Causal Inference Using Semi-parametric Imputation

Andrea Piesse[1], Laura Alvarez-Rojas[1], David Judkins[1]
William R. Shadish[2]
[1]Westat, 1600 Research Boulevard, Rockville, MD 20850
[2]University of California, School of Social Sciences, Humanities and Arts,
5200 North Lake Road, Merced, CA 95343

Abstract
In nonrandomized studies, selection bias may confound the relationship between treatment and outcome. Imputation is one method for addressing selection bias, though it is not widely used. With this approach, a potential outcome is imputed for each treatment level not received and the association between treatment and outcome is estimated using both reported and imputed outcomes. Multiple imputations may be used to account for the impact of the imputation process on variances. This paper analyzes data from a four-arm comparison study (Shadish, Clark, and Steiner, 2008) where students were first randomly assigned to a randomized experiment or an observational study. Using the randomized experiment as a benchmark, we examine treatment effects estimated by applying semi-parametric multiple imputation to the observational study and compare them to effects estimated using other analytic approaches. Issues with variance estimation are discussed.

Key Words: Causal inference, potential outcomes, counterfactual, hot deck, multiple imputation, bootstrap

1. Introduction

A randomized experiment is a study design in which subjects are randomly assigned to treatment groups. Apart from chance imbalance, the distributions of all covariates (measured and unmeasured) are expected to be the same across the treatment groups. For this reason, statistically significant differences in outcomes between the groups are likely due to treatment and the randomized experiment is considered the gold standard for causal inference.

However, there may be practical or ethical issues that prevent the use of random treatment assignment for a given study. In these situations, we have a quasi-experimental design (sometimes referred to as an observational study or nonrandomized experiment). Often treatment is self-selected by the study subjects and as a result, there is no guarantee that covariate distributions will be the same across the treatment groups. Relationships between treatment and outcomes may be confounded by selection bias and there is a need for some form of adjustment when estimating causal effects.

1.1 Potential Outcomes Framework and Notation
Causal effects are comparisons among the outcomes that a study subject would have under different treatment conditions. In other words, the notion of causality pertains to how some form of treatment, exposure, or intervention would change a subject's

outcomes. To characterize a treatment effect in a randomized experiment, Neyman (1923) introduced multiple outcomes for each experimental unit. Rubin (1974, 1978a) proposed a similar notation in the context of observational studies, producing a framework that is often referred to as Rubin's causal model. Let T denote treatment status, where for simplicity we assume that $T = 0$ or 1, and let Y denote an outcome of interest. Using the potential outcome framework, let $Y(0)$ denote the outcome if $T = 0$ and $Y(1)$ the outcome if $T = 1$. Suppose that we want to estimate the average causal effect of T on Y,

$$\Delta = E\left[Y(1) - Y(0)\right].$$

Because no study subject can receive multiple treatments at the same time, one of the two potential outcomes is missing for each subject. The missing (i.e., unobserved) potential outcomes are sometimes called counterfactual outcomes. Therefore, the fundamental problem of causal inference (Holland, 1986) is that for no subject do we observe the causal effect of treatment.

2. Analytic Methods for Observational Studies

2.1 A Missing Data Problem

The potential outcome framework makes it clear that causal inference may be regarded as a missing-data problem, as illustrated below.

Treatment status	$Y(0)$	$Y(1)$
$T = 0$	Observed	Not observed
$T = 1$	Not observed	Observed

Figure 1: Illustration of missing-data status among potential outcomes with two treatment levels

Two standard methods of dealing with missing data are weighting and imputation. Using the weighting approach, quantities involving $Y(0)$ are estimated by weighting up the study subjects with $T = 0$ to also represent the subjects with $T = 1$. Similarly, estimates involving $Y(1)$ are obtained by weighting up the subjects with $T = 1$ to also represent the subjects with $T = 0$. The weighting adjustments are typically performed within groups of subjects having similar estimated treatment propensities. The average treatment effect is then estimated by

$$\hat{\Delta} = \frac{\sum_{i \text{ where } T=1} w_i' Y_i(1)}{\sum_{i \text{ where } T=1} w_i'} - \frac{\sum_{i \text{ where } T=0} w_i' Y_i(0)}{\sum_{i \text{ where } T=0} w_i'},$$

where w_i' is the adjusted weight for study subject i. This approach is sometimes referred to as inverse-probability weighting based on estimated propensity scores or counterfactual projection weighting.

Using the imputation approach, missing values for the counterfactual or unobserved potential outcomes are replaced by imputed values. Here, the imputation process is typically performed within groups of subjects having similar characteristics and/or predicted outcomes. The average treatment effect in then estimated by

$$\hat{\Delta} = \frac{\sum_{i \text{ where } T=1} w_i Y_i(1) + \sum_{i \text{ where } T=0} w_i Y_i^{imp}(1)}{\sum_i w_i} - \frac{\sum_{i \text{ where } T=1} w_i Y_i^{imp}(0) + \sum_{i \text{ where } T=0} w_i Y_i(0)}{\sum_i w_i},$$

where $Y_i^{imp}(t)$ is the imputed potential outcome under $T = t$, and w_i is the regular survey weight for study subject i.

2.2 Motivation for Imputation Approach

Numerous other analytic techniques exist and have been used to estimate causal effects. One of the most common is ANCOVA whereby outcomes are regressed on treatment group indicators, eligible covariates, and relevant interactions. Inference about treatment effects is based on the regression coefficients of the treatment indicator variables (and of any interaction terms involving treatment). Other methods attempt to match study subjects from different treatment groups, using available covariates or a summary measure such as the estimated propensity score in an effort to match subjects that are "similar" apart from the treatment received. Estimation of treatment effects is then based on the differences between outcomes within the matched pairs. Stratifying the sample by propensity score and conducting a stratified analysis of outcome differences attempts something similar, except that the matching is at a coarser level. In addition, it is possible to combine aspects of these different analytic techniques in an attempt to make the results more closely resemble those of a randomized experiment.

Aside from general interest in the performance of imputation for causal inference relative to some of the more common approaches we have just mentioned, there may be reasons to prefer the imputation approach. One is that imputed potential outcomes can simplify the communication of findings to stakeholders. For example, proper interpretation of a logistic regression coefficient may be more difficult for a client than simply comparing estimates of the percentage of the target population that would exhibit a certain trait in the complete absence of treatment versus treatment saturation. A second reason to consider the imputation approach is that the impact of missing data on survey estimates can be reflected using multiple imputation.

First proposed by Rubin (1978b), multiple imputation can be used to capture the added uncertainty in treatment effect estimates that is due to missing data. The method involves performing $M \geq 2$ independent imputations to create M complete data sets. The multiple imputation estimator of the treatment effect, Δ, is the average of the estimators obtained from each completed data set,

$$\hat{\Delta}_{MI} = \frac{1}{M} \sum_{m=1}^{M} \hat{\Delta}_m .$$

Discussion of the multiple imputation variance estimator can be found in Section 3.

If one decides to adopt the imputation approach to causal inference, a choice of imputation method must be made. In general, a wide variety of imputation procedures exists: random or deterministic; cyclic or not; weighted or unweighted; single or multiple; nearest neighbour or random within cell; Bayesian or semi-parametric, etc. In the limited applications of imputation for causal inference, it would appear that Bayesian imputation methods are most frequently used. However, we preferred to investigate the performance of a semi-parametric approach for two main reasons.

Westat has already developed a highly-automated, semi-parametric imputation software product known as AutoImpute (see, for example, Krenzke and Judkins, 2008), which has been found to perform well when dealing with item nonresponse and in comparison to other software packages (Judkins et al., 2007). AutoImpute is based on iterative cycling through p-partition hot decks, starting with a simple hot deck to fill in values for all missing items. Subsequent passes through the data set re-impute each item sequentially, with donors chosen via model-based estimates of the item being imputed, until specified convergence criteria are reached. This approach is a semi-parametric analogue of the parametric conditional imputation methods in the software packages IVEWare (Raghunathan et al., 2001) and MICE (Van Buuren and Oudshoorn, 1999). Therefore, AutoImpute is clearly less dependent on parametric assumptions than Bayesian imputation methods and is easy to implement. Because AutoImpute was originally developed to handle regular item nonresponse, its extension to the imputation of potential outcomes also provides a way to fully integrate uncertainty due to missing data in outcomes *and* covariates, into the estimation of causal effects.

3. Variance Estimation

It is well known that treating imputed values as if they were reported values leads to variance estimates that understate the true variances of survey estimates. This "naïve" approach underestimates the degree of uncertainty because it ignores the variability due to nonresponse. When imputation is applied to potential outcomes this is of particular relevance due to the large amount of missing data.

There are three main approaches to valid variance estimation using imputed data: explicit formulae that incorporate nonresponse; resampling methods designed to take account of the imputation procedure; and multiple imputation. Explicit formulae may be derived by making assumptions about the missing data mechanism or the model for the distribution of Y in the population (e.g., Särndal, 1992; Chen and Shao, 2000; Brick et al., 2004; Kim and Rao, 2010). Because of the vulnerability of such approaches to model violations, resampling methods are another alternative. These include the adjusted jackknife of Rao and Shao (1992), the fractionally weighted hot deck of Kim and Fuller (2004), and the bootstrap approach of Shao and Sitter (1996). Most of these proposed methods discuss only univariate variance estimation and may rely on imputation cell information such as the mean of the respondent values in each cell. However, this approach breaks down in the causal inference setting because treatment effects are necessarily multivariate in nature, being the difference (or some other function) of two or more potential outcomes. As a consequence, no imputation cell has a respondent!

Whilst its use may come at the expense of being computationally more expensive, the Shao-Sitter bootstrap appears to be one method that will lead to valid variance estimation in the context of imputed potential outcomes. The approach involves drawing B

independent bootstrap samples with replacement from the original sample, carrying out the same imputation procedure on each of the bootstrap samples, and applying the usual bootstrap mean and imputation formulae. The bootstrap estimator of the treatment effect, Δ, is the average of the estimators obtained from each completed bootstrap data set,

$$\hat{\Delta}_{BT} = \frac{1}{B}\sum_{b=1}^{B}\hat{\Delta}_b \ ,$$

and the variance of $\hat{\Delta}_{BT}$ is given by

$$\hat{V}_{BT} = \frac{1}{(B-1)}\sum_{b=1}^{B}\left(\hat{\Delta}_b - \hat{\Delta}_{BT}\right)^2 .$$

Multiple imputation (see Section 2.1) is another potentially viable approach to variance estimation for imputed potential outcomes. The variance of $\hat{\Delta}_{MI}$ (the multiple imputation estimator of the treatment effect) is the sum of the average within-imputation variance and the between-imputation variance, with a bias correction for the finite number of multiply imputed data sets,

$$\hat{V}_{MI} = \frac{1}{M}\sum_{m=1}^{M}\hat{V}_m + \left(1+\frac{1}{M}\right)\frac{1}{(M-1)}\sum_{m=1}^{M}\left(\hat{\Delta}_m - \hat{\Delta}_{MI}\right)^2 .$$

The multiple imputation method has risen in popularity due to the relative ease with which it can be applied, however there are situations in which it does not produce valid variance estimates. For example, if hot-deck imputation is used and the donor pool for a respondent is the same for all M data sets, the method is not a "proper" multiple imputation procedure (Rubin, 1978b). In this case, the true variance is underestimated even with an infinite number of imputed data sets, and the degree of underestimation may be considerable if a large amount of data are being imputed.

