MULTIVARIATE BEHAVIORAL RESEARCH
https://doi.org/10.1080/00273171.2019.1657790

Routledge
Taylor & Francis Group

An Approach to Addressing Multiple Imputation Model Uncertainty Using Bayesian Model Averaging

David Kaplan and Sinan Yavuz

University of Wisconsin–Madison

ABSTRACT

This paper considers the problem of imputation model uncertainty in the context of missing data problems. We argue that so-called "Bayesianly proper" approaches to multiple imputation, although correctly accounting for uncertainty in imputation model parameters, ignore the uncertainty in the imputation model itself. We address imputation model uncertainty by implementing Bayesian model averaging as part of the imputation process. Bayesian model averaging accounts for both model and parameter uncertainty, and thus we argue is fully Bayesianly proper. We apply Bayesian model averaging to multiple imputation under the fully conditional specification approach. An extensive simulation study is conducted comparing our Bayesian model averaging approach against normal theory-based Bayesian imputation not accounting for model uncertainty. Across almost all conditions of the simulation study, the results reveal the extent of model uncertainty in multiple imputation and a consistent advantage to our Bayesian model averaging approach over normal-theory multiple imputation under missing-at-random and missing-completely-at random in terms of Kullback-Liebler divergence and mean squared prediction error. A small case study is also presented. Directions for future research are discussed.

KEYWORDS

Missing data; Bayesian model averaging; Multiple imputation

Introduction

Multiple imputation (MI) (Rubin, 1987) is arguably the gold-standard in addressing problems of missing data. Taken as a generic idea, multiple imputation can be implemented under three general approaches (Enders, 2010; Little & Rubin, 2002; van Buuren, 2012): (a) monotone data imputation, (b) joint modeling, and (c) fully conditional specification. For monotone data imputation, a series of univariate methods are constructed to impute the missing data. The joint modeling and fully conditional specification approaches are for more general patterns of missing data. Specifically, under the joint modeling approach, imputations are obtained from a multivariate model applied to the data. Under the fully conditional specification approach (also referred to as *chained equations* or *sequential regressions*), imputation is conducted by specifying conditional univariate regressions across iterated conditional models (van Buuren, 2012).

The orientation of this paper is from the viewpoint of an imputer who is tasked with providing an imputed data set to a secondary user. This orientation is quite common in large-scale survey research. Thus, we situate our discussion of multiple imputation within the framework of *congenial* missing data problems. The concept of congeniality in missing data problems was introduced by Meng (1994, see also Rubin, 1996; Kaplan & Su, 2018). In outlining the steps in conducting a large-scale survey, Meng (1994) pointed out that each step in the construction of a large-scale survey inherits information from the previous step. That is, the data file that a researcher uses is the result of a set of design steps which includes, in important ways, decisions that are made regarding the imputation of missing data. In many cases, as Meng (1994) notes, the imputer charged with decisions regarding missing data imputation has little or no contact with the secondary user of the data. Thus, if analysts are interested in conducting a secondary statistical analysis using the data, then their statistical models might have little in common with the model used to impute the missing data and this "disconnect" can lead to serious biases. Quoting Meng (1994, p. 539)

> "… *uncongeniality* … essentially means that the analysis procedure does not correspond to the imputation model. The uncongeniality arises when

CONTACT David Kaplan david.kaplan@wisc.edu University of Wisconsin-Madison, Educational Psychology, 1025 W. Johnson St., Madison, WI 53706, USA.

> the analyst and the imputer have access to different amounts and sources of information, and have different assessments (e.g., explicit model, implicit judgement) about both responses and non-responses. If the imputer's assessment is far from reality, then, as Rubin (1995)[1] wrote, "all methods for handling nonresponse are in trouble" based on such an assessment; all statistical inferences need underlying key assumptions to hold at least approximately. If the imputer's model is reasonably accurate, then following the multiple-imputation recipe prevents the analyst from producing inferences with serious nonresponse biases."

The problem of uncongeniality has led to the general principle that one should include as many variables as possible in the imputation model for the missing data (see e.g. Rubin, 1996). However, to quote Murray (2018, pg. 7),

> "Unless the analyst is the imputer, congeniality is less a condition we should try to satisfy than one we should try to fail gracefully … "

That is, although we recognize that congeniality is a critical issue, for the present paper, we are focusing more on the problem of capturing imputation model uncertainty than having captured the correct imputation model.

Proper versus Bayesianly proper imputations

Overarching the various methods of multiple imputation, Rubin (1987) introduced the idea of so-called *proper* imputations. The notion of proper imputations relates to the asymptotic properties of the statistics of interest obtained from the imputation process over an infinitely large number of imputed data sets. Following Rubin (1987), let m represent the number of multiply imputed (completed) data sets ($m = 1, 2, ..., M$); $\hat{Q}_m$ represent a statistic such as a mean which estimates the corresponding parameter Q on each of the m completed data sets; and U_m be the sampling variability of $\hat{Q}_m$ for each of the m completed data sets. Furthermore, let $\bar{Q}_m$ and let $\bar{U}_m$ be the corresponding averages across the data sets, and let B_m be the variance among the m completed data sets.

Next, consider the notion of *randomization-based* inference, in which the question concerns the behavior of the statistic of interest under a randomization distribution and a specified null hypothesis. Under randomization-based inference, proper imputations imply that the statistics $(\bar{Q}_\infty, \bar{U}_\infty, B_\infty)$ yield valid inferences for the corresponding complete-data values $(\hat{Q}, U)$. Valid inferences can be obtained from Rubin's (1987) combining rules. In essence, if Rubin's (1987) combining rules yield consistent and asymptotically normal estimators then the imputations are proper (Rubin, 1987).

The randomization-based justification for multiple imputation outlined by Rubin (1987) is relevant for frequentist or design-based frameworks for statistical inference. Rubin also provides a Bayesian justification for multiple imputation, and this is more fully discussed in Schafer (1997). Under the Bayesian framework, let Y^{mis} represent observations on Y that are missing, and let Y^{obs} represent observations on Y that are observed. Then, assuming some complete data model and priors for the model parameters, say θ, the posterior predictive distribution of the missing data can be written as

$$p(Y^{mis}|Y^{obs}) = \int p(Y^{mis}|Y^{obs}, \theta)p(\theta|Y^{obs})d\theta, \qquad (1)$$

where $p(\theta|Y^{obs})$ is the posterior distribution of the parameters given the observed data. The posterior distribution of the parameters can, in turn, be decomposed into the product of the data distribution and the prior distribution of the model parameters via Bayes' theorem – namely,

$$p(\theta|Y^{obs}) \propto p(Y^{obs}|\theta)p(\theta), \qquad (2)$$

where $p(Y^{obs}|\theta)$ is the distribution of the observed data given the model parameters θ, and $p(\theta)$ is the prior distribution on the model parameters. Thus, as long as imputations are the result of independent realizations of equation (1), they are said to be *Bayesianly proper* (Schafer, 1997).

Bayesianly proper imputations address uncertainty in the imputation process through the prior distributions placed on the model parameters in equation (2). However, model parameters are not the only sources of uncertainty in the imputation process. The central argument of this paper, more fully developed below, is that there is uncertainty in the choice of the imputation model itself and this uncertainty is not being accounted for in conventional Bayesianly proper multiple imputation. Thus, for multiple imputation to be fully Bayesianly proper, it is necessary to account for imputation model uncertainty. We address imputation model uncertainty by adding a Bayesian model averaging component to multiple imputation.

The organization of this paper is as follows. In the next section, we provide an overview of Bayesian model averaging following closely the discussion in Hoeting, Madigan, Raftery, and Volinsky (1999) and

[1] *Note: footnote ours*. This paper was eventually published as Rubin (1996).

more recently, Kaplan and Lee (2018). This is then followed by a review of the method of chained equations that we use as our framework for multiple imputation. Next, we introduce our approach that combines Bayesian model averaging with multiple imputation via chained equations which we refer to as *miBMA – Multiple Imputation under Bayesian Model Averaging*. This is followed by our simulation study design, followed by the results. We then provide a small case study using real data from the 2015 cycle of the Program on International Study Assessment (PISA) (OECD, 2016). Finally, we close the paper with a discussion of possible extensions of our approach to imputation under the generalized linear model.

