers. For example, researchers can choose to put prior values on individual cell values, include features for time series or time series cross-sectional data, implement transformations, include many types of prior information, and evaluate results and judge model fit using a wide range of diagnostics.

References

Bityukov, SI, VV Smirnova, NV Krasnikov and VA Taperechkina. 2006. Statistically dual distributions in statistical inference. In *Statistical Problems in Particle Physics, Astrophysics and Cosmology: proceedings of PHYSTAT05, Oxford, UK, 12-15 September 2005*. pp. 102–105. http://arxiv.org/abs/math/0411462v2.

Blackwell, Matthew, James Honaker and Gary King. 2015*a*. "Replication data for: A Unified Approach To Measurement Error And Missing Data: Details And Extensions.". http://dx.doi.org/10.7910/DVN/29610 IQSS Dataverse Network [Distributor].

Blackwell, Matthew, James Honaker and Gary King. 2015*b*. "A Unified Approach to Measurement Error and Missing Data: Overview." *Sociological Methods and Research* .

Honaker, James and Gary King. 2010. "What to do About Missing Values in Time Series Cross-Section Data." *American Journal of Political Science* 54(2, April):561–581. http://gking.harvard.edu/files/abs/pr-abs.shtml.

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

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(3):815–835.

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(1, March):49–69. http://gking.harvard.edu/files/abs/evil-abs.shtml.

Newey, Whitney K. and Daniel McFadden. 1994. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, ed. Robert F. Engle and Daniel L. McFadden. Elsevier pp. 2111–2245.

Schafer, Joseph L. 1997. *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Schafer, Joseph L. and Maren K. Olsen. 1998. "Multiple Imputation for multivariate Missing-Data Problems: A Data Analyst's Perspective." *Multivariate Behavioral Research* 33(4):545–571.

Seaman, Shaun, John Galati, Dan Jackson and John Carlin. 2013. "What Is Meant by "Missing at Random"?" *Statistical Science* 28(2):257–268.

Stefanski, L. A. 2000. "Measurement Error Models." *Journal of the American Statistical Association* 95(452):1353–1358.

Wu, C F Jeff. 1983. "On the Convergence Properties of the EM Algorithm." *The Annals of Statistics* 11(1, March):95–103.