Listwise Deletion is Evil: What to Do About Missing Data in Political Science

Gary King James Honaker Anne Joseph Kenneth Scheve

Department of Government
Harvard University[1]

August 19, 1998

[1] Littauer Center North Yard, Cambridge Massachusetts 02138. Gary King: King@Harvard.Edu, http://GKing.Harvard.Edu, (617) 495-2027; James Honaker: tercer@latte.harvard.edu, http://data.fas.harvard.edu/gov-dept/graduate/tercer; Anne Joseph: ajoseph@fas.harvard.edu; Kenneth Scheve: kscheve@latte.harvard.edu, http://www.fas.harvard.edu/~scheve/. We thank Tim Colton and Mike Tomz for participating in several of our meetings during the early stages of this project; Chris Achen, Jim Alt, Micah Altman, Mike Alvarez, John Barnard, Neal Beck, Larry Bartels, Ted Brader, Charles Franklin, Rob Van Houweling, Jas Sekhon, Brian Silver, and Chris Winship for helpful discussions; Joe Schafer for a prepublication copy of his extremely useful book; Mike Alvarez, Paul Beck, Adam Berinsky, John Brehm, Tim Colton, Russ Dalton, Jorge Domínguez, Bob Huckfeldt, and Jay McCann for their data; and the National Science Foundation (SBR-932121), the Centers for Disease Control and Prevention (Division of Diabetes Translation), The National Institutes on Aging, the World Health Organization, and the Global Forum for Health Research for research support to Gary King.

Abstract

We propose a remedy to the substantial discrepancy between the way political scientists analyze data with missing values and the recommendations of the statistics community. With a few notable exceptions, statisticians and methodologists have agreed on a widely applicable approach to many missing data problems based on the concept of "multiple imputation," but most researchers in our field and other social sciences still use far inferior methods. Indeed, we demonstrate that the threats to validity from current missing data practices rival the biases from the much better known omitted variable problem. As it turns out, this discrepancy is not entirely our fault, as the computational algorithms used to apply the best multiple imputation models have been slow, difficult to implement, impossible to run with existing commercial statistical packages, and demanding of considerable expertise on the part of the user (even experts disagree on how to use them).

In this paper, we adapt an existing algorithm, and use it to implement a general-purpose, multiple imputation model for missing data. This algorithm is between 65 and 726 times faster than the leading method recommended in the statistics literature and is very easy to use. We also quantify the considerable risks of current political science missing data practices, illustrate how to use the new procedure, and demonstrate the advantages of our approach to multiple imputation through simulated data as well as via replications of existing research. We also offer easy-to-use public domain software that implements our approach.

1 Introduction

On average, about half of the respondents who participate in sample surveys do not give answers to one or more questions analyzed in the average survey-based political science article. Almost all analysts contaminate their data at least partially by making up some answers for these respondents (such as by coding "don't know" on party identification questions as "independent"), and approximately 94% use listwise deletion to eliminate entire observations (losing about one-third of their data on average) when any one variable remains missing after the first procedure.[1] Of course, similar problems with missing data occur in non-survey research as well.

In this paper, we address the discrepancy between the treatment of missing data in political science and the well-developed body of statistical theory that recommends against precisely what we do. Even if the answers we make up for nonresponding respondents are right on average, the procedure considerably overestimates the certainty with which we know those answers. Consequently, estimated standard errors will be too small. Listwise deletion discards a third of cases on average, both deleting the few nonresponses as well as the many responses in the deleted cases. This discards too many babies with just a bit of bathwater. The result is a vast waste of valuable information at best and severe selection bias at worst.

Some researchers are able to avoid the problems missing data can cause by using sophisticated statistical models optimized for their particular applications (such as censoring or truncation models; see Section 4). When possible, it is best to adapt one's statistical model specially to deal with missing data, as suggested by the two superb political science books on the subject (Achen, 1986; Brehm, 1993). Unfortunately, doing so in some situations puts heavy burdens on the investigator since optimal models for missing data are highly specialized and so often require unfamiliar methods that differ with each application and may not be programmed in standard statistical software packages.[2]

Our complementary approach is to try to raise the floor on the quality of widely applicable and easy-to-use methods for missing data. We hope to change the default method of coping with missing data in political science, from making up answers in combination with listwise deletion to another method based on the concept of "multiple imputation" that is nearly as easy to use but avoids the statistical problems of current practices (Rubin, 1977). Multiple imputation methods have been around for about two decades, and are

[1] The numbers in this paragraph come from our content analysis of the last five years (1993–97) of the *American Political Science Review*, the *American Journal of Political Science*, and the *British Journal of Political Science*. In these articles, 203 scholarly analyses — 24% of all articles and about half of all quantitative articles — use some form of survey analysis, and 176 of these were mass rather than elite surveys. Only 19% of authors were explicit about how they dealt with missing values; by also asking investigators, looking up codebooks, checking computer programs, or making educated guesses based on partial information provided, we were able to gather sufficient information in 77% of the articles. The situation surely is not better in the articles without adequate reporting, and so both missing data practices and reporting problems are serious concerns that need to be addressed. Our more casual examinations of other journals in political science and other social sciences do not reveal any obvious differences from our results here.

[2] This paper is about *item nonresponse* — when respondents answer some questions and not others (or in general when scattered individual cells in a data matrix are missing). A related issue is *unit nonresponse* — when some of the chosen sample cannot be located or refuse to be interviewed. Brehm (1993) and Bartels (1998) demonstrate that, with some interesting exceptions, the types of unit nonresponse common in political science data sets do not introduce much bias in our analyses. Globetti (1997) and Sherman (1998) show that item nonresponse is a comparatively more serious issue in our field. The many other types of missing data can often be seen as a combination of item and unit nonresponse. Some examples include entire variables missing from one of a series of cross-sectional surveys (Franklin, 1989; Gelman, King, and Liu, 1998), matrix sampling (Raghunathan and Grizzle, 1995), panel attrition, etc.

now the choice of most statisticians at least in principle, but they have not made it into the toolbox of more than a few statisticians or social scientists. The problem is only in part a lack of information. A bigger issue is that although this method is easy to use in theory, it requires in practice computational algorithms that can take many hours or days to run and cannot be fully automated. For example, because they rely on concepts of stochastic (rather than deterministic) convergence, knowing when the iterations are complete and the program should be stopped is still something of an art form about which there is little consensus among experts in the field. For these reasons and others, to our knowledge no commercial statistical software packages include a correct implementation of multiple imputation.[3]

In addition to formalizing the risks of current approaches to missing data in our field, demonstrating how listwise deletion is an inferential problem of comparable magnitude to the much better known omitted variable bias, and showing political scientists how to use better methods, we adapt an existing algorithm and apply it to this problem. This algorithm runs about 65 to 726 times faster for the same imputation model as the leading algorithm used by scholars in the missing data literature, handles more variables effectively, does not rely on stochastic convergence, produces statistically independent imputations, and is easily automated. We plan to release the software we developed that implements our method, although the approach should also be easy for commercial software companies to include in their products as well. This should make it relatively easy for researchers to substitute multiple imputation methods for existing practices, and then to continue to use whatever statistical method they would have if all their data were observed. Methods can be designed in the context of specific data sets to outperform those we discuss, but often at the cost of additional time spent by researchers learning or developing new models (and in some cases a loss of robustness). Our goal is to improve statistical practice, in practice.

We begin with a review of three specific assumptions one can make about missing data in Section 2. Then in Section 3, we demonstrate analytically the severe disadvantages of listwise deletion, problems that exist even under the rosiest of possible assumptions (with mathematical details set aside in Appendix A). Section 4 summarizes some available methods of analyzing data with missing values, and Section 5 introduces a statistical model to create imputations. Algorithms to implement multiple imputation models are discussed in Section 6. In Section 7, we provide systematic Monte Carlo evidence that shows how well our method compares with standard approaches to missing data in political science, and how it is equivalent to the standard approach now used in statistics except that it runs in a very small fraction of the time. Section 8 then reports on several replications of prior research to show how assumptions about and methods for missing data can affect our conclusions about government and politics. Section 9 concludes.

2 Assumptions about Missingness

To determine when different methods are applicable, we outