Multivariate Behavioral Research, 49:505–517, 2014

ISSN: 0027-3171 print / 1532-7906 online
DOI: 10.1080/00273171.2014.928492

Routledge
Taylor & Francis Group

Bayesian Model Averaging for Propensity Score Analysis

David Kaplan and Jianshen Chen
Department of Educational Psychology, University of Wisconsin–Madison

This article considers Bayesian model averaging as a means of addressing uncertainty in the selection of variables in the propensity score equation. We investigate an approximate Bayesian model averaging approach based on the model-averaged propensity score estimates produced by the R package *BMA* but that ignores uncertainty in the propensity score. We also provide a fully Bayesian model averaging approach via Markov chain Monte Carlo sampling (MCMC) to account for uncertainty in both parameters and models. A detailed study of our approach examines the differences in the causal estimate when incorporating noninformative versus informative priors in the model averaging stage. We examine these approaches under common methods of propensity score implementation. In addition, we evaluate the impact of changing the size of Occam's window used to narrow down the range of possible models. We also assess the predictive performance of both Bayesian model averaging propensity score approaches and compare it with the case without Bayesian model averaging. Overall, results show that both Bayesian model averaging propensity score approaches recover the treatment effect estimates well and generally provide larger uncertainty estimates, as expected. Both Bayesian model averaging approaches offer slightly better prediction of the propensity score compared with the Bayesian approach with a single propensity score equation. Covariate balance checks for the case study show that both Bayesian model averaging approaches offer good balance. The fully Bayesian model averaging approach also provides posterior probability intervals of the balance indices.

The distinctive feature that separates Bayesian statistical inference from its frequentist counterpart is its focus on describing and modeling all forms of uncertainty. The primary focus of uncertainty within Bayesian inference concerns prior knowledge about model parameters. In the Bayesian framework, all unknown parameters are assumed to be random, described by probability distributions. Bayesian inference encodes background knowledge about the unknown parameters in the form of the prior distribution (Gelman, Carlin, Stern & Rubin, 2003).

Within the Bayesian framework, parameters are not the only unknown elements. In fact, the Bayesian framework recognizes that models themselves possess uncertainty insofar as a particular model is typically chosen based on prior knowledge of the problem at hand and the variables that have been used in previously specified models. This form of uncertainty often goes unnoticed. Quoting Hoeting, Madigan, Raftery and Volinsky (1999),

> Standard statistical practice ignores model uncertainty. Data analysts typically select a model from some class of models and then proceed as if the selected model had generated the data. This approach ignores the uncertainty in model selection, leading to over-confident inferences and decisions that are more risky than one thinks they are. (p. 382)

An internally consistent Bayesian framework for model building and estimation must also account for model uncertainty. The current approach to addressing the problem of uncertainty lies in the method of *Bayesian model averaging* (BMA).

In this article, we investigate the use of Bayesian model averaging in propensity score analysis (Rosenbaum & Rubin, 1983) for quasi-experimental or observational studies. The organization of this article is as follows. In the next section we briefly review the two-step Bayesian propensity score approach proposed by Kaplan and Chen (2012). Next, we

Correspondence concerning this article should be addressed to David Kaplan, Department of Educational Psychology, University of Wisconsin–Madison, Madison, WI 53706. E-mail: dkaplan@education.wisc.edu

outline the method of Bayesian model averaging with an additional discussion of Occam's window (Madigan & Raftery, 1994)—a method used to reduce the overall size of the model space. This is followed by the proposed approaches to Bayesian model averaging within the propensity score framework. Here we outline two approaches to Bayesian model averaging using existing packages in R (R Development Core Team, 2011). The first approach adopts the model-averaged propensity score estimates provided by the R package *BMA* (Raftery, Hoeting, Volinsky, Painter & Yeung, 2009). The second approach approach employs a Markov chain Monte Carlo (MCMC) procedure to obtain the posterior propensity scores in the selected models via the R packages *BMA* and *MCMCpack* (Martin, Quinn & Park, 2010). The former approach is approximately Bayesian because the model selection criterion is Bayesian but the posterior mean estimates in the selected models are approximated by maximum likelihood estimates, whereas the latter approach is fully Bayesian insofar as it utilizes Bayesian model selection criteria but also simulates the posterior distribution of the propensity score using MCMC, allowing for the incorporation of priors into the selected propensity score models. After the elaboration of the methods, we outline the design of two simulation studies and their results. This is followed by the design and results of a real-data study case. The article concludes with a discussion.

BAYESIAN APPROACHES TO PROPENSITY SCORE ANALYSIS

In their seminal paper, Rosenbaum and Rubin (1983) proposed propensity score analysis as a practical tool for reducing selection bias through balancing treatment and control groups on measured covariates. Since then, a variety of propensity score techniques have been developed for both the estimation and the application of the propensity score. Models for estimating the propensity score equation have included parametric logit regression with chosen interaction and polynomial terms (e.g., Dehejia & Wahba, 1999; Hirano & Imbens, 2001), and generalized boosting modeling (McCaffrey, Ridgeway & Morral, 2004). Methods for estimating the treatment effect while accounting for the propensity score include stratification, weighting, matching, and regression adjustment.

Rubin (1985) argued that a Bayesian approach to propensity score analysis should be of great interest to the applied Bayesian analyst, and yet propensity score estimation within the Bayesian framework was not addressed until relatively recently. Hoshino (2008) developed a quasi-Bayesian estimation method for general parametric models, such as latent variable models, and developed a MCMC algorithm to estimate the propensity score. McCandless, Gustafson and Austin (2009) provided a practical Bayesian approach to propensity score stratification, estimating the propensity score and the treatment effect and sampling from the joint posterior distribution of model parameters via an MCMC algorithm. The marginal posterior probability of the treatment effect can then be obtained based on the joint posterior distribution. Similar to the McCandless et al. (2009) study, An (2010) presented a Bayesian approach that jointly models both the propensity score equation and outcome equation at the same time and extended this one-step Bayesian approach to propensity score regression and single nearest neighbor matching methods.

A consequence of the Bayesian joint modeling procedure utilized by McCandless et al. (2009) and An (2010) is that the posterior distribution of the propensity score may be affected by the outcome variable that is observed after treatment assignment, resulting in biased propensity score estimation. This is especially problematic if the relationship between the outcome and the propensity score is misspecified (McCandless, Douglas, Evans & Smeeth, 2010). To solve this problem, McCandless et al. (2010) utilized an approximate Bayesian technique introduced by Lunn, Best, Spiegelhalter, Graham and Neuenschwander (2009) for preventing undesirable feedback between the propensity score model and outcome model components. Specifically, McCandless et al. (2010) included the posterior distribution of the propensity score parameters as covariate input in the outcome model so that the flow of information between the propensity score and the outcome is restricted. This so-called *sequential Bayesian propensity score analysis* yields treatment effect estimates that are comparable to estimates obtained from frequentist propensity score analysis. Nevertheless, as McCandless et al. (2010) pointed out, their method is only approximately Bayesian and also encounters the difficulty that the Markov chain is not guaranteed to converge.

In order to maintain a fully Bayesian framework while overcoming the conceptual and practical difficulties of the joint modeling methods of McCandless et al. (2009) and An (2010), a two-step Bayesian propensity score approach (BPSA) was recently developed by Kaplan and Chen (2012) that can incorporate prior information on the model parameters of both the propensity score equation and outcome model equation. Consistent with Bayesian theory (see, e.g., Finetti, 1974), specifying prior distributions on the model parameters is a natural way to quantify uncertainty— here in both the propensity score and outcome equations. We develop the Bayesian model averaging approaches based on the Kaplan & Chen's approach to Bayesian propensity score analysis and discuss its method and properties in the next section.

A TWO-STEP BAYESIAN PROPENSITY SCORE ANALYSIS

As noted earlier, a recent paper by Kaplan and Chen (2012) advanced a two-step approach to Bayesian propensity score analysis that was found to quite accurately estimate the treatment effect while at the same time preventing undesirable feedback between the propensity score model and the outcome model. We apply Bayesian model averaging to

the Kaplan & Chen model and describe that model in this section.

Specification of the Two-Step Bayesian PSA Model

In the Kaplan and Chen (2012) two-step Bayesian propensity score approach (BPSA), the propensity score model specified was the following logit model.

$$Log\left(\frac{e(x)}{1-e(x)}\right) = \alpha + \beta x, \tag{1}$$

where α is the intercept, β refers to a vector of slopes, and x represents a set of chosen covariates. For this step, Kaplan & Chen use the R package *MCMClogit* (Martin et al., 2010) to sample from the posterior distributions of α and β using a random walk Metropolis algorithm. After the posterior propensity scores are obtained, a Bayesian outcome model is fit in the second step to estimate the treatment effect via various propensity score methods such as stratification, weighting, and optimal full matching.

To illustrate the Kaplan and Chen (2012) approach, consider a posterior sampling procedure of a chosen Bayesian logit model with $1,000$ iterations and a thinning interval of 1. Then for each observation, there will be $m = 1,000$ posterior propensity scores $e(x)$ calculated using propensity score model parameters α and β as follows:

$$e(x) = \frac{exp(\alpha + \beta x)}{1 + exp(\alpha + \beta x)}. \tag{2}$$

Based on each posterior propensity score, there will be $J = 1,000$ posterior draws of the treatment effect γ generated from its posterior distribution. Assuming that y is the outcome and T is the treatment indicator, Kaplan and Chen (2012) then provide the following treatment effect estimator,

$$E(\gamma \mid x, y, T) = m^{-1}J^{-1}\sum_{i=1}^{m}\sum_{j=1}^{J}\gamma_j(\eta_i), \tag{3}$$

where $J^{-1}\sum_{j=1}^{J}\gamma_j(\eta_i)$ is the posterior sample mean of γ in the Bayesian outcome model based on the i^{th} set of propensity scores $\eta_i, i = 1, \ldots, m$ and $j = 1, \ldots, J$. This posterior sample mean is then averaged over m sets of posterior propensity scores. The posterior variance of γ is then based on the total variance formula,

$$Var(\gamma \mid x, y, T) = m^{-1}\sum_{i=1}^{m}\sigma^2_{\gamma(\eta_i)} + (m-1)^{-1} \sum_{i=1}^{m}\left\{\mu_{\gamma(\eta_i)} - m^{-1}\sum_{i=1}^{m}\mu_{\gamma(\eta_i)}\right\}^2, \tag{4}$$

where

$$\sigma^2_{\gamma(\eta_i)} = (J-1)^{-1}\sum_{j=1}^{J}\left\{\gamma_j(\eta_i) - J^{-1}\sum_{j=1}^{J}\gamma_j(\eta_i)\right\}^2 \tag{5}$$

is the posterior sample variance of γ in the Bayesian outcome model under the i^{th} set of propensity scores and

$$\mu_{\gamma(\eta_i)} = J^{-1}\sum_{j=1}^{J}\gamma_j(\eta_i), \tag{6}$$

is the posterior sample mean of γ in the same Bayesian outcome model. Notice that two sources of variation are present in Equation (4). The first source of variation is the average of the posterior variances of γ across the posterior samples of propensity scores, represented by the first part of the right hand side of Equation (4), and the second source of variation comes from the variance of the posterior means of γ obtained across the posterior samples of propensity scores, estimated by the second part of the right-hand side of Equation (4) (Kaplan & Chen, 2012).

A graphical display of the Kaplan and Chen (2012) two-step Bayesian propensity score model is shown in Figure 1. On the left-hand side, Step 1 shows a Bayesian logistic regression with model parameters α and β as in Equation (2). For each of the 1,000 draws from the posterior distribution of the propensity score model parameters, 1,000 draws are then obtained from the posterior distribution of treatment effect γ.

Kaplan and Chen (2012) conducted three simulation studies as well as a small case study comparing frequentist propensity score analysis with the two-step Bayesian alternative focusing on the estimated treatment effect and variance estimates. The effects of different sample sizes, true treatment effects, and choice of priors on the treatment effect and variance estimates were also evaluated. Consistent with Bayesian theory, Kaplan and Chen's findings showed that lower prior precision of the treatment effect is desirable when no prior information is available in order to obtain estimates similar to frequentist results but wider intervals that account for propensity score uncertainty; or, higher prior precision is preferable when accurate prior information regarding treatment effects is attainable in order to obtain more accurate and precise treatment effect estimates. For the case of small sample size, the Bayesian approach shows slight superiority in the estimation of the treatment effect compared with the frequentist counterpart.

A further study of the covariate balance properties of the Kaplan and Chen (2012) approach was given in a case study by Chen and Kaplan (2014). Their results revealed that both Bayesian and frequentist propensity score approaches substantially reduced initial imbalance as expected, and their performance on covariate balance was similar with regard to the standardized mean/proportion differences and variance ratios in the treatment group and control group. Similar

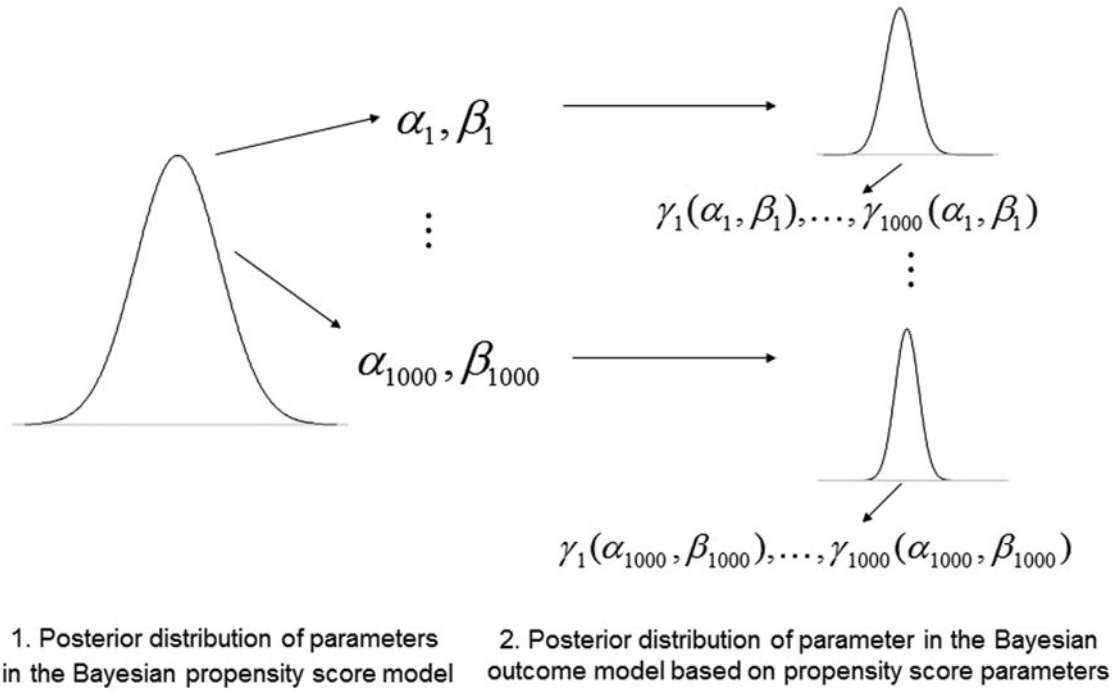

FIGURE 1 Graphical display of two-step Bayesian propensity score analysis.

performance was also found with respect to the 95% bootstrap intervals and posterior probability intervals. That is, although the frequentist propensity score approach provided slightly better covariate balance for the propensity score stratification and weighting methods, the two-step Bayesian approach offered slightly better covariate balance under optimal full matching method. Results of the Chen and Kaplan simulation study indicated similar findings. In addition, the Bayesian propensity score approach with informative priors showed equivalent balance performance compared with the Bayesian approach with noninformative priors, indicating that the specification of the prior distribution did not greatly influence the balance properties of the two-step Bayesian approach. The optimal full matching method, on average, offered the best covariate balance compared with stratification and weighting methods for both Bayesian and frequentist propensity score approaches. Chen and Kaplan also found that the two-step Bayesian approach under optimal full matching with either noninformative or informative priors provided, on average, the smallest standardized mean/proportion difference and variance ratio of the covariates between the treatment and control groups.

Chen and Kaplan (2014) argued that a benefit of conducting Bayesian propensity score analysis is that one can obtain the posterior distribution of the propensity score and thus the posterior distribution of corresponding balance indices (e.g., Cohen's *d* and variance ratio) so that the variation in balance indices can be studied in addition to the point estimates to assist in balance checking. Good balance is achieved if both the point estimates and the posterior probability intervals of the balance indices fall into the desirable range.

The Bayesian propensity score approaches described in the preceding paragraphs all assume that the propensity score model itself is, in some sense, fixed. Returning to the quote by Hoeting, Madigan, Raftery and Volinsky (1999), we argue that it is incorrect to treat the propensity score equation as fixed. Rather, as a model for treatment selection, it is reasonable to assume that many possible models could have been chosen. Thus, we argue that a full accounting of uncertainty in propensity score analysis should also address model uncertainty, and thus the purpose of this article is to explore Bayesian model averaging in the propensity score context.

In their paper, Kaplan and Chen (2012) noted that their approach addressed uncertainty in the parameters of the propensity score equation and outcome equation in a sensible fashion. However, they also noted that their approach did not address uncertainty in the choice of covariates to be used in the propensity score equation. The problem of covariate choice has been discussed in Steiner, Cook and Shadish (2011) and Steiner, Cook, Shadish and Clark (2010), who demonstrate important strategies for covariate selection that directly concern the assumption of strong ignorability of treatment assignment. However, uncertainty with regard to strong ignorability is captured in the disturbance term of the propensity score equation. As stated in the introduction, the purpose of this article is to provide two Bayesian approaches to addressing uncertainty in the choice of a propensity score model via Bayesian model averaging.

BAYESIAN MODEL AVERAGING

To begin, consider a quantity of interest such as a future observation or a parameter. Following the notation given in Madigan and Raftery (1994, see also Hoeting et al., 1999), we denote this quantity as Δ. Next, consider a set of competing models M_k, $k = 1, 2, \ldots, K$, that are not necessarily nested. For this study, M_k could be k possible propensity score models. The posterior distribution of Δ given data y can be written as

$$p(\Delta|y) = \sum_{k=1}^{K} p(\Delta|M_k, y)p(M_k|y), \tag{7}$$

where $p(\Delta|M_k, y)$ is the posterior distribution of the quantity of interest given model M_k and $p(M_k|y)$ is the posterior probability of model M_k written as

$$p(M_k|y) = \frac{p(y|M_k)p(M_k)}{\sum_{l=1}^{K} p(y|M_l)p(M_l)}, \qquad l = 1, 2, \ldots, K. \tag{8}$$

An important feature of Equation (8) is that $p(M_k)$ will likely be different for different models. The term $p(y|M_k)$ can be expressed as an integrated likelihood

$$p(y|M_k) = \int p(y|\theta_k, M_k)p(\theta_k|M_k)d\theta_k, \tag{9}$$

where $p(y|\theta_k, M_k)$ is the likelihood under model M_k and $p(\theta_k|M_k)$ is the prior density of θ_k under model M_k (Madigan & Raftery, 1994).

It can be difficult to obtain an analytic solution to the integrated likelihood in Equation (9) because it is often of very high dimension. To resolve this problem one can use an approximation to the Bayesian information criterion (BIC), written as

$$2\log p(y|M_k) \approx 2\log p(y|\hat{\theta}_k) - d_k \log(n) = -\mathrm{BIC}_k, \tag{10}$$

where d_k is the number of independent parameters in M_k and $\hat{\theta}_k$ is the maximum likelihood estimator for the parameters in M_k. Thus, BMA provides an approach for combining models specified by researchers. The advantage of BMA has been discussed in Madigan and Raftery (1994) who showed that BMA provides better predictive performance than that of a single model. Given that the propensity score is the predicted probability of treatment assignment given a set of covariates, we hypothesize that BMA should provide estimated propensity scores at least as good as that obtained from assuming a fixed single propensity score equation. We evaluate the quality of the BMA approach to estimating the propensity score by comparing treatment effects across methods and examine covariate balance in a case study.

Occam's Window

As pointed out by Hoeting et al. (1999), BMA is difficult to implement. In particular, they noted that the number of terms in Equation (7) can be quite large, the corresponding integrals are hard to compute (though possibly less so with the advent of MCMC), the specification of $p(M_k)$ may not be straightforward, and choosing the class of models to average over is also challenging. The problem of reducing the overall number of models that one could incorporate in the summation of Equation (7) has led to a solution based on the notion of *Occam's window* (Madigan & Raftery, 1994).

To motivate the idea behind Occam's window, consider the problem of finding the best subset of predictors in a linear regression model. Following closely the discussion given in Madigan and Raftery (1994), we consider an initially large number of predictors, but perhaps the goal is to find a subset that provides accurate predictions.[1] As noted in the earlier quote by Hoeting et al. (1999), the concern in drawing inferences from a single "best" model is that the choice of a single set of predictors ignores uncertainty in model selection. Occam's window (Madigan & Raftery, 1994) provides such an approach for Bayesian model averaging by reducing the subset of models under consideration.

The Occam's window algorithm proceeds in two steps (Madigan & Raftery, 1994). In the first step, models are eliminated if they predict the data less well than the model that provides the best predictions. Formally, consider a set of models M_k, $k = 1 \ldots K$, and a cutoff value C chosen in advance by the analyst. Then, we can create the set $\mathcal{A}'$ such that

$$\mathcal{A}' = \left\{ M_k : \frac{\max_l\{p(M_l|y)\}}{p(M_k|y)} \leq C \right\}. \tag{11}$$

We see that Equation (11) compares the 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 (11) is less than or equal to a chosen value C, then it is to be included in the model averaging.