4. Application Using Real Data

To assess the performance of semi-parametric multiple imputation for causal inference, we used real data from a study of volunteer undergraduate students taking introductory psychology classes at a large mid-southern public university. In the first example of a four-arm within-study comparison design, Shadish, Clark, and Steiner (2008) pre-tested these students in different domains before randomly assigning them either to a randomized experiment with two treatment arms (learning about vocabulary or mathematics) or to a quasi-experiment with self-selection into these same two arms. After training, the mathematics and vocabulary scores of all participants were assessed.

Data from this study afford researchers the opportunity to test how well the results of nonrandomized experiments, with proper adjustments, can approximate the results of randomized experiments. While the study design has its own limitations in terms of generalizability, its strength lies in the absence of conditions that might otherwise confound such comparisons. Aside from the assignment to type of experiment (randomized/nonrandomized) and the type of training (mathematics/vocabulary) assigned

or self-selected, all other features of the study were held constant. For example, students in the randomly and non-randomly formed groups received their vocabulary or mathematics training at the same sessions and were always tested in the same way at the same time.

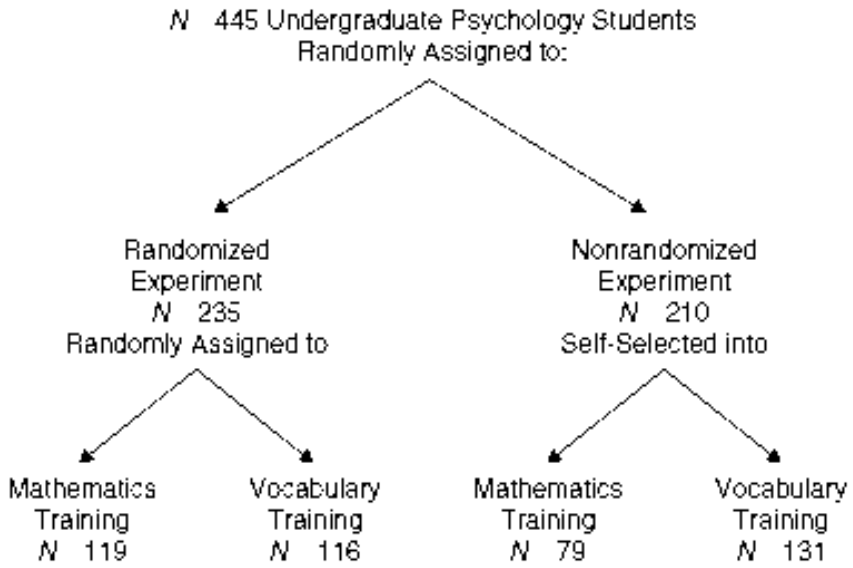

Figure 2: Overall design of the four-arm within-study comparison

It is important to stress that we used data only from the nonrandomized arms of the study (210 students) to estimate the effects of mathematics training on mathematics scores and vocabulary training on vocabulary scores. The same treatment effects estimated from the randomized experiment (235 students) data were treated as a gold standard against which to compare the results of the semi-parametric imputation but otherwise played no role in our analyses.

Careful attention was given to the collection of baseline variables that might predict a student's choice of training and/or post-treatment mathematics and vocabulary scores. The full set of measured covariates includes: Pre-treatment mathematics and vocabulary scores, sex, age, race (black/white/other), marital status, father's and mother's degree of education, parents' annual income, high school GPA, ACT, college GPA, college credit hours, major field of study, number of prior mathematics courses, mathematics anxiety rating scale, liking of mathematics, liking of literature, preference for literature over mathematics, Beck depression inventory, agreeableness, conscientiousness, emotional stability, extroversion, and openness to experience.

To estimate the treatment effects using data from the nonrandomized experiment, we used AutoImpute to impute the following four potential outcomes:

- The post-treatment mathematics score for the 79 students who received mathematics training if instead they had received vocabulary training;
- The post-treatment vocabulary score for the 79 students who received mathematics training if instead they had received vocabulary training;

- The post-treatment mathematics score for the 131 students who received vocabulary training if instead they had received mathematics training; and
- The post-treatment vocabulary score for the 131 students who received vocabulary training if instead they had received mathematics training.

The AutoImpute runs were set up to impute all four potential outcomes at once, allowing the potential mathematics score to enter the imputation model for the potential vocabulary score (and vice versa). However, we did not allow the potential outcomes given mathematics training to be eligible predictors in the imputation models for the potential outcomes given vocabulary training (nor vice versa). This decision was made on the basis that there are no students with *observed* data both under mathematics training and under vocabulary training. Furthermore, the hope was that any associations between these potential outcomes (e.g., due to innate ability) would be captured through the other covariates in the imputation models. Following Shadish, Clark, and Steiner (2008), two sets of baseline measures were considered as eligible covariates in different AutoImpute runs. One set of covariates included all the baseline measures described above, with the exception of parents' income which was deemed unreliable due to students' lack of exact knowledge about their parents' incomes and a high missing rate. The other covariate set included only "predictors of convenience": Sex, age, marital status, and race.

The first set of AutoImpute runs used Shadish, Clark, and Steiner's pre-imputed[1] version of the study data to facilitate comparison with results presented in their paper. A second set of AutoImpute runs was conducted using the unimputed study data set to examine the ability of our method to account simultaneously for item nonresponse and unobserved potential outcomes.

A "doubly robust" approach was built into our semi-parametric imputation method in the following manner. Each student's propensity to choose mathematics (instead of vocabulary) training was estimated using a logistic regression model. Based on the estimated propensity scores, students were grouped into propensity quintiles. Propensity quintile was then included as an eligible covariate in the imputation models. For the set of AutoImpute runs conducted using the pre-imputed study data, propensity quintile was also used as the soft boundary variable in the procedure's initial hot deck. This was done in the hope of producing a better set of starting values from which the iterative imputation proceeds. It was not possible to use propensity quintile when forming the initial hot-deck cells for the unimputed study data set due to missing values. For these runs, we used age as the soft boundary variable in AutoImpute's initial hot deck because there was complete response to this item.

Treatment effects $\left(\hat{\Delta}\right)$ and standard errors (S.E.) were estimated in two different ways: one using Rubin's multiple imputation approach and the other using the Shao-Sitter approach with 500 bootstrap samples[2] drawn from the 210 students in the nonrandomized experiment (see Section 3).

In addition to varying the set of eligible imputation predictors and presence/absence of missing values in the input data set, we explored the effects of several imputation parameters on the treatment effect estimates. The first of these was the maximum number

[1] Here "pre-imputed" refers to the observed survey items, not potential outcomes.
[2] The bootstrap samples were stratified by the choice of mathematics or vocabulary training.

of times a student could be used as a donor, *donMax*. The second was the maximum number of internal AutoImpute iterations, *numIter*, and the third was the number of multiple imputations, *nMI*. A point to note is that the value of *donMax* required for imputation of potential outcomes will typically be higher than would be used in other settings. For a sample of size *n*, amongst which the smallest treatment group size is $n_{\min}$, the minimum value of *donMax* must be at least $(n - n_{\min}) / n_{\min}$.

5. Results

Tables 1–4 each show the results of applying AutoImpute to the pre-imputed study data set. Tables 1 and 2 present summaries for the effect of mathematics training on mathematics scores – Table 1 is based on results using the full covariate predictor set and Table 2 is based on results using only convenient predictors. The information in Tables 3 and 4 is similarly arranged but relates to the effect of vocabulary training on vocabulary scores. Note that in each table, the statistics in the first row represent the gold standard treatment effect estimated from the randomized experiment data (using covariate adjustment to address chance imbalance between the treatment groups). All other estimates are based on the nonrandomized experiment.

Table 1: Effect of Mathematics Training on Mathematics Scores – estimated using full covariate set and pre-imputed study data set

donMax	*numIter*	*nMI*	*MI* $\hat{\Delta}$ *(S.E.)*	*BT* $\hat{\Delta}$ *(S.E.)*
Randomized with covariate adjustment			*4.01 (0.35)*	
5	5	5	4.10 (0.29)	4.03 (0.51)
20	5	5	4.11 (0.29)	4.02 (0.45)
5	10	5	4.11 (0.29)	3.99 (0.48)
5	5	10	4.08 (0.30)	

Table 2: Effect of Mathematics Training on Mathematics Scores – estimated using convenient predictors and pre-imputed study data set

donMax	*numIter*	*nMI*	*MI* $\hat{\Delta}$ *(S.E.)*	*BT* $\hat{\Delta}$ *(S.E.)*
Randomized with covariate adjustment			*4.01 (0.35)*	
5	5	5	*4.96 (0.42)*	*5.02 (0.53)*
20	5	5	*4.95 (0.37)*	*5.01 (0.51)*
5	10	5	*4.98 (0.37)*	*5.02 (0.52)*
5	5	10	*5.00 (0.38)*	

A comparison between Tables 1 and 2, and between Tables 3 and 4, makes it clear that using the full set of covariates as eligible predictors in the imputation models produces treatment effects that are closer to the gold standard treatment effects estimated from the randomized experiment. Within-table comparisons show that variation in the imputation parameters (donor maximum, number of AutoImpute iterations, and number of multiple imputations) leads to differences in treatment effects that are of much smaller order than the differences induced by the choice of eligible covariates in the imputation models, and that are insignificant in magnitude given the size of the standard errors.

Table 3: Effect of Vocabulary Training on Vocabulary Scores – estimated using full covariate set and pre-imputed study data set

donMax	*numIter*	*nMI*	$MI\,\hat{\Delta}$ *(S.E.)*	$BT\,\hat{\Delta}$ *(S.E.)*
Randomized with covariate adjustment			8.25 (0.37)	
5	5	5	8.28 (0.29)	8.30 (0.49)
20	5	5	8.16 (0.30)	8.24 (0.50)
5	10	5	8.24 (0.31)	8.27 (0.47)
5	5	10	8.19 (0.29)	

Table 4: Effect of Vocabulary Training on Vocabulary Scores – estimated using convenient predictors and pre-imputed study data set

donMax	*numIter*	*nMI*	$MI\,\hat{\Delta}$ *(S.E.)*	$BT\,\hat{\Delta}$ *(S.E.)*
Randomized with covariate adjustment			8.25 (0.37)	
5	5	5	8.77 (0.37)	8.71 (0.50)
20	5	5	8.79 (0.34)	8.68 (0.48)
5	10	5	8.72 (0.36)	8.69 (0.48)
5	5	10	8.74 (0.38)	

In all cases, there is a noticeable difference between the standard errors of the effect estimates produced using the multiple imputation approach and the bootstrap method. While gains in precision due to the averaging over multiple imputations might lead us to expect smaller standard errors using this rather than the bootstrap approach, inspection of the within- and between-imputation variance components revealed that the "improper" nature of the hot-deck imputation routine internal to AutoImpute was the main reason for underestimated variances using the multiple imputation approach. The hot-deck procedure selected donors without replacement and in a manner designed to equalize the number of times each donor was used (to the extent possible). When using semi-parametric imputation for causal inference and multiple imputation for variance estimation, it would be preferable to select donors with replacement and without attempting to ensure that each donor is used approximately the same number of times; this should lead to better estimates of the between-imputation variance. Alternatively, one could adopt an adjustment to the basic hot-deck procedure that makes it "proper" for use with multiple imputation, such as the approximate Bayesian bootstrap (Rubin and Schenker, 1986).

Tables 5 and 6 are similar to Tables 1 and 3 but show the results of applying AutoImpute to the unimputed study data set. No gold standard analogs of the treatment effects were estimated based on the unimputed study data set, however the within-table conclusions to be drawn from these imputation runs mimic those noted above. In comparing the performance of the semi-parametric procedure to impute potential outcomes only versus potential outcomes *and* missing covariate data, there are some differences. It would appear that simultaneous imputation of missing covariates and potential outcomes leads to slightly larger treatment effect estimates for the mathematics outcome, and multiple imputation standard errors that are larger and closer to their bootstrap counterparts. A possible explanation for the latter observation may be that the additional number of items with missing values combined with the sequential and iterative nature of the imputation procedure, led to an increased likelihood of different donor pools for a given student's missing data item across multiple imputations. In other words, the imputation procedure

may be closer to a "proper" imputation method as required for the validity of the multiple imputation variance estimator.

Table 5: Effect of Mathematics Training on Mathematics Scores – estimated using full covariate set and unimputed study data set

donMax	*numIter*	*nMI*	$MI\,\hat{\Delta}$ *(S.E.)*	$BT\,\hat{\Delta}$ *(S.E.)*
5	5	5	4.26 (0.43)	4.13 (0.49)
20	5	5	4.36 (0.35)	4.13 (0.47)
5	10	5	4.22 (0.43)	4.16 (0.48)
5	5	10	4.37 (0.38)	

Table 6: Effect of Vocabulary Training on Vocabulary Scores – estimated using full covariate set and unimputed study data set

donMax	*numIter*	*nMI*	$MI\,\hat{\Delta}$ *(S.E.)*	$BT\,\hat{\Delta}$ *(S.E.)*
5	5	5	8.22 (0.35)	8.20 (0.49)
20	5	5	8.27 (0.29)	8.18 (0.52)
5	10	5	8.33 (0.32)	8.21 (0.52)
5	5	10	8.23 (0.31)	

Finally, Table 7 shows a comparison of treatment effects estimated using AutoImpute and a selection of different analytic methods (see Section 2.2). A comparison between the selection presented here as well as other analytic approaches can be found in Table 1 of Shadish, Clark, and Steiner (2008). Excluding the gold standard estimates in the first row, all methods were applied to the nonrandomized pre-imputed study data set and the full set of baseline covariates. To simplify the presentation, we included effect estimates from the imputation approach for only one of the imputation parameter combinations (*donMax* = 5, *numIter* = 5, and *nMI* = 5). The standard errors shown for the AutoImpute approach are those produced using the Shao-Sitter bootstrap method.

Table 7: Comparison of Treatment Effects by Analytic Method – estimated using full covariate set and pre-imputed study data set

Design and Analytic Method	*Effect of Mathematics Training on Mathematics Scores*			*Effect of Vocabulary Training on Vocabulary Scores*		
	$\hat{\Delta}$	*% Bias Reduction*	*S.E.*	$\hat{\Delta}$	*% Bias Reduction*	*S.E.*
Randomized with covariate adjustment	4.01	—	.35	8.25	—	.37
Nonrandomized with no adjustment	5.01	0%	.55	9.00	0%	.51
Nonrandomized with PS stratification	3.72	71%	.57	8.15	86%	.60
Nonrandomized with PS weighting	3.67	66%	.71	8.22	96%	.66
Nonrandomized with ANCOVA	3.85	84%	.44	8.21	94%	.43
*Nonrandomized with AutoImpute**	4.10	91%	.51	8.28	96%	.49

**donMax* = 5, *numIter* = 5, *nMI* = 5

One measure used to compare the analytic methods is the percentage reduction in the bias of the treatment effect estimate, relative to the of bias of the unadjusted estimate obtained from the nonrandomized experiment. With respect to this measure, the semi-parametric imputation estimator performs as well as (for the vocabulary outcome) and better than (for the mathematics outcome) any of the other alternatives shown in Table 7. In terms of standard errors, the imputation estimator comes second only to the ANCOVA approach, and the same is true when the comparison is based on the root mean-square errors (not shown here). These results suggest that further investigation into the use of AutoImpute for causal inference may be worthwhile.

Overall, our results are consistent with the findings of Shadish, Clark, and Steiner (2008). Namely, that the choice of variables to be used in adjusting the estimate of the effect of training on outcomes is more important than the analytic method of adjustment. We also conclude that semi-parametric imputation of potential outcomes produces acceptable point estimates of treatment effects, at least for the data analyzed in this study. Estimation of valid variances using this approach is more of a challenge. It would appear that a bootstrap method such as that proposed by Shao and Sitter (1996) may be required to compensate for the use of hot-deck imputation. However, AutoImpute offers a highly automated, relatively fast and convenient method for causal inference using survey data, especially when imputation also is required for regular item nonresponse.

6. Discussion and Further Research

We are encouraged by the results of this investigation into the use of semi-parametric imputation for causal inference and anticipate directions for further research. Here we used a real-life example but the observational data set did not include survey weights. A more realistic approach would be to study the performance of the semi-parametric imputation method assuming a complex sample design. To investigate the extent to which the findings reported herein generalize to other data sets, we need to study the performance of AutoImpute for causal inference using simulated data for which the true effects are known. There is also important work to be done in determining the best hot-deck procedure and variance estimator to use in the causal inference setting.

Acknowledgements

The authors thank Graham Kalton, David Morganstein, and Mike Brick of Westat for their support and valuable discussions during the early phases of the research.

References

Brick, J.M., Kalton, G., and Kim, J.K. (2004). Variance estimation with hot deck imputation using a model. *Survey Methodology*, 30, 57–66.

Chen, J. and Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*, 16, 113–141.

Holland, P.W. (1996). Statistics and causal inference. *Journal of the American Statistical Association*, 81, 945–970.

Judkins, D., Krenzke, T., Piesse, A., Fan, Z., and Haung, W.C. (2007). Preservation of skip patterns and covariate structure through semi-parametric whole questionnaire imputation. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 3211–3218.

Judkins, D., Piesse, A., and Krenzke, T. (2008). Multiple semi-parametric imputation. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 48–58.

Krenzke, T. and Judkins, D. (2008). Filling in the blanks: Some guesses are better than others: Illustrating the impact of covariate selection when imputing complex survey items. *CHANCE*, Vol. 21, No. 3, 7–13.

Kim, J.K. and Fuller, W.A. (2004). Inference procedures for hot deck imputation. *Biometrika*, 91, 559–578.

Kim, J.K. and Rao, J.K.N. (2009). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, 96, 917–932.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essays on Principles, Section 9. Translated in *Statistical Science*, 5, 465–480, 1990.

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 21, 85–95.

Rao, J.N.K. and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811–822.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688–701.

Rubin, D.B. (1978a). Bayesian inference for causal effects: the role of randomization. *Annals of Statistics*, 6, 34–58.

Rubin, D.B. (1978b). Multiple imputation in sample surveys – a phenomenological Bayesian approach to nonresponse. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 20–34.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, D.B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable non-response. *Journal of the American Statistical Association*, 81, 366–374.

Särndal, C.E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18, 241–252.

Shadish, W.R., Clark, M.H., and Steiner, P.M. (2008). Can nonrandom experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, Vol. 103, No. 484, Applications and Case Studies, 1334–1343.

Shao, J. and Sitter, R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, Vol. 91, No. 435, Theory and Methods, 1278–1288.

Van Buuren, S. and Oudshoorn, C.G.M. (1999). *Flexible Multivariate Imputation by MICE*. Technical report, TNO Prevention and Health, Leiden.