Overview of Bayesian model averaging

As mentioned earlier, Bayesian imputation under the normal linear model addresses uncertainty only through the specification of prior distributions on the model parameters and does not account for uncertainty in the choice of imputation models. To account for parameter and model uncertainty requires the method of *Bayesian model averaging*.

Bayesian model averaging has had a long history of theoretical developments and practical applications. Early work by Leamer (1978) laid the foundation for Bayesian model averaging. Fundamental theoretical work on Bayesian model averaging was conducted in the mid-1990s by Madigan and his colleagues (e.g., Madigan & Raftery, 1994; Raftery, Madigan, & Hoeting, 1997; Hoeting et al., 1999). Additional theoretical work was conducted by Clyde (1999). Draper (1995) has discussed how model uncertainty can arise even in the context of experimental designs, and Kass and Raftery (1995) provide a review of Bayesian model averaging and the costs of ignoring model uncertainty. A more recent review of the general problem of model uncertainty can be found in Clyde and George (2004). Bayesian model averaging has been implemented in the R software programs "BMA" (Raftery, Hoeting, Volinsky, Painter, & Yeung, 2015) and "BMS" (Zeugner & Feldkircher, 2015).

In addition to theoretical developments, Bayesian model averaging has been applied to a wide variety of content domain. A perusal of the extant literature shows Bayesian model averaging applied to economics (e.g., Fernández, Ley, & Steel, 2001), bioinformatics of gene expression (e.g., Yeung, Bumgarner, & Raftery, 2005), weather forecasting (e.g., Sloughter, Gneiting, & Raftery, 2013), and causal inference within the propensity score framework (Kaplan & Chen, 2014; Zigler & Dominici, 2014), to name just a few. A recent extension of Bayesian model averaging to structural equation modeling can be found in Kaplan and Lee (2015), and an overview of Bayesian model averaging with applications to education policy research can be found in Kaplan and Lee (2018).

Bayesian model averaging: Methods

Following Madigan and Raftery (1994, see also; Kaplan & Lee, 2018), consider a quantity of interest such as the prediction of a missing value. We will denote this quantity as Υ. Next, consider a set of competing imputation models M_k, $k = 1, 2, \ldots, K$ that are not necessarily nested. The posterior distribution of Υ given data y can be written as a mixture distribution,

$$p(\Upsilon|y) = \sum_{k=1}^{K} p(\Upsilon|M_k)p(M_k|y), \tag{3}$$

where $p(M_k|y)$ is the posterior probability of model M_k given the data y written as

$$p(M_k|y) = \frac{p(y|M_k)p(M_k)}{\sum_{l=1}^{K} p(y|M_l)p(M_l)}, \quad l \neq k. \tag{4}$$

where the first term in the numerator on the right-hand side of equation (4) is the probability of the data given model k, also referred to as the *integrated likelihood* written as

$$p(y|M_k) = \int p(y|\theta_k, M_k)p(\theta_k|M_k)d\theta_k, \tag{5}$$

where $p(\theta_k|M_k)$ is the prior distribution of the parameters θ_k under model M_k (Raftery et al., 1997). The posterior model probabilities can be considered mixing weights for the mixture distribution given in equation (3) (Clyde & Iversen, 2015). The second term $p(M_k)$ on the right-hand side of equation (4) is the prior imputation model probability for model k, allowing each imputation model to have a different prior probability based on past performance of that imputation model or a belief regarding which of the models might be the true model. The denominator of equation (4) ensures that $p(M_k|y)$ integrates to 1.0.

An important feature of equation (4) is that $p(M_k|y)$ captures the posterior (post-data) uncertainty in a given imputation model and will likely vary across models. Herein lies the problem of model selection in the context of multiple imputation; given the choice of a particular imputation model, the analyst effectively ignores the uncertainty in other models

that could have been used for imputation. Of course, equation (4) could be used as a method for model selection, simply choosing the imputation model with the largest posterior model probability. However, we argue that to settle on a particular imputation model – even one with the largest posterior model probability among a set of competing models – still ignores the uncertainty inherent in the model choice problem. As such, we hypothesize that our approach to multiple imputation based on using BMA to address imputation model uncertainty will perform better than choosing the model with the largest posterior model probability or simply ignoring model uncertainty altogether, as is the case with conventional Bayesian MI. Indeed, Raftery, Madigan, and Hoeting (1997) show that BMA provides superior predictive validity compared to that of any single model measured by a logarithmic scoring rule.

Parameter and model priors for BMA

In the context of BMA, we are required to specify priors on the model parameters as well as on the model space. For model parameters, we use the default priors in the "BMA" package – namely, the *unit information prior* (Kass & Wasserman, 1995). Following Raftery (1998, pp 3–6, see also Kaplan & Lee, 2018), the unit information prior is a weakly informative prior that is diffused over the region of the likelihood where parameter values are considered mostly plausible, but not overly spread out. This is accomplished by forming the prior based on the maximum likelihood estimate of the parameter mean, with variance equal to the expected information from one observation.[2] The default prior on the model space is $1/M$, where M is the number of models, reflecting the belief that no imputation model is to be favored as the true model to be used for imputation a priori.

Computational issues

As pointed out by Hoeting et al. (1999), Bayesian model averaging is difficult to implement. In particular, they note that the number of terms in equation (3) can be quite large, the corresponding integrals are hard to compute, the specification of $p(M_k)$ may not be straightforward, and choosing the class of models to average over is also challenging. To address the problem of computing equation (5) the Laplace method, which has been used productively for the computation of Bayes factors (Kass & Raftery, 1995), can be used and this will lead to a simple BIC approximation under certain circumstances (Tierney & Kadane, 1986; Raftery, 1996).[3]

To address the problem of reducing the overall number of imputation models, we use the so-called *Occam's window* algorithm (Madigan & Raftery, 1994) as implemented in the R program "BMA" (Raftery et al., 2015). Following closely the discussion given in Raftery et al. (1997), in the context of multiple imputation, one might start with a very large number of predictors; but the goal is to narrow down this large set of predictors to a small number of predictors that provide accurate predictions of the missing data. As noted in the earlier quote by Hoeting et al. (1999), the concern in drawing inferences from a single "best" imputation model is that the choice of a single set of predictors of the missing data ignores uncertainty in model selection.

The algorithm proceeds in two steps (Raftery et al., 1997). In the first step, imputation models are eliminated from equation (3) if they predict the missing data much less well than the model that provides the best predictions based on a "caliper" value C chosen in advance by the analyst. The caliper C sets the "width" of Occam's window. Formally, consider again a set of imputation models M_k, $k = 1...K$. Then, the set A' is defined as

$$A' = \left\{ M_k : \frac{max_l\{p(M_l|y)\}}{p(M_k|y)} \leq C \right\}. \tag{6}$$

Equation (6) compares the imputation model with the largest posterior model probability, $max_l\{p(M_l|y)\}$, to a given model $p(M_k|y)$. If the ratio in equation (6) is greater than the chosen value C, then it is discarded from the set A' of models to be included in the model averaging. Notice that the set of models contained in A' is based on Bayes factor values.

The set A' now contains imputation models to be considered for model averaging. In the second, optional step, imputation models are discarded from A' if they receive less support from the data than

[2]Note that the unit information prior is equivalent to Zellner's g-prior (Zellner et al. 1986), where $g = 1/N \ldots$, and where N is the sample size. See also Fernández et al. (2001).

[3]The Laplace method of integrals is based on a Taylor expansion of a function $f(u)$ of a q-dimensional vector u. The approximation is $\int e^{f(u)}du \simeq 2(\pi)^{q/2}|A|^{1/2}\exp\{f(u^*)\}$, where u^* is the value of u at which f attains its maximum, and A is minus the inverse of the Hessian of f evaluated at u^*. Following Raftery (1996, pg. 253), when the Laplace method is applied to equation (5), we obtain the approximation $p(y|M_k) \simeq (2\pi)^{q_k}|A_k|^{1/2}p(y|\tilde{\theta}_k, M_k)p(\tilde{\theta}_k, M_k)$, where q_k is the dimension of θ_k, $\tilde{\theta}_k$ is the posterior mode of θ_k, and A_k is minus the inverse of the Hessian of $\log\{p(y|\theta_k, M_k)p(\theta_k|M_k)\}$, evaluated at the posterior mode $\tilde{\theta}_k$.

simpler sub-models. Formally, models are further excluded from equation (3) if they belong to the set

$$B = \left\{ M_k : \exists M_l \in A', M_l \subset M_k, \frac{p(M_l|y)}{p(M_k|y)} > 1 \right\}. \quad (7)$$

Equation (7) states that there exists a model M_l within the set A' and where M_l is simpler than M_k. If a complex model receives less support from the data than a simpler sub-model – again based on the Bayes factor – then it is included in B. Notice that the second step corresponds to the principal of Occam's razor (Madigan & Raftery, 1994).

With step 1 and step 2, the computational problem of BMA is simplified by replacing equation (3) with

$$p(\Upsilon|y, A) = \sum_{M_k \in A} p(\Upsilon|M_k, y) p(M_k|y, A), \quad (8)$$

where A is the relative complement of A' and B That is, the imputation models under consideration for Bayesian model averaging are those that are in A' but not in B.

Madigan and Raftery (1994) then outline an approach to the choice between two models to be considered for Bayesian model averaging. To make the approach clear, consider the case of just two models M_1 and M_0, where M_0 is the simpler of the two models. This could the case where M_0 contains fewer predictors than M_1 in a regression analysis. In terms of log-posterior odds, if the log-posterior odds are positive, indicating support for M_0, then we reject M_1. If the log-posterior odds are large and negative, then we reject M_0 in favor of M_1. Finally, if the log-posterior odds lie in between the pre-set criterion, then both models are retained.

Multiple imputation via chained equations

Our approach for combining BMA and multiple imputation lies within the chained equations framework developed by van Buuren (2012) and implemented in the R program "mice" (van Buuren & Groothuis-Oudshoorn, 2010). The chained equations approach works as follows: First, a target variable is chosen among the set of variables to be imputed and this will be the first variable that the algorithm encounters containing missing data. The missing data on this first target variable is replaced by a random value from the observed data on that target variable. The algorithm then proceeds to assign these "placeholder" values to each variable in the set to be imputed. After the placeholders are assigned, the chained equations algorithm chooses the first variable as the dependent variable and all other variables as predictor variables and runs the chosen imputation method – in our case Bayesian linear regression under the normal model (Schafer, 1997).[4] After imputation of the missing data for the first variable, the algorithm moves to the next variable in the dataset and again runs the imputation method of choice. After the algorithm cycles through all the variables, the procedure is repeated based on a pre-set number of iterations. Once all iterations are completed, the resulting imputed file constitutes the first imputed data set. The process then repeats itself until the desired number of imputations are obtained. The algorithm can run these sequences simultaneously m number of times obtaining m imputed data sets. This is the algorithm used in "mice" which we will use for our analyses below.

Bayesian imputation under the normal linear model

For this paper, we will follow closely the notation of van Buuren (2012). Let X_{obs} represent observed predictors, X_{mis} represent missing predictors, y_{obs} represent the observed outcome, and y_{mis} represent the missing outcome. Further, imputed values of X and y will be represented as $\dot{X}$ and $\dot{y}$. The "mice" package uses standard non-informative priors for each parameter (van Buuren, 2012). As noted earlier, after estimating $\dot{y}$, the value imputed for the missing case and all subsequent missing cases in the data constitutes the first iteration. Bayesian imputation under the normal linear model using the "mice.impute.norm" function in the program "mice" proceeds as follows:

1. Obtain the cross products matrix $S = X'_{obs}X_{obs}$.
2. Calculate $V = [S + diag(S)\kappa]^{-1}$, where κ is a constant ridge parameter fixed to be close to zero. For this paper, $\kappa = 0.0001$.
3. Calculate $\hat{\beta} = VX'_{obs}y_{obs}$.
4. Draw a random variate $\dot{g} \sim \chi^2_\nu$ with $\nu = n_1 - q$, where n_1 is the number of observed rows in X and q is the number of variables in X.
5. Calculate $\dot{\sigma}^2 = (y_{obs} - X_{obs}\hat{\beta})'(y_{obs} - X_{obs}\hat{\beta})/\dot{g}$.
6. Draw q independent $N(0, 1)$ variates and arrange in the vector $\dot{z}_1$.
7. Calculate $V^{1/2}$ via a Cholesky decomposition.
8. Calculate $\dot{\beta} = \hat{\beta} + \dot{\sigma}\dot{z}_1 V^{1/2}$.

[4]Note that the "mice" program is flexible enough to allow different imputation methods to be chosen for different scales of variables. We address this issue and its implications for missing data with Bayesian model averaging in the Discussion section.

9. Draw n_0 independent $N(0, 1)$ variates and arrange in the vector $\dot{z}_2$, where n_0 is the number of missing rows in X.
10. Calculate the n_0 values $\dot{y} = X_{mis}\dot{\beta} + \dot{z}_2\dot{\sigma}$.

For this paper, we refer to Bayesian multiple imputation under the normal linear model as *NORM*.

FCS v. Joint modeling

Our choice for embedding BMA within the FCS framework is primarily due to the flexibility of the approach, and this is especially true when missing data appear in variables with very different metrics – e.g. missing on binary or polytomous variables. We discuss future research on this topic in the Discussion section, but for this paper, we consider only continuous normal variables. In addition, our view is that there is model uncertainty present for any given target variable chosen for imputation. Thus, our view is that our approach would capture a greater amount of model uncertainty than that which could be captured by employing BMA under multivariate imputation.

It is worth considering, however, the conditions under which multiple imputation under the multivariate model and under FCS provide comparable results. In a study of the stationary distribution of iterative imputation (i.e. FCS), Liu, Gelman, Hill, Su, and Kropko (2014) point out that the stationary distribution resulting from the Bayesian approach to FCS will not necessarily provide results that converge to any multivariate distribution. Liu et al. (2014) argue that if the families of the conditional models used in FCS are *compatible*, then FCS is asymptotically equivalent to the joint Bayesian model as long as the MCMC algorithm converges to a stationary distribution for each conditional model. Under that condition, Rubin's combining rules (Rubin, 1987) are valid for conditional models.

As defined by Liu et al. (2014), compatibility refers to the ability to map the parameters of the iterative model to the parameters of the joint model. If this mapping cannot be accomplished, then the iterative and joint models are not compatible. However, in that case, what is essential is that the conditional models are valid for the target variables of imputation. In Example 2 of Liu et al. (2014) they show that for continuous data, the parameters of the FCS models can be mapped onto the parameters of the joint model, and hence the distributions under FCS are compatible to those under the joint Bayesian models. In general, however, it is difficult to establish compatibility when the conditional models are quite different for different target variables (e.g., normal, logistic, multinomial, etc.). For the present study, however, we generate data according to a multivariate normal distribution and then use FCS with BMA under the normal model. Moreover, we monitor convergence of the algorithm at each step. Thus, we argue that our addition of BMA to FCS is compatible to what would be obtained under a multivariate model.

Multiple imputation under Bayesian model averaging

In this section, we outline our proposed *miBMA* approach. As noted earlier, we implement our approach using Bayesian normal linear regression implemented in the "mice.impute.norm" function in "mice". It should be noted that an approach similar to the method in this paper was proposed by Mitra and Dunson (2010).[5] In their paper, Mitra and Dunson focused on using BMA within a stochastic search variable selection (SSVS) (George & McCulloch, 1993) algorithm to conduct variable selection with data that are missing on the predictors. They argued that SSVS cannot be directly implemented when there is missing data in the predictor set, and common approaches such as listwise deletion are not efficient and can introduce bias. To address this issue, Mitra and Dunson (2010) developed an extension of the SSVS algorithm that simultaneously imputes missing data and conducts Bayesian variable selection via BMA. Their results using simulation and real data analyses showed the benefits of model averaging over imputation models in terms of out-of-sample predictive performance.

Of relevance to our paper, Mitra and Dunson (2010) did not propose BMA as a method for multiple imputation, per se. They do, however, allude to our approach by noting that one could impute missing data via MCMC at each step of the SSVS algorithm, thus accounting for uncertainty in the predictors included in the model. Our paper differs from Mitra and Dunson (2010) primarily in that we are focusing on BMA as a *method* for creating a multiply imputed data sets within a fully chained equations framework prior to any data analysis and not proposing BMA as a method for variable selection within any specific model. Thus, as we discuss in the introduction, we are drawing a distinction between the *imputer* who creates

[5]We thank an anonymous reviewer for bringing this paper to our attention.

the data set, and the *analyst* who is proposing a model for a specific substantive question (see, Meng, 1994).

The chained equations approach just described accounts for parameter uncertainty, and as discussed earlier, is Bayesianly proper (Schafer, 1997). However, as we argued in the introduction of the paper, Bayesian multiple imputation approaches do not account for uncertainty in the imputation model and hence are not fully Bayesianly proper. Our approach, in its essence, simply adds a Bayesian model averaging component to each cycle of the chained equations approach. Thus, as each variable takes its turn as the target variable for imputation, Bayesian model averaging is applied to the imputation model for that target variable. In doing so, imputation model uncertainty (as well as parameter uncertainty) is accounted for across all variables and iterations.

More specifically, the *miBMA* algorithm can be outlined as follows where steps 3 and 4 indicate our *miBMA* addition to the *NORM* algorithm.

1. Obtain the cross products matrix $S = X'_{obs}X_{obs}$.
2. Calculate $V = [S + diag(S)\kappa]^{-1}$, where κ is a ridge parameter.
3. Specify the imputation model $P(y_{mis}|y_{obs}, X_{obs})$ and allow maximum model uncertainty by choosing a large caliper value C for Occam's window.
4. Use $p(\beta_w|y_{obs}, X_{obs}, A) = \sum_{M_k \in A} p(\beta_w|M_k, y_{obs}, X_{obs})p(M_k|y_{obs}, X_{obs}, A)$ to calculate the averaged weighted regression coefficients, β_w, and where A was defined earlier as the subset of retained models from BMA.
5. Draw a random variate $\dot{g} \sim \chi^2_\nu$ with $\nu = n_1 - q$.
6. Calculate $\dot{\sigma}^2 = (y_{obs} - X_{obs}\hat{\beta}_w)'(y_{obs} - X_{obs}\hat{\beta}_w)/\dot{g}$.
7. Draw q independent $N(0, 1)$ variates and arrange in the vector $\dot{z}_1$.
8. Calculate $V^{1/2}$ via a Cholesky decomposition.
9. Calculate $\dot{\beta}_w = \hat{\beta}_w + \dot{\sigma}\dot{z}_1 V^{1/2}$.
10. Draw n_0 independent $N(0, 1)$ variates and arrange in the vector $\dot{z}_2$.
11. Calculate the n_0 values $\dot{y} = X_{mis}\dot{\beta}_w + \dot{z}_2\dot{\sigma}$.

Simulation study design

We investigate the impact of imputation model uncertainty under six different design conditions: (a) sample size – small (100), medium (1000) and large (5000), (b) the size of the correlations between variables – low (.2) and moderately high (.6), the percentage of missing data – low missing (20%), medium missing (40%) and high missing (60%), (d) missing mechanism – MAR and MCAR, and (e) method of imputation – *miBMA* and *NORM*. In total, 72 conditions are produced. We generate 10 variables under the assumption of multivariate normality. Missing data are generated for 3 out of 10 variables under an MAR mechanism. To generate missing under MAR, we sort variable 1 from highest to lowest. Those observations in the top, say, 20% of variable 1 are removed from the first MAR variable (variable 4). Similarly, variable 2 is sorted from highest to lowest and used to generate missingness in the second MAR variable (variable 5), and so on. Variables 7 through 10 have complete data. The process is repeated for 40% and 60% missing. We also generate missing data on 5 out of 10 variables under MCAR mechanism on the initial (full) data set by randomly removing data using the function "sample" in base program of R (R Core Team, 2017). The data are then multiply imputed separately under *miBMA* and *NORM*.

Ten iterations and 20 imputations are completed for each method. We use a large caliper value for Occam's window and let the result have up to 250 different models.[6] Each condition of the design is replicated 500 times, and thus we are exploring the frequentist properties of our BMA approach to multiple imputation (Little, 2006, 2011).

Simulation study evaluation: Kullback–Leibler divergence and mean squared prediction error

To evaluate our BMA approach for multiple imputation against the imputation model with largest posterior model probability and against normal-theory Bayesian multiple imputation, we use the *Kullback-Leibler divergence* (K-L divergence) measure (Kullback & Leibler, 1951; Kullback, 1959, 1987). The K-L divergence is related to Boltzmann's (1877) concept of entropy in physics and Shannon's (1948) notion of entropy in communication theory. Following the discussion given in Burnham and Anderson (2002), consider two probability distributions: f, which is assumed to be fixed, and g, which is assumed to vary over the space of candidate models, each defined by a parameter vector (or scalar) θ. For this study, f represents the true complete-data distribution and g represents the distribution of the data after imputation. Then, the K-L divergence between the two distributions can be written as

$$I(f, g) = \int f(x) log\left(\frac{f(x)}{g(x|\theta)}\right) dx \qquad (9)$$

[6]The default number of models in the "BMA" package is 150.

Table 1. Average over 500 replications of the cumulative Posterior Model Probabilities for miBMA.

			MAR			MCAR		
			MAR1	MAR2	MAR3	MCAR1	MCAR3	MCAR5
N = 100	Cor 0.2	20%	0.581	0.578	0.584	0.572	0.574	0.574
		40%	0.614	0.615	0.616	0.609	0.611	0.610
		60%	0.639	0.638	0.640	0.637	0.638	0.637
	Cor 0.6	20%	0.562	0.564	0.573	0.567	0.568	0.571
		40%	0.613	0.614	0.619	0.609	0.609	0.610
		60%	0.653	0.654	0.654	0.645	0.644	0.640
N = 1000	Cor 0.2	20%	0.719	0.725	0.720	0.706	0.701	0.699
		40%	0.828	0.829	0.830	0.837	0.839	0.838
		60%	0.914	0.914	0.917	0.923	0.924	0.925
	Cor 0.6	20%	0.854	0.852	0.855	0.867	0.865	0.869
		40%	0.888	0.887	0.888	0.899	0.891	0.897
		60%	0.940	0.938	0.939	0.961	0.961	0.962
N = 5000	Cor 0.2	20%	0.985	0.986	0.987	0.981	0.982	0.982
		40%	0.974	0.973	0.974	0.972	0.972	0.972
		60%	0.985	0.985	0.984	0.994	0.993	0.994
	Cor 0.6	20%	1.000	1.000	1.000	1.000	1.000	1.000
		40%	1.000	1.000	1.000	0.999	0.999	0.999
		60%	0.997	0.997	0.997	0.992	0.993	0.993

Note: miBMA = Multiple imputation under Bayesian model averaging.

where we interpret $I(f, g)$ as the "information lost when g is used to approximate f i.e., $I(f, g)$ is the distance from g to f (Burnham & Anderson, 2002, p. 51).

We argue that the K-L divergence is an appropriate measure for judging the quality of our *miBMA* approach. Specifically, it is not expected that any missing data imputation method provide accurate point estimates of the missing data, but rather that the method provides reasonable distributions of plausible missing data values. Indeed, Equation (1) shows that the underlying nature of multiple imputation is the predictive distribution. To assess the adequacy of predictive distributions, it is useful to use *scoring rules* which provide measures of the accuracy of probabilistic predictions (see e.g., Winkler et al., 1996; Jose, Nau, & Winkler, 2008; Merkle & Steyvers, 2013; Gneiting & Raftery, 2007). The K-L divergence measure is a proper scoring rule (Dawid & Musio, 2015) and is equivalent to a logarithmic scoring rule (see Theorem 4.1 in Mitra & Dunson, 2010). The model with the lowest K-L divergence measure is deemed best in the sense that the information lost when approximating the true complete-data distribution with the distribution based on the particular imputation method is lowest. To calculate K-L divergence results for both simulation and the case study, we use the "entropy" package (Hausser & Strimmer, 2014) in the R programing environment (R Core Team, 2017).