In the second step, models are discarded from consideration if they receive less support from the data than simpler submodels. Formally, we consider a set $\mathcal{B}$, where

$$\mathcal{B} = \left\{ M_k : \exists M_l \in \mathcal{A}', M_l \subset M_k, \frac{p(M_l|y)}{p(M_k|y)} > 1 \right\}. \tag{12}$$

Equation (12) states that there exists a model M_l within the set $\mathcal{A}'$ and where M_l is simpler than M_k. If the simpler model receives more support from the data than the more complex model, then it is included in the set $\mathcal{B}$. Notice that the second step corresponds to the principle of Occam's razor (Madigan & Raftery, 1994).

[1] In the frequentist framework the notion of "best subset regression" is controversial because of concern over capitalization on chance. However, in the Bayesian framework with its focus on predictive accuracy, finding the best subset of predictors is less of a problem.

With Step 1 and Step 2, the problem of Bayesian model averaging is simplified by replacing equation (7) with

$$p(\Delta|y, \mathcal{A}) = \sum_{M_k \in \mathcal{A}} p(\Delta|M_k, y)p(M_k|y, \mathcal{A}), \qquad (13)$$

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

Madigan and Raftery (1994) then outlined the approach to choosing between two models to be considered for BMA. Specifically, now consider just two models M_1 and M_0, where M_0 is the smaller of the two models. This could be the case where M_0 contains fewer predictors than M_1 in a regression analysis. 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 odd lies in between the preset criterion, then both models are retained.

METHODS

An excellent and freely available open source program for Bayesian model averaging is the R package *BMA* (Raftery et al., 2009). We apply the Bayesian model averaging method utilized in the *BMA* program (Raftery et al., 2009) to obtain the model-averaged propensity score estimate, which uses maximum likelihood estimates to approximate posterior means of the selected models and then average them with posterior model probabilities as weights. This approach is referred to as *BMA-Approx* in this article. However, the *BMA* program cannot directly provide posterior samples of the propensity score, and so the uncertainty of propensity score cannot be incorporated into the outcome equation. Therefore, in this article, we propose a fully Bayesian MCMC methodology of Bayesian model averaging for propensity score analysis, accounting for both model uncertainty and parameter uncertainty in the propensity score model. For these two BMA approaches, both frequentist and Bayesian outcome models are examined. In addition, the two-step Bayesian propensity score approach developed by Kaplan and Chen (2012) is evaluated and compared with the BMA-based propensity score approaches. The R programs *MCMC logit* and *MCMCregress* within *MCMCpack* (Martin et al., 2010) are implemented to obtain posterior propensity score and treatment effect, respectively. Methods of implementation include propensity score stratification (e.g., Rosenbaum & Rubin, 1983, 1984), weighting (e.g., Hirano & Imbens, 2001; Lunceford & Davidian, 2004), optimal full matching (e.g., Hansen, 2004; Rosenbaum, 1989), and regression adjustment (e.g., Rubin, 1979; Schafer & Kang, 2008).

Bayesian Model Averaging (BMA) Approaches to Propensity Score Analysis

For the BMA-Approx propensity score approach, we first estimate the propensity score using the "bic.glm" routine in the *BMA* package. Specifically, a weighted average of the approximate posterior mean estimates, with the posterior model probabilities as weights, is used to create the model-averaged propensity score estimates. Then, based on the propensity score estimates, the treatment effects are estimated in the outcome model via propensity score stratification, weighting, optimal full matching, and regression adjustment methods. The benefit of this BMA approach is that it accounts for the uncertainty in the selection of propensity score models and also the program runs very fast. However, the approximate BMA approach is limited in that each unit of analysis has only one estimated propensity score and thus the uncertainty of the propensity score itself is ignored in the treatment effect estimation. The posterior distribution of the propensity score cannot be directly obtained and has to be normally approximated using the posterior mean and standard deviation estimates provided by the *BMA* package.

To address the problem associated with ignoring uncertainty in the propensity score, we propose a fully MCMC Bayesian model averaging procedure within the propensity score framework. We refer to this approach as *BMA-MCMC* later. For our approach, we use the R software packages *BMA* (Raftery et al., 2009) and *MCMCpack* (Martin et al., 2010). We have found these programs the most flexible for our needs, but it is, of course, possible that our approach could be implemented using other existing programs, or through original R programming.

The detailed steps of our approach are as follows:

Step 1: Select propensity score models (covariates) using the R program *BMA*. In some cases where a large quantity of models with small posterior model probabilities are selected, certain cumulative posterior model probabilities (e.g., top 50%, 70%, and 90%) can be used to limit the selected models to the most crucial ones.

Step 2: Use the Bayesian logistic regression program *MCMClogit* within *MCMCpack* (Martin et al., 2010) to obtain the posterior distribution of the propensity score for each selected model. If prior research or expert opinion can be brought to bear on the process of selection into treatment, then priors in the Bayesian logit model can be specified in this step.

Step 3: Sample from the posterior distribution of the propensity score in each model with the model's posterior probability as weight to get the final posterior distribution of the propensity score.

Step 4: Based on each posterior draw of the propensity score, a treatment effect estimate and a variance estimate can be

produced through the outcome model via stratification, weighting, optimal full matching, or regression adjustment. In this study, there are 1,000 posterior draws of the propensity score and thus 1,000 treatment effect and variance estimates are obtained.

Step 5: The final treatment effect estimate and variance estimate can be calculated by Equation (3) and Equation (4), respectively, as proposed by Kaplan and Chen (2012).

DESIGN OF SIMULATION STUDIES

We conduct two simulation studies to assess the performance of the two Bayesian model averaging approaches to covariate choice in propensity score analysis. In both simulation studies, data are generated with one continuous outcome, one binary treatment, and 10 covariates with different distributions. The details of our procedure are as follows:

1. Independently generate random variables $x_1, x_2, \ldots, x_{10}$ as 10 covariates under sample size $n = 200$, such as

$$\begin{aligned} x_1 &\sim N(0,1) & x_6 &\sim Bernoulli(0.3) \\ x_2 &\sim Poisson(2) & x_7 &\sim N(-1,3) \\ x_3 &\sim Bernoulli(0.5) & x_8 &\sim N(2,2) \\ x_4 &\sim N(0,2) & x_9 &\sim N(1,0.8) \\ x_5 &\sim Bernoulli(0.6) & x_{10} &\sim N(2,1). \end{aligned}$$

   These distributions are chosen to imitate different types of covariates found in practice such as continuous variables, count data, and dichotomous (e.g., yes/no) variables.
2. Let $x = (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10})'$ be the vector of covariates. Obtain the true propensity scores by the model

$$e_i(x) = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}, \tag{14}$$

   where the generating coefficients are $\alpha = 0$ and $\beta = (0.6, 0.1, -0.3, -0.4, 0.2, -0.3, -0.2, 0.2, 0.5, 0.3)$.
3. Calculate the treatment assignment vector T by comparing the propensity score $e_i(x)$ with a random variable U_i generated from the $Uniform\,(0,1)$ distribution, where $i = 1, \ldots, n$. Assign $T_i = 1$ if $U_i \leq e_i(x)$, $T_i = 0$ otherwise.
4. Generate outcomes $y_1, \ldots, y_n$ using the model

$$y_i = \lambda x_i + \gamma T_i + \epsilon_i, \tag{15}$$

   where $\lambda = (0.2, 0.1, 0.2, -0.1, -0.2, 0.2, -0.2, 0.1, 0.2, 0.1)$, $\epsilon \sim N(0,1)$ and γ is the true treatment effect taking the value of 5.
5. Data $= \{(y_i, x_i, T_i), i = 1, \ldots, n; n = 200\}$.
6. Replicate the aforementioned steps 200 times.

Note that Step 6 provides an evaluation of the frequentist properties of the proposed BMA propensity score methods. The issue of providing a frequentist evaluation of Bayesian methods in general is based on original ideas of Box (1980) and Rubin (1984) and was developed by Little (2006, 2011, 2013), who referred to the idea as *calibrated Bayes*.

In simulation Study 1, all 10 generating covariates are used in the propensity score estimation model for all the methods investigated in this article. Specifically, for the Bayesian model averaging approach using model-averaged propensity score estimates provided by the original *BMA* program, we evaluate the effects of Occam's window on the accuracy and precision of the treatment effect estimation. When Occam's window is on, the default value 20 in the *BMA* program is used as the maximum ratio for excluding models in Occam's window. When Occam's window is off, all the models are selected and used for obtaining the averaged propensity score estimates.

For the fully MCMC Bayesian model averaging propensity score approach, three different cumulative posterior model probabilities (50%, 70%, and 90%) are adopted to investigate whether the number and quality of the selected models affect the treatment effect estimation. Also, both noninformative and informative prior distributions are utilized in the propensity score model at Step 2, where the generating coefficients in the propensity score model are used as informative hyperparameters for the prior distributions.

In simulation Study 2, only 8 out of 10 generating covariates are included in the propensity score model to examine the performance of the proposed Bayesian model averaging methods when the propensity score model is misspecified. The covariates x_9 and x_{10} are excluded for the propensity score estimation. For the sake of brevity, we only explore the BMA-MCMC approach for models with the top 50% cumulative posterior model probabilities and with noninformative priors.

RESULTS OF SIMULATION STUDY 1

With the exception of a small number of cases described later, all models showed convergence of the Gibbs sampler based on inspection of trace and density plots. Table 1 shows that, overall, Occam's window does not influence treatment effect estimation for the approximate Bayesian model averaging approach. The proposed fully Bayesian model averaged propensity score methods perform similarly when the models with the top 50%, 70%, or 90% cumulative posterior model probabilities are used for propensity score estimation. In other words, these findings indicate that models using the top 50% cumulative posterior model probabilities contribute the most to the parameter estimation and bias reduction.[2]

[2]For this article, we are not computing a percentage under-or over estimation to represent "bias". We are simply comparing the obtained treatment effects from the simulations with the true parameter value.

TABLE 1
Treatment (Trt) Effect and Standard Error (SE) Estimates in Simulation Study 1
(Avg. Trt Effect(SD of Trt Effect Estimates)/Avg. SE(SD of SE Estimates))

Method	Stratification	Opt. Matching	Regression	Weighting
BMA-Approx+OLS				
Occam's window on	4.93(.19)/.23(.03)	4.98(.18)/.21(.02)	4.99(.17)/.20(.01)	4.92(.21)/.18(.01)
Occam's window off	4.94(.20)/.23(.03)	4.99(.17)/.21(.02)	4.99(.17)/.20(.01)	4.91(.21)/.18(.01)
BMA-Approx+Bayes				
Occam's window on	4.93(.19)/.23(.03)	4.99(.18)/.21(.02)	4.99(.17)/.19(.01)	NA
Occam's window off	4.93(.20)/.23(.03)	4.99(.18)/.21(.02)	4.99(.17)/.19(.01)	NA
BMA-MCMC+OLS				
(Noninformative)				
Top 50%	4.89(.19)/.25(.04)	4.92(.17)/.24(.02)	4.93(.17)/.21(.02)	4.99(.24)/.25(.06)
Top 70%	4.88(.18)/.26(.04)	4.92(.17)/.24(.02)	4.93(.17)/.22(.02)	4.97(.23)/.25(.06)
Top 90%	4.88(.18)/.26(.04)	4.92(.17)/.24(.02)	4.92(.17)/.22(.02)	4.97(.23)/.25(.06)
BMA-MCMC+Bayes				
(Noninformative, 50%)	4.88(.18)/.25(.04)	4.93(.17)/.24(.02)	4.93(.17)/.21(.02)	NA
BMA-MCMC+OLS				
(Informative)				
Top 50%	4.89(.18)/.26(.04)	4.93(.17)/.24(.02)	4.93(.17)/.22(.02)	4.98(.23)/.25(.06)
Top 70%	4.88(.18)/.26(.04)	4.92(.17)/.24(.02)	4.93(.17)/.22(.02)	4.97(.23)/.25(.06)
Top 90%	4.88(.18)/.26(.04)	4.92(.17)/.24(.02)	4.92(.17)/.22(.02)	4.97(.23)/.25(.06)
Two-step BPSA	4.88(.18)/.26(.04)	4.93(.17)/.25(.02)	4.93(.17)/.22(.02)	NA

Note. The Bayesian propensity score weighting approach with Bayesian outcome model is not discussed here due to the absence of Bayesian weighted regression in the propensity score literature. BPSA = Bayesian propensity score analysis; BMA-Approx+OLS = Approximate Bayesian model averaging with ordinary least squares outcome model; BMA-Approx+Bayes = Approximate Bayesian model averaging with Bayesian outcome model; BMA-MCMC+OLS = Fully Bayesian model averaging with OLS outcome model; BMA-MCMC+Bayes = Fully Bayesian model averaging with Bayesian outcome model.

Priors on the propensity score model parameters have little impact on the treatment effect estimation, but it is important to point out that this finding could be due to the particular choice of priors in this study.

Among all the methods examined, the proposed BMA-MCMC approach based on models using the top 50% cumulative posterior model probabilities provides slightly less biased treatment effect estimate and the more accurate uncertainty estimate for the weighting approach. On average, BMA-Approx provides slightly less biased treatment effect estimates. However, BMA-Approx slightly underestimates the uncertainty in the treatment effect for the weighting method (0.18 vs. 0.21 and 0.18 vs. 0.24, respectively). The proposed BMA-MCMC approach overall produces larger uncertainty estimates as expected, but it tends to overestimate the variation. This might be due to the small number of replications in the simulation study. When we increased the number of data replications to 500 for the BMA-MCMC propensity score model with the ordinary least squares (OLS) outcome model, the treatment effect and uncertainty estimates were very similar to the results with 200 data replications. The R code for simulation study 1 is available upon request.

RESULTS OF SIMULATION STUDY 2

From Table 2, we observe that all of the Bayesian propensity score methods are quite robust to the misspecification of propensity score model. Specifically, the BMA-Approx approach provides slightly less biased estimates for optimal full matching and weighting methods, whereas the BMA-MCMC approach offers somewhat better treatment effect estimate for the weighting method. In terms of the uncertainty estimates, the BMA-Approx slightly underestimates the variation in the treatment effect. Out of all the Bayesian methods investigated, the proposed BMA-MCMC approach under weighting and the BMA-Approx approach under optimal matching offer the closest standard error estimates to the approximately true uncertainty (0.24 vs. 0.23 and 0.20 vs. 0.19, respectively). The R code for simulation Study 2 is also available upon request.

DESIGN OF THE CASE STUDY

In order to further investigate the properties of the proposed BMA procedures for propensity score analysis, we conducted a case study using real data from the Early Childhood Longitudinal Study Kindergarten cohort of 1998 (ECLS-K; National Center for Education Statistics, 2001). The sampled children attended either full-day or part-day kindergarten programs and had diverse socioeconomic and racial/ethnic backgrounds. Also, a number of variables assessing early childhood (pre-K) experiences were collected in the ECLS-K, thus propensity score approaches can be fruitfully applied. This article investigates the treatment effect of full-day

TABLE 2
Treatment (Trt) Effect and Standard Error (SE) Estimates in Simulation Study 2
(Avg. Trt Effect(SD of Trt Effect Estimates)/Avg. SE(SD of SE Estimates))

Method	Stratification	Opt. Matching	Regression	Weighting
BMA-Approx+OLS	4.95(.19)/.22(.03)	4.99(.19)/.20(.02)	5.01(.17)/.19(.01)	4.95(.21)/.18(.01)
BMA-Approx+Bayes	4.95(.19)/.22(.03)	5.00(.19)/.21(.02)	5.01(.17)/.19(.01)	NA
BMA-MCMC+OLS (Noninf, 50%)	4.91(.18)/.24(.03)	4.95(.17)/.23(.02)	4.96(.17)/.20(.02)	5.00(.23)/.24(.06)
BMA-MCMC+Bayes (Noninf, 50%)	4.91(.18)/.23(.03)	4.96(.18)/.23(.02)	4.96(.17)/.20(.02)	NA
Two-step BPSA	4.92(.18)/.24(.03)	4.96(.17)/.23(.02)	4.97(.17)/.20(.02)	NA

Note. BPSA = Bayesian propensity score analysis; BMA-Approx+OLS = Approximate Bayesian model averaging with ordinary least squares outcome model; BMA-Approx+Bayes = Approximate Bayesian model averaging with Bayesian outcome model; BMA-MCMC+OLS = Fully Bayesian model averaging with OLS outcome model; BMA-MCMC+Bayes = Fully Bayesian model averaging with Bayesian outcome model.

versus part-day kindergarten attendance on children's reading achievement at the end of 1998 fall kindergarten.

A sample of 1,000 children were randomly selected proportional to the number of children in full-day or part-day kindergarten in the population. This resulted in 538 children in full-day programs and 462 children in part-day programs. Fourteen covariates were chosen for estimating the propensity scores, including gender, race, mother's employment status, child's age at kindergarten entry, child's age at first nonparental care, primary type of nonparental care, both parent language to child, number of siblings, family composition, mother's employment between child's birth and kindergarten, number of nonparental care arrangements (pre-K), social economic status, parent's expectation of child's degree, and how often parent reads to child. Missing data were handled via the multivariate imputation by chained equations using the R function mice in the *mice* package (Van Buuren & Groothuis-Oudshoorn, 2011) with one iteration. The diagnostics were performed as default and the chains converged.

Similar to the first simulation study, we examine the effects of Occam's window on the accuracy and precision of the treatment effect estimation for the approximate Bayesian model averaging propensity score approach. When Occam's window is on, the default value 20 in the *BMA* program is utilized as the maximum ratio for excluding models in Occam's window, and this yields four selected models with the posterior model probabilities 0.31, 0.29, 0.23, and 0.17, respectively. When Occam's window is off, there are 184 models being selected and used for obtaining the model averaged propensity score estimates. For the BMA-MCMC approach, we utilize all four selected models when Occam's window is on. Here, however, we do not examine the BMA-MCMC approach when Occam's window is off because there are 184 selected models in total and the estimation of the Bayesian logit models encounter problems for many models with small posterior model probabilities that include very few variables. In this case, certain cumulative posterior model probabilities (e.g., top 50%) can be used to restrict the selected models to the most predictive ones or Occam's window can stay open.

Covariate Balance

For this case study, we also examine the covariate balance performance of the Bayesian model averaging approaches. The continuous covariates and categorical covariates are evaluated separately. The balance indices used in this study are the standardized mean/proportion difference (Cohen's d; Cohen 1988) and variance ratio for each continuous covariate/each level of categorical covariates between treatment group and control group. Specifically, the standardized mean difference for a continuous covariate is obtained by

$$B_1 = (\bar{x_t} - \bar{x_c})/\sqrt{(s_t^2 + s_c^2)/2}, \tag{16}$$

where $\bar{x_t}$ and $\bar{x_c}$ are the sample mean of each covariate in treatment and control groups, respectively, and s_t^2 and s_c^2 are corresponding sample variances. The variance ratio for a continuous covariate, R_1, is defined as s_t^2/s_c^2. All the categorical covariates are dummy coded. Then for each categorical level, we evaluate the standardized difference in proportions between different treatment conditions, consistent with Harder, Stuart and Anthony (2010). The standardized proportion difference is calculated by

$$B_2 = (\hat{p}_t - \hat{p}_c)/\sqrt{[\hat{p}_t(1 - \hat{p}_t) + \hat{p}_c(1 - \hat{p}_c)]/2}, \tag{17}$$

where $\hat{p}_t$ and $\hat{p}_c$ are proportions of participants in the treatment group and control group, respectively, for a specific level of categorical covariates. The variance ratio for a certain categorical level, R_2, is calculated by $\hat{p}_t(1 - \hat{p}_t)/\hat{p}_c(1 - \hat{p}_c)$.

In addition to the aforementioned point estimates, the BMA-MCMC approach provides the 95% posterior probability intervals (PPI) of the standardized mean/proportion difference and the 95% PPI of the variance ratio based on the posterior distribution of the propensity score parameters. For each posterior propensity score drawn from the posterior distribution, we can obtain a point estimate of each balance index. As we have 1,000 posterior draws of the propensity score, a distribution of the balance index with 1,000 points can be obtained and the mean of this posterior distribution

provides the point estimate of the covariate balance and the 2.5 and 97.5 percentiles to form the corresponding 95% PPI.

Note that the optimal full matching method matches subjects as fine as possible to minimize the total distance among all the matched strata. Optimal full matching yields more than 300 optimal strata in our case study, many of which have only one subject in the treatment group or in the control group such that the calculation of covariate variance within each stratum becomes infeasible. Also, it is hard to obtain covariate variance for treatment group and control group, respectively, in the regression adjustment method where the posterior propensity scores are utilized as a covariate in the outcome model. So we use the unadjusted covariate variances in the treatment group and in the control group to calculate Cohen's d for the optimal matching and regression adjustment methods and do not show the variance ratio for these two methods.

Prediction of the Propensity Score

This case study also assessed the predictive performance of the two Bayesian model averaging propensity score approaches and compared them with the Bayesian propensity score approach with a single propensity score equation. As we are interested in the prediction of the propensity score for a binary treatment scenario, that is, the predicted probability of being in the full-day kindergarten, we adopted the *Brier score* (Brier, 1950) to evaluate the quality of the predicted posterior propensity scores. The Brier score has been used to evaluate the predictive performance of Bayesian model averaging in the context of binary responses (Yeung, Bumgarner & Raftery, 2005). The literature on Bayesian model averaging suggests that predictive performance under Bayesian model averaging should be better than prediction obtained under any single model (Madigan & Raftery, 1994).

In this study, we first randomly split the data set into two halves, one training data set and one testing data set. The propensity score model parameters are estimated using the training data set and then the predicted posterior propensity scores are obtained based on the testing data set and the estimated propensity score model parameters. The posterior mean of the predictive propensity scores for each student is denoted as p_i. Let T_i denote the treatment selection, half-day or full-day kindergarten program, for student i, where T_i = 0 or 1. Then the Brier score is defined as $\sum_{i=1}^{n}(T_i - p_i)^2$, which is the sum of squares of the differences between the treatment selection and the predicted probability over n students. The smaller the model's Brier score across competing models, the better prediction the model makes. To account for the randomness of data splitting, we randomly split the data set into two equal halves 100 times and then take the average of the Brier score across 100 data splittings for the models investigated here.

RESULTS OF THE CASE STUDY

We conducted a Bayesian regression with the full-day or half-day kindergarten selection as the only predictor for children's reading achievement in the fall kindergarten and obtained the unadjusted treatment effect 1.74 with standard error 0.77. This result serves as a "negative benchmark" to check whether the propensity score approaches investigated in this study adjust the bias due to self-selection of the treatment conditions. For comparison purposes, we also fit a Bayesian covariate-adjusted regression with the full-day or half-day kindergarten selection and all the 14 covariates as predictors. The treatment effect estimate is 2.30 with standard error 0.72.

The treatment effect and standard error estimates of the case study shown in Table 3 indicate that all the propensity score methods adjusted the selection bias to some extent. Bayesian model averaging approaches provide results comparable to other propensity score approaches. The effects of Occam's window vary across different propensity score methods. Turning Occam's window on or off does not unduly influence treatment effect estimation for stratification, weighting, and regression adjustment. However, turning Occam's window on or off does slightly affect the treatment effect estimates under the optimal full matching method. In addition, using OLS or Bayesian outcome models has little impact on treatment effect estimation. The R code for the BMA-Approx and BMA-MCMC propensity score analyses for the case study is available upon request.

The covariate balance results of the case study are presented in Table 4. Overall, the average absolute standardized mean/proportion differences across all the covariates are within the ideal range ($\pm$ 0.1 standard deviation) for all the investigated propensity score approaches, indicating that the propensity score adjustments effectively reduced the selection bias in the covariates. Bayesian model averaging approaches provide comparably good covariate balance compared to the two-step Bayesian propensity score approach without model averaging. According to the balance criteria presented in Rubin (2001), the average variance ratios in the treatment group and control group across all the covariates fall in the acceptable range (between 1/2 and 2) for all the examined propensity score approaches. The average variance ratios for all the examined Bayesian approaches are in the desirable range (between 4/5 and 5/4).

In addition to the mean estimates, the BMA-MCMC approach and the two-step Bayesian propensity score approach also offer the 95% posterior probability intervals of the balance indices (Cohen's d and variance ratio here) so that the variation in the estimated balance indices can be captured. For instance, although the mean Cohen's d estimates of the Bayesian propensity score approaches are quite similar for stratification, optimal matching, regression, and weighting, the 95% posterior probability intervals of the Cohen's d and

TABLE 3
Treatment Effect (Trt) and Standard Error (SE) Estimates in the Case Study (Trt(SE))

Method	Stratification	Opt. Matching	Regression	Weighting
Propensity score methods				
BMA-Approx+Bayes				
Occam's window on	2.23(.75)	2.56(.86)	2.51(.78)	NA
Occam's window off	2.23(.75)	2.52(.83)	2.51(.78)	NA
BMA-Approx+OLS				
Occam's window on	2.25(.78)	2.55(.83)	2.51(.79)	2.52(.76)
Occam's window off	2.25(.78)	2.48(.86)	2.51(.79)	2.52(.76)
BMA-MCMC+Bayes				
Occam's window on	2.08(.81)	2.42(.87)	2.47(.81)	NA
BMA-MCMC+OLS				
Occam's window on	2.09(.84)	2.42(.86)	2.47(.82)	2.57(.84)
Two-step BPSA	2.02(.83)	2.22(.95)	2.23(.84)	NA
Other methods	Estimate(S.E.)			
Bayes unadjusted regression	1.74(.77)			
Bayes covariate-adjusted regression	2.30(.72)			

Note. BPSA = Bayesian propensity score analysis; BMA-Approx+OLS = Approximate Bayesian model averaging with ordinary least squares outcome model; BMA-Approx+Bayes = Approximate Bayesian model averaging with Bayesian outcome model; BMA-MCMC+OLS = Fully Bayesian model averaging with OLS outcome model; BMA-MCMC+Bayes = Fully Bayesian model averaging with Bayesian outcome model.

variance ratio for the weighting method are wider than the intervals for stratification, optimal matching, and regression methods, implying more uncertainty in the estimation of balance indices for the weighting method.

In terms of the predictive performance, across 100 data splittings, the average Brier Scores of the BMA-Approx approach and the BMA-MCMC approach are 121.83 and 121.84, respectively, whereas the average Brier Score of the two-step Bayesian propensity score approach with a single propensity score equation is 122.34. Thus, as theoretically expected, both Bayesian model averaging propensity score approaches showed advantage in the prediction of the propensity score compared with the Bayesian propensity score approach that ignored model uncertainty. The R code for the covariate balance check and the predictive performance check of the case study are available upon request.

SUMMARY AND CONCLUSION

The purpose of this article is to provide a Bayesian model averaging approach within the propensity score framework. In particular, we propose a fully Bayesian approach to propensity score analysis to account for uncertainty in the propensity score model itself. In this sense, our fully Bayesian model averaging approach addresses the problem of covariate choice as a problem of model choice and directly recognizes that

TABLE 4
Average Absolute Standardized Mean/Proportion Difference (Cohen's *d*) and Variance Ratio Between Treatment and Control Group Across All Covariates in the Case Study

Method	Stratification	Opt. Matching	Regression	Weighting
BMA-Approx. approach				
Cohen's *d*	0.06	0.05	0.05	0.09
Variance ratio	1.01			0.82
BMA-MCMC approach				
Cohen's *d*	0.07	0.06	0.06	0.10
95% PPI of Cohen's *d*	(0.06, 0.09)	(0.05, 0.07)	(0.05, 0.07)	(0.06, 0.15)
Variance ratio	1.01			1.01
95% PPI of variance ratio	(0.96, 1.07)			(0.61, 1.93)
Two-step BPSA				
Cohen's *d*	0.05	0.05	0.04	0.08
95% PPI of Cohen's *d*	(0.04, 0.07)	(0.04, 0.07)	(0.03, 0.06)	(0.05, 0.12)
Variance ratio	1.00			1.00
95% PPI of variance ratio	(0.95, 1.06)			(0.52, 2.13)

Note. The adjusted variances in the treatment group and in the control group are not available for the optimal full matching and regression adjustment. Therefore, the Cohen's *d* for these two methods are calculated based on the unadjusted variances and the variance ratios of these two methods are not shown. PPI = Posterior probability interval.

such uncertainty needs to be accounted for in the creation of the propensity score.

To sum up, our simulation studies reveal that both Bayesian model averaging propensity score methods perform similarly when models/covariates using the top 50%, 70%, or 90% cumulative posterior model probabilities are used for propensity score estimation. For this study, priors on the propensity score model parameters were shown to have little impact on the treatment effect estimation. This might not always hold true, particularly in the case where high precision is placed on poorly elicited prior values or in the case of noninformative priors in the context of small sample size problems. Our new BMA-MCMC approach using models that have the top 50% cumulative posterior model probabilities provides slightly less biased treatment effect estimates for the weighting approach compared with the BMA-Approx approach. Our proposed BMA-MCMC approach overall produces somewhat larger uncertainty estimates as expected, but it tends to overestimate the variation. As noted earlier, we increased the number of data replications to 500 for the BMA-MCMC propensity score model with the frequentist ordinary least squares outcome model, but the treatment effect and uncertainty estimates are very similar to the results with 200 data replications. This slight overestimation in uncertainty has been observed in other Bayesian research (e.g., Yuan & MacKinnon, 2009) and might be because the simulation studies follow the frequentist framework, thus favoring a frequentist outcome.

With regard to omitted variables in the propensity score equation, we found that all propensity score methods were robust to misspecification of the propensity score model. Although differences are small, the BMA-Approx approach provides the least biased estimates for optimal full matching and regression adjustment methods, whereas the BMA-MCMC approach offers the best treatment effect estimate for the weighting method. In terms of the uncertainty estimates, the BMA-Approx under the weighting approach underestimates the variation in the treatment effect. Out of all the Bayesian methods investigated, our proposed BMA-MCMC weighting approach and the BMA-Approx under optimal matching approach offer the closest standard error estimates to the approximately true uncertainty.

We also conducted a real-data study, where the ECLS-K data set is used for examining the recovery of the treatment effect estimates in the quasi-experiment via different propensity score methods. In addition, the effects of Occam's window on the treatment effect estimates are evaluated. Results reveal that the impact of Occam's window varies slightly across different propensity score methods, but overall, Bayesian model averaging propensity score approaches perform well and are comparable to other propensity score approaches and some other bias-reduction methods such as covariate-adjusted regression. The proposed Bayesian model averaging approaches also provide good covariate balance and better prediction of the propensity score compared with the Bayesian approach that ignores model uncertainty.

A number of open questions remain that were beyond the scope of this article to explore. In particular, we chose either accurate priors (based on the generating coefficients of the propensity score model) or noninformative priors based on the uniform distribution. It would be important to explore the sensitivity of our results to other choices of priors—particularly the case where priors are poorly elicited. This would reflect the case of demonstrating relative certainty around incorrect prior values. Our simulations and case study were also examined under simple default conditions regarding the width of Occam's window. Here too it would be important to explore the sensitivity of our approach to widely varying choices of the width of Occam's window. Finally, future studies should examime the case where all models have small posterior model probabilities. A question that can be raised is whether Bayesian model averaging in the case of small posterior model probabilities still provides improved posterior prediction compared with the use of any one model.

To conclude, this article addresses covariate selection in propensity score analysis as a problem of model selection. From a Bayesian perspective, the problem of model selection is a problem of model uncertainty. At present, the main approach to addressing model uncertainty in Bayesian inference is through Bayesian model averaging. Thus, our article accomplishes two goals. First, we provide an extension of Kaplan and Chen's (2012) two-step propensity score approach by addressing the problem of model selection from a Bayesian standpoint. Second, we provide a new fully Bayesian approach to model averaging within the propensity score context and compare it with an existing approximate approach to Bayesian model averaging. Overall, the differences are slight, and so choosing either the BMA-Approx approach or the BMA-MCMC approach are not likely to dramatically affect subsequent inferences. To conclude, we offer a Bayesian propensity score analysis that accounts fully for model and parameter uncertainty and can be adopted by those who wish to do causal inference within the Bayesian paradigm.

FUNDING

The research reported in this article was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D110001 to the University of Wisconsin–Madison. The opinions expressed are those of the authors and do not necessarily represent views of the Institute or the U.S. Department of Education.

REFERENCES

An, W. H. (2010). Bayesian propensity score estimators: Incorporating uncertainties in propensity scores into causal inference. *Sociological Methodology*, *40*, 151–189.

Box, G. E. P. (1980). Sampling and Bayes inference in scientific modeling and robustness (with discussion). *Journal of the Royal Statistical Society, Series A*, *143*, 383–430.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*, 1–3.

Chen, J., & Kaplan, D. (2014). Covariate balance in bayesian propensity score approaches forobservational studies. *Journal for Research on Educational Effectiveness*. doi: 10.1080/19345747.2014.911396

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences,* (2nd ed). Hillsdale, NJ: Erlbaum.

de Finetti, B. (1974). *Theory of probability: Vols. 1 and 2*. New York NY: Wiley.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, *94*, 1053–1062.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis,* (2nd ed). LondonUK: Chapman & Hall.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, *99*, 609–618.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, *15*, 234–249.

Hirano, K., & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, *2*, 259–278.

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

Hoshino, T. (2008). A Bayesian propensity score adjustment for latent variable modeling and MCMC algorithm. *Computational Statistics & Data Analysis*, *52*, 1413–1429.

Kaplan, D., & Chen, J. (2012). A two-step Bayesian approach for propensity score analysis: Simulations and case study. *Psychometrika*, *77*, 581–609.

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

Little, R. J. (2011). Calibrated Bayes, for statistics in general, and missing data in particular. *Statistical Science*, *26*, 162–174.

Little, R. J. (2013). In praise of simplicity not mathematistry! Ten simple powerful ideas for the statistical scientist. *Journal of the American Statistical Association*, *108*, 359–369.

Lunceford, J., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, *23*, 2937–2960.

Lunn, D., Best, N., Spiegelhalter, D., Graham, G., & Neuenschwander, B. (2009). Combining MCMC with "sequential" PKPD modelling. *Journal of Pharmacokinetics and Pharmacodynamics*, *36*, 19–38.

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*, 1535–1546.

Martin, A. D., Quinn, K. M., & Park, J. H. (2010). *Markov chain Monte Carlo (MCMC) package*. Retrieved from http://mcmcpack.wustl.edu/

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*, 403–425.

McCandless, L. C., Douglas, I. J., Evans, S. J., & Smeeth, L. (2010). Cutting feedback in Bayesian regression adjustment for the propensity score. *The International Journal of Biostatistics, 6*, Article 16.

McCandless, L. C., Gustafson, P., & Austin, P. C. (2009). Bayesian propensity score analysis for observational data. *Statistics in Medicine*, *28*, 94–112.

National Center for Education Statistics. (2001). *Early childhood longitudinal study: Kindergarten class of 1998-99: Base year public-use data files user's manual* (Tech. Rep. No. NCES 2001-029). Washington, DC: U.S. Department of Education.

R Development Core Team. (2011). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org (ISBN 3-900051-07-0)

Raftery, A. E., Hoeting, J., Volinsky, C., Painter, I., & Yeung, K. Y. (2009, September 18). *Bayesian model averaging (BMA), version 3.12*. Retrieved from http://www2.research.att.com/volinsky/bma.html

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, *84*, 1024–1032.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*, 516–524.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, *74*, 318–328.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, *12*, 1151–1172.

Rubin, D. B. (1985). The use of propensity scores in applied Bayesian inference. *Bayesian Statistics*, *2*, 463–472.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, *2*, 169–188.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, *13*, 279–313.

Steiner, P. M., Cook, T. D., & Shadish, W. R. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics*, *36*(2), 213–236.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, *15*, 250–267.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1–67.

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*, 2394–2402.

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods*, *14*, 301–322.