In addition to evaluating the accuracy of our BMA approach to multiple imputation using K-L divergence, we also calculate mean squared prediction error (MSPE) for this simulation study as follows.

$$MSPE = \sum_{m=1}^{20}\left[\sum_{i=1}^{10}\left(\sum_{n=1}^{N}(\dot{y}_n - y_{obs_n})^2/N\right)/10\right]/20 \quad (10)$$

where $\dot{y}_n$ are the predicted missing values, y_{obs_n} are the observed values, $n = \{1, ..., N\}$ represents number of missing cases which is a factor in the study, i indexes the number of iterations, and m indexes the number of imputations. It is well known that MSPE is a measure of both the bias and variance of an estimator for an unknown quantity. The approach with the lowest MSPE is preferred. It should be noted however, that BMA is not optimized for MSPE, but rather is a measure of model fit (Yao, Vehtari, Simpson, & Gelman, 2018).

All analyses for this simulation study are conducted within the R programing environment (R Core Team, 2017). The R code and data for this simulation study is available at http://bise.wceruw.org/index.html.

Results of simulation study

With 10 variables there are 2^{10} possible sub-models that the algorithm must explore. Table 1 provides the average posterior model probabilities over 500 replications under each condition of our study. In particular, we show the posterior model probability summed over all the sub-models. For example, for Cor 0.2/20% missing and $N = 100$ condition, we find average posterior model probabilities 0.581, 0.578 and 0.584 for the first, second and third variable with missing cases (labeled MAR1, MAR2 and MAR3, respectively in Table 1). Under MCAR, results are very similar to MAR; 0.572, 0.574 and 0.574 for MCAR1, MCAR3 and MCAR5 variables.[7] To interpret this finding, note that the posterior model probabilities do not all sum to 1.0. This suggests that there is a considerable amount of imputation model uncertainty that is not accounted for by conventional methods of multiple imputation. We note that as sample increases, model uncertainty decreases.

We also notice an interaction between the sample size conditions and the conditions of variable correlations and percent missing. In particular, for the $N = 100$ condition, the percent missing among the predictors has a greater impact on model uncertainty than the correlations, with somewhat larger posterior model probabilities when the percent missing is higher. Similar results are found for the $N = 1000$ and $N = 5000$ case.

[7]In the interest of space, we only show the results for MCAR1, MCAR3, and MCAR5. The results for MCAR2 and MCAR4 results are very similar under all conditions.

Table 2. K-L Divergence and MSPE Results under miBMA and NORM for MAR Mechanism.

				K-L Divergence			MSPE		
		Missing	Method	MAR 1	MAR 2	MAR 3	MAR 1	MAR 2	MAR 3
N = 100	Cor 0.2	20%	miBMA	0.048	0.050	0.048	2.050	2.054	2.031
			NORM	0.055	0.057	0.055	2.125	2.110	2.114
		40%	miBMA	0.078	0.083	0.082	2.102	2.116	2.091
			NORM	0.087	0.094	0.088	2.335	2.348	2.316
		60%	miBMA	0.106	0.108	0.109	2.256	2.290	2.253
			NORM	0.115	0.120	0.116	2.982	2.975	2.896
	Cor 0.6	20%	miBMA	0.056	0.058	0.058	1.085	1.076	1.068
			NORM	0.060	0.061	0.061	1.099	1.085	1.089
		40%	miBMA	0.066	0.073	0.070	1.109	1.120	1.108
			NORM	0.073	0.077	0.075	1.208	1.217	1.202
		60%	miBMA	0.080	0.084	0.085	1.203	1.215	1.196
			NORM	0.091	0.096	0.095	1.535	1.556	1.510
N = 1000	Cor 0.2	20%	miBMA	0.018	0.019	0.018	1.780	1.781	1.778
			NORM	0.018	0.020	0.019	1.758	1.760	1.760
		40%	miBMA	0.029	0.032	0.029	1.792	1.792	1.791
			NORM	0.030	0.033	0.029	1.781	1.780	1.781
		60%	miBMA	0.041	0.042	0.041	1.812	1.812	1.812
			NORM	0.041	0.044	0.042	1.823	1.819	1.824
	Cor 0.6	20%	miBMA	0.030	0.028	0.030	0.916	0.914	0.916
			NORM	0.030	0.029	0.030	0.907	0.907	0.907
		40%	miBMA	0.031	0.030	0.031	0.927	0.926	0.928
			NORM	0.031	0.030	0.032	0.921	0.920	0.921
		60%	miBMA	0.034	0.032	0.033	0.943	0.942	0.942
			NORM	0.033	0.033	0.033	0.943	0.943	0.944
N = 5000	Cor 0.2	20%	miBMA	0.013	0.015	0.016	1.733	1.740	1.734
			NORM	0.013	0.015	0.016	1.732	1.737	1.734
		40%	miBMA	0.020	0.022	0.024	1.744	1.746	1.743
			NORM	0.020	0.022	0.024	1.742	1.743	1.742
		60%	miBMA	0.027	0.030	0.031	1.754	1.757	1.756
			NORM	0.028	0.030	0.031	1.752	1.754	1.754
	Cor 0.6	20%	miBMA	0.022	0.022	0.022	0.894	0.896	0.894
			NORM	0.022	0.022	0.023	0.895	0.897	0.895
		40%	miBMA	0.023	0.023	0.023	0.899	0.900	0.899
			NORM	0.022	0.022	0.022	0.901	0.901	0.901
		60%	miBMA	0.023	0.022	0.023	0.905	0.907	0.906
			NORM	0.023	0.023	0.023	0.907	0.907	0.908

Note: miBMA = Multiple imputation under Bayesian model averaging; NORM = Multiple imputation using Bayesian linear regression; K-L Test: Kullback-Leibler divergence test; MSPE: Mean squared prediction error.

Tables 2 and 3 provide the main results of this paper – namely the K-L divergence measures and MSPE for the difference between the distributions of the complete data and the distributions of the imputed missing data under *miBMA* and *NORM* across all conditions of the study. These K-L divergences and MSPE values are averaged over 500 replications. In Table 2, for sample size $N = 100$, the results indicate a noticeable benefit of accounting for model uncertainty using *miBMA* across all conditions when missing data are MAR. We find that *miBMA* outperforms *NORM* in terms of recovering the original distributions as measured by K-L divergence and MSPE. We also find that with sample sizes $N = 1000$ and $N = 5000$, K-L divergence and MSPE values are almost identical. We also note that for sample size $N = 1000$ and Cor 0.2/20%-40% the MSPEs are slightly lower for *NORM*. The results for MCAR follows a very similar pattern. Table 3 provides the results for the MCAR mechanism.

Case study design and results

We apply *miBMA* to United States data from the 2015 cycle of PISA (OECD, 2016). Launched in 2000 by the Organization for Economic Cooperation and Development (OECD), PISA is a triennial international survey which aims to evaluate education systems worldwide by testing the skills and knowledge of 15-year-old students. In 2015, over half a million students, statistically representative of 28 million 15-year-olds in 72 countries and economies, took an internationally agreed-upon two-hour test. Students were assessed in science, mathematics, reading, collaborative problem solving and financial literacy. In addition to these so-called "cognitive outcomes", policymakers and researchers alike have begun to focus increasing attention on the nonacademic contextual aspects of schooling. Context questionnaires provide important variables for models predicting cognitive outcomes and these variables have become important outcomes in their own right - often referred to as "non-

Table 3. K-L Divergence and MSPE Results under miBMA and NORM for MCAR Mechanism.

				K-L Divergence			MSPE		
		Missing	Method	MCAR1	MCAR3	MCAR5	MCAR1	MCAR3	MCAR5
N = 100	Cor 0.2	20%	miBMA	0.044	0.044	0.042	2.003	2.013	1.956
			NORM	0.048	0.048	0.049	2.005	2.019	1.986
		40%	miBMA	0.075	0.076	0.080	2.063	2.082	2.057
			NORM	0.082	0.085	0.084	2.127	2.140	2.117
		60%	miBMA	0.104	0.106	0.110	2.244	2.255	2.226
			NORM	0.114	0.115	0.114	2.391	2.432	2.397
	Cor 0.6	20%	miBMA	0.033	0.037	0.036	1.040	1.048	1.017
			NORM	0.035	0.039	0.038	1.031	1.040	1.023
		40%	miBMA	0.063	0.065	0.064	1.079	1.089	1.075
			NORM	0.068	0.069	0.067	1.102	1.109	1.098
		60%	miBMA	0.091	0.090	0.092	1.178	1.187	1.168
			NORM	0.097	0.092	0.094	1.248	1.269	1.253
N = 1000	Cor 0.2	20%	miBMA	0.017	0.016	0.016	1.771	1.774	1.774
			NORM	0.017	0.016	0.016	1.750	1.752	1.753
		40%	miBMA	0.031	0.030	0.031	1.787	1.796	1.796
			NORM	0.032	0.031	0.032	1.771	1.774	1.775
		60%	miBMA	0.044	0.042	0.042	1.815	1.822	1.821
			NORM	0.045	0.043	0.044	1.797	1.802	1.803
	Cor 0.6	20%	miBMA	0.016	0.015	0.015	0.910	0.912	0.911
			NORM	0.015	0.015	0.015	0.901	0.902	0.902
		40%	miBMA	0.029	0.028	0.029	0.925	0.929	0.929
			NORM	0.030	0.027	0.030	0.916	0.918	0.918
		60%	miBMA	0.042	0.040	0.041	0.948	0.951	0.950
			NORM	0.042	0.040	0.042	0.935	0.938	0.939
N = 5000	Cor 0.2	20%	miBMA	0.011	0.010	0.011	1.741	1.735	1.737
			NORM	0.011	0.011	0.012	1.738	1.734	1.736
		40%	miBMA	0.021	0.020	0.021	1.753	1.752	1.751
			NORM	0.021	0.020	0.021	1.747	1.747	1.748
		60%	miBMA	0.028	0.030	0.030	1.770	1.768	1.770
			NORM	0.028	0.030	0.031	1.762	1.760	1.762
	Cor 0.6	20%	miBMA	0.011	0.010	0.011	0.895	0.892	0.893
			NORM	0.011	0.010	0.011	0.894	0.892	0.893
		40%	miBMA	0.020	0.019	0.020	0.905	0.904	0.904
			NORM	0.020	0.019	0.021	0.904	0.904	0.904
		60%	miBMA	0.026	0.030	0.030	0.920	0.919	0.920
			NORM	0.027	0.030	0.030	0.917	0.916	0.917

Note: miBMA = Multiple imputation under Bayesian model averaging; NORM = Multiple imputation using Bayesian linear regression; K-L Test: Kullback-Leibler divergence test; MSPE: Mean squared prediction error.

cognitive outcomes" (Heckman & Kautz, 2012; Kuger, Klieme, Jude, & Kaplan, 2016). PISA also assesses these non-cognitive outcomes via a one-half hour internationally agreed-upon context questionnaire (OECD, 2016).

The dependent variable for this example is the first plausible value of science achievement (PV1SCIE).[8] Nine predictor variables were selected and these variables are described in Table 4. The data contain unit missing data and the missing data patterns and Little's MCAR test results are given in Table 5. The dependent variable does not include any missing cases because plausible values are already imputed (von Davier, 2013). The total size of the sample is 5712. We find that HISEI (index of highest parental occupational status) has the highest number of missing values (7.9%) and ESCS (Index of economic, social and cultural status) has the lowest number (1.3%).

After determining the number of unique missing data patterns in the data set, we multiply impute the data using (a) *miBMA* and (b) *NORM*. Each method has 20 imputations and 10 iterations. After imputation, we apply both frequentist and Bayesian linear regression for the analyses. We pooled the results differently for the frequentist regression and Bayesian regression analyses in accordance to best practices (Zhou & Reiter, 2010). For the frequentist regression analysis, ordinary least squares estimates of the model parameters are obtained for each imputed data set and the results are pooled according to Rubin's rules (Rubin, 1987). For Bayesian linear regression, and for each imputation approach, we generate 2 chains of 50,000 samples after 5,000 burn-in by Gibbs sampling implemented via the "rjags" program (Plummer, 2016) for each imputed data set. A vague normal prior is used for the PV1SCIE and weakly-informative priors

[8]We recognize that when using plausible values in the analysis of large-scale educational assessments, it is more appropriate to use all plausible values and combine them using Rubin's (1987) rules. However, extending our miBMA approach to the plausible value framework is beyond the scope of this paper.

are used for the model parameters. This generates 2 million draws from the posterior distribution of the model parameters. These draws are combined to form the posterior distribution, for which the expected a posterior estimate, Monte Carlo standard deviations, and 95% posterior probability intervals are obtained.

All analyses for this case study were conducted within the R programing environment (R Core Team, 2017). The R code and data for this case study is available at http://bise.wceruw.org/index.html.

Table 4. Variables for Case Study.

Variable Name	Explanation
JOYSCIE	Enjoyment of science (WLE)
INSTSCI	Instrumental motivation (WLE)
SCIEEFF	Science self-efficacy (WLE)
SCIEACT	Index science activities (WLE)
HISEI	Index of highest parental occupational status
ANXTEST	Personality: Test Anxiety (WLE)
MOTIVAT	Student Atttidudes, Preferences and Self-related beliefs: Achieving motivation (WLE)
HEDRES	Home educational resources (WLE)
ESCS	Index of economic, social and cultural status (WLE)
PV1SCIE	First plausible value of the PISA 2015 science assessment

Table 5. Number and percent of missing in each variable.

Variable	Number Missing	Percent Missing
JOYSCIE	203	0.036
INSTSCI	288	0.050
SCIEEFF	324	0.057
SCIEACT	300	0.053
HISEI	453	0.079
ANXTEST	119	0.021
MOTIVAT	122	0.021
HEDRES	97	0.017
ESCS	74	0.013
PV1SCIE	0	0

Note: Number of missing data patterns = 67; Little's (1988) MCAR Test: chi-square = 1062.183, df = 421, $p<.05$.

Table 6 shows for each variable, the largest posterior model probability for each imputation. Here again, we see the extent of uncertainty in the imputation model. For example, for the first imputation, we find that the posterior model probabilities across the variables range from a low of 0.480 for ESCS to a high of 1.0 for INSTSCI.

Results for the frequentist and Bayesian regression analyses are displayed in Tables 7 and 8, respectively. We do not observe any systematic differences between frequentist and Bayesian results regardless of imputation method, due likely to the large sample size and non-informative priors used in Bayesian linear regression. However, we do observe sizable differences in the results between *miBMA* and *NORM*. Given the relative advantage of the *miBMA* approach with respect to K-L divergence and MSPE found in the simulation study, these results suggest that it is reasonable to account for imputation model uncertainty when conducting multiple imputation.

Discussion

An important issue that emerges from this work concerns the expansion of our method to handling missing data among variables that are not normally distributed or from data structures derived from clustered sampling (multilevel) designs. Indeed, an important feature of the chained equations approach, as implemented in "mice" (van Buuren & Groothuis-Oudshoorn, 2010) is the ability to choose different

Table 6. Model posterior probabilities for each imputation.

Imp.	JOYSCIE PMP	INSTSCI PMP	SCIEEFF PMP	SCIEACT PMP	HISEI PMP	ANXTEST PMP	MOTIVAT PMP	HEDRES PMP	ESCS PMP
1	0.809	1.000	0.561	0.668	0.533	0.653	0.675	0.806	0.480
2	0.701	1.000	0.680	0.502	0.761	0.181	0.413	0.852	0.430
3	0.739	0.944	0.335	1.000	1.000	0.341	0.424	0.710	0.556
4	0.686	1.000	0.540	0.784	0.674	0.443	0.257	0.397	0.808
5	0.610	0.860	0.609	0.931	0.588	0.416	0.407	0.391	0.696
6	0.656	1.000	0.747	0.880	0.584	0.699	0.328	0.788	0.370
7	0.416	1.000	0.681	0.598	0.952	0.378	0.273	0.491	0.820
8	0.877	1.000	0.513	1.000	1.000	0.289	0.837	0.740	0.444
9	0.404	0.870	0.818	1.000	1.000	0.673	0.602	0.454	0.503
10	0.557	0.845	0.516	0.718	1.000	0.591	0.292	0.490	0.304
11	0.844	0.852	0.714	0.905	0.669	0.325	0.373	0.700	0.759
12	0.404	1.000	0.767	0.849	0.752	0.619	0.415	0.220	0.708
13	0.526	0.868	1.000	0.869	0.938	0.517	0.429	0.321	1.000
14	0.473	1.000	0.637	0.695	0.456	0.250	0.471	0.693	0.577
15	0.444	0.936	0.649	0.589	0.800	0.712	0.352	0.791	0.617
16	0.710	1.000	0.942	1.000	0.513	0.282	0.324	0.342	0.735
17	0.847	0.887	0.737	0.821	0.839	0.533	0.812	0.269	0.487
18	0.334	1.000	0.931	0.865	0.528	0.342	0.732	0.760	0.805
19	0.655	0.947	0.790	0.792	0.525	0.333	0.238	0.536	0.613
20	0.520	1.000	0.523	0.590	0.684	0.465	0.490	0.554	0.909
Mean	0.611	0.950	0.685	0.803	0.740	0.452	0.457	0.565	0.631

Note: Imp = Imputation number.

Table 7. Frequentist linear regression under two missing data approaches.

	miBMA					NORM				
	Beta	SE	*t*	*df*	*p*	Beta	SE	*t*	*df*	*p*
Intercept	481.52	5.88	81.88	623.14	0.00	480.99	5.66	84.98	1406.53	0.00
JOYSCIE	24.59	1.32	18.69	2141.46	0.00	24.74	1.29	19.18	3997.79	0.00
INSTSCI	−7.55	1.43	−5.28	1248.96	0.00	−7.55	1.41	−5.36	1814.68	0.00
SCIEEFF	7.75	1.08	7.18	842.93	0.00	7.76	1.05	7.42	1890.95	0.00
SCIEACT	−7.80	1.16	−6.70	1546.12	0.00	−7.73	1.16	−6.64	1489.60	0.00
hisei	0.13	0.11	1.12	495.32	0.26	0.14	0.11	1.27	1247.66	0.20
ANXTEST	−8.84	1.25	−7.09	1474.56	0.00	−8.77	1.22	−7.21	3117.50	0.00
MOTIVAT	3.10	1.31	2.37	3296.50	0.02	3.04	1.31	2.32	3076.46	0.02
HEDRES	−0.87	1.29	−0.68	3133.75	0.50	−0.91	1.29	−0.70	2904.77	0.48
ESCS	26.43	2.66	9.92	890.29	0.00	26.21	2.60	10.06	1488.29	0.00

Table 8. Bayesian linear regression under two missing data approaches.

	miBMA				NORM			
Predictors	Beta	SD	PPI_{low}	PPI_{high}	Beta	SD	PPI_{low}	PPI_{high}
Intercept	481.48	5.80	470.03	492.77	480.96	5.64	469.91	491.98
JOYSCIE	24.56	1.30	22.00	27.11	24.73	1.29	22.21	27.25
INSTSCI	−7.51	1.42	−10.29	−4.71	−7.55	1.40	−10.29	−4.79
SCIEEFF	7.71	1.07	5.61	9.79	7.76	1.04	5.71	9.79
SCIEACT	−7.77	1.15	−10.01	−5.51	−7.73	1.16	−10.00	−5.46
HISEI	0.13	0.11	−0.09	0.34	0.14	0.11	−0.07	0.35
ANXTES	−8.84	1.24	−11.27	−6.41	−8.77	1.21	−11.15	−6.39
MOTIVAT	3.13	1.32	0.56	5.72	3.04	1.31	0.48	5.59
HEDRES	−0.86	1.28	−3.37	1.65	−0.90	1.28	−3.42	1.62
ESCS	26.41	2.64	21.21	31.55	26.19	2.59	21.10	31.26

Note: S.D. = Standard deviation, PPI = Posterior probability interval.

imputation methods for different variables. So, for example, if a variable with missing data is deemed to be normally distributed, then "mice.impute.norm" can be chosen for that variable. On the other hand, if a variable is dichotomous or polytomous, then the "mice" routines "mice.impute.logreg", or "mice.impute.polr", respectively, can be used. Or, in a more generic fashion, a routine such as predictive mean matching via "mice.impute.pmm" could be used. Finally, in the presence of multilevel data with missing data at both levels, a routine such as "mice.impute.2lnorm" in "mice" could be used. We believe that it would be relatively straightforward to extend our method to dichotomous data using the "bic.glm" function in "BMA" which provides a link function to dichotomous data (Raftery et al., 2015). However, extending our approach to polytomous regression, predictive mean matching, or two-level imputation requires additional research and development.

In addition, an important assumption underlying BMA is that the true imputation model, say, M_T is one of the models in the set of imputation models M_k, $k = 1, 2, \ldots, K$. This assumption is referred to as the *M−closed* framework, discussed in Bernardo and Smith (2000) and Clyde and Iversen (2015), and outlined in Kaplan and Lee (2018) in the context of BMA applications to education research.

Following the discussion in Bernardo and Smith (2000, pg. 385, see also Kaplan & Lee, 2018), the *M−closed* framework can be contrasted with the *M−completed* framework and the *M−open* framework. In the *M−closed* framework, it makes sense to assign prior probabilities that M_T is in the space of imputation models. In fact, this is the framework that underlies the standard approach to BMA discussed in this paper; prior probabilities are assigned to the set of imputation models (typical the indifference prior $1/M$) encoding one's belief that each imputation model is equally likely to be the true model. The application of the indifference prior is the conventional default used in this paper (Raftery et al., 2015). In the *M−completed* and *M−open* frameworks M_T is not in the set of models M_k, which are simply considered proxies to be compared. As such, the assignment of prior probabilities makes less sense and the question comes down to how imputation models are to be chosen and averaged if the true imputation model does not exist within the set of possible models.

The simulated data ~~is~~ this paper are, by definition, generated from a true model that we know, and therefore the analytic model operates in the *M−closed* framework. Moreover, in the case of missing data problems in large-scale survey operations, the notion of "congeniality" (Meng, 1994) as discussed in the introduction might serve as a warrant for operating in the *M−closed* framework. Nevertheless, the distinction among these modeling frameworks is quite important, and indeed, recent work by Clyde and Iversen (2015) have used a decision-theoretic framework that allows BMA within the *M−open* framework. These issues warrant further investigation.

Finally, we motivated our approach from the point of view of the inputer and analyst as separate individuals, where the inputer is tasked with creating a complete dataset for secondary analyses by an analyst. Of course, it is common for the inputer and analyst to be one and the same person. In such a situation,

we would hope that the analyst would shy away from questionable ad hoc approaches to handling missing data and instead implement some form of multiple imputation. We view our approach as a valuable addition to the generic notion of multiple imputation via chained-equations but we recognize that our approach might not be required for all situations, in particular when there are a very small number of variables involved in the analysis. However, our simulations do reveal an advantage to our approach even for modest size predictor sets.

To conclude, the purpose of this paper was to offer a new approach to multiple imputation that accounts for uncertainty in the imputation model via Bayesian model averaging. We argue that by addressing imputation model uncertainty directly within the chained equation approach to multiple imputation, our approach is fully Bayesianly proper in the sense of Schafer (1997). The results of our simulation study and case study demonstrate the extent of imputation model uncertainty that can occur in multiple imputation. We find that across almost all conditions of the simulation study, our *miBMA* approach confers a noticeable advantage over normal-theory based multiple imputation in terms of reproducing the complete-data distribution as measured by K-L divergence and mean squared prediction error. Thus, in general, we find that accounting for imputation model uncertainty yields superior missing data imputation performance compared to normal-theory based multiple imputation which ignores model uncertainty. Our simulation results and the findings from our case study support the argument that adding BMA to multiple imputation is a prudent approach to handling missing data in practice.

Article information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was not supported.

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: This paper is based, in part, on the first author's 2018 presidential address to the Society for Multivariate Experimental Psychology. The authors would like to thank Associate Editor Sarah Depaoli and two anonymous reviewers for their comments on prior versions of this manuscript.

References

Bernardo, J., & Smith, A. F. M. (2000). *Bayesian Theory*. New York: Wiley.

Boltzmann, L. (1877). Über die Beziehung zwichen dem Hauptsatze derzwe: Ten mechanischen Wärmtheorie und der Wahrscheinlichkeitreschung respective den Sätzen über das Wärmegleichgewicht. *Wiener Berichte*, *76*, 373–435.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer.

Clyde, M. A. (1999). Bayesian model averaging and model search strategies. In *Bayesian statistics 6* (pp. 157–185). Oxford: Oxford University Press.

Clyde, M. A., & George, E. I. (2004). Model uncertainty. *Statistical Science*, *19*, 81–94. doi:10.1214/088342304000000035

Clyde, M. A., & Iversen, E. S. (2015). Bayesian model averaging in the M-open framework. In *Bayesian theory and applications* (pp. 483–498). Oxford: Oxford University Press.

Dawid, A. P., & Musio, M. (2015). Bayesian model selection based on proper scoring rules. *Bayesian Analysis*, *10*(2), 479–499. doi:10.1214/15-BA942

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion.). *Journal of the Royal Statistical Society (Series B)*, *57*, 55–98.

Enders, C. K. (2010). *Applied missing data analysis*. New York: The Guilford Press.

Fernández, C., Ley, E., & Steel, M. F. J. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, *100*(2), 381–427. doi:10.1016/S0304-4076(00)00076-2

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881–889. doi:10.2307/2290777

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. doi:10.1198/016214506000001437

Hausser, J., & Strimmer, K. (2014). entropy: Estimation of entropy, mutual information and related quantities [Computer software manual]. Retrieved from https://

CRAN.R-project.org/package=entropy (R package version 1.2.1)

Heckman, J. J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics*, *19*(4), 451–464. doi:10.1016/j.labeco.2012.05.014

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*, 382–417. doi:10.1214/ss/1009212814

Jose, V. R. R., Nau, R. F., & Winkler, R. L. (2008). Scoring rules, generalized entropy, and utility maximization. *Operations Research*, *56*(5), 1146–1157. doi:10.1287/opre.1070.0498

Kaplan, D., & Chen, J. (2014). Bayesian model averaging for propensity score analysis. *Multivariate Behavioral Research*, *49*(6), 505–517. doi:10.1080/00273171.2014.928492

Kaplan, D., & Lee, C. (2015). Bayesian model averaging over directed acyclic graphs with implications for the predictive performance of structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(3), 343–353. doi:10.1080/10705511.2015.1092088

Kaplan, D., & Lee, C. (2018). Optimizing prediction using Bayesian model averaging: Examples using large-scale educational assessments. *Evaluation Review*, *42*(4), 423. doi:10.1177/0193841X187611

Kaplan, D., & Su, D. (2018). On imputation for planned missing data in context questionnaires using plausible values: A comparison of three designs. *Large-Scale Assessments in Education*, *6*(1), 1–31. Retrieved from doi:10.1186/s40536-018-0059-9

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. doi:10.2307/2291091

Kass, R. E., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, *90*(431), 928–934. doi:10.2307/2291327

Kuger, S., Klieme, E., Jude, N., & Kaplan, D. (2016). *Assessing contexts of learning world-wide – Extended context assessment frameworks*. Dordrecht: Springer.

Kullback, S. (1959). *Information theory and statistics*. New York: John Wiley and Sons.

Kullback, S. (1987). The Kullback-Leibler distance. *The American Statistician*, *41*, 340–341.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. doi:10.1214/aoms/1177729694

Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data*. New York: Wiley.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*(404), 1198–1202. doi:10.1080/01621459.1988.10478722

Little, R. J. A. (2006). Calibrated Bayes: A Bayes/frequentist roadmap. *The American Statistician*, *60*(3), 213–223. doi:10.1198/000313006X117837

Little, R. J. A. (2011). Calibrated Bayes, for statistics in general, and missing data in particular. *Statistical Science*, *26*(2), 162–174. doi:10.1214/10-STS318

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: John Wiley & Sons.

Liu, J., Gelman, A., Hill, J., Su, Y.-S., & Kropko, J. (2014). On the stationary distribution of iterative imputations. *Biometrika*, *101*(1), 155–173. doi:10.1093/biomet/ast044

Madigan, D., & Raftery, A. E. (1994). Model selection and accounting for model uncertainly in graphical models using Occam's window. *Journal of the American Statistical Association*, *89*(428), 1535–1546. doi:10.2307/2291017

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, *9*(4), 538–558 doi:10.1214/ss/1177010269

Merkle, E. C., & Steyvers, M. (2013). Choosing a strictly proper scoring rule. *Decision Analysis*, *10*(4), 292–304. doi:10.1287/deca.2013.0280

Mitra, R., & Dunson, D. (2010). Two-level stochastic search variable selection in GLMS with missing predictors. *The International Journal of Biometrics*, *6*, 1–41. Retrieved from doi:10.2202/1557-4679.1173

Murray, J. S. (2018). Multiple imputation: A review of practical and theoretical findings. *Statistical Science*, *33*(2), 142. doi:10.1214/18-STS644

OECD. (2016). *PISA 2015 Results (Volume* I*)*. Retrieved from https://www.oecd-ilibrary.org/content/publication/9789264266490-en

Plummer, M. (2016). RJAGS: Bayesian graphical models using MCMC [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=rjags (R package version 4-6)

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika*, *83*(2), 251–266. doi:10.1093/biomet/83.2.251

Raftery, A. E. (1998). *Bayes factors and the BIC: Comment on Weakliem (Tech. Rep. No. 347)*. Seattle, WA: University of Washington, Department of Statistics.

Raftery, A. E., Hoeting, J., Volinsky, C., Painter, I., & Yeung, K. Y. (2015, June 22). *Bayesian Model Averaging (BMA), Version 3.12*. http://www2.research.att.com/volinsky/bma.html.

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, *92*(437), 179–191. doi:10.2307/2291462

Rubin, D. B. (1987). *Multiple imputation in nonresponse surveys*. Hoboken, NJ: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*(434), 473–489. doi:10.1080/01621459.1996.10476908

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York: Chapman & Hall/CRC.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x

Sloughter, J. M., Gneiting, T., & Raftery, A. E. (2013). Probabilistic wind vector forecasting using ensembles and Bayesian model averaging. *Monthly Weather Review*, *141*, 2107–2119. doi:10.1175/MWR-D-12-00002.1

Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, *81*(393), 82–86. doi:10.2307/2287970

van Buuren, S., & Groothuis-Oudshoorn, K. (2010, January). Multivariate Imputation by Chained Equations, Version 2.3. http://www.multiple-imputation.com/.

van Buuren, S. (2012). *Flexible imputation of missing data*. New York: Chapman & Hall.

von Davier, M. (2013). Imputing proficiency data under planned missingness in population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. Boca Raton, IL: Chapman Hall/CRC.

Winkler, R. L., Muñoz, J., Cervera, J. L., Bernardo, J. M., Blattenberger, G., Kadane, J. B., … Ríos-Insua, D. (1996). Scoring rules and the evaluation of probabilities. *Test*, *5*(1), 1–60. doi:10.1007/BF02562681

Yao, Y., Vehtari, A., Simpson, D. & Gelman, A. (2018). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, *13*, 917–1007.

Yeung, K. Y., Bumgarner, R. E., & Raftery, A. E. (2005). Bayesian model averaging: Development of an improved multi-class, gene selection, and classification tool for microarray data. *Bioinformatics*, *21*(10), 2394–2402. doi:10.1093/bioinformatics/bti319

Zellner, A., (1986). On assessing prior distributions and Bayesian regression analysis with g prior distributions. In P. Goel, & A. Zellner (Eds.), *Bayesian inference and decision techniques: Essays in honor of Bruno de Finetti. Studies in Bayesian Econometrics* (pp. 233–243). New York: Elsevier.

Zeugner, S., & Feldkircher, M. (2015). Bayesian model averaging employing fixed and flexible priors: The BMS package for R. *Journal of Statistical Software*, *68*(4), 1–37. doi:10.18637/jss.v068.i04

Zhou, X., & Reiter, J. P. (2010). A note on Bayesian inference after multiple imputation. *The American Statistician*, *64*(2), 159–163. doi:10.1198/tast.2010.09109

Zigler, C. M., & Dominici, F. (2014). Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects. *Journal of the American Statistical Association*, *109*(505), 95–107. doi:10.1080/01621459.2013.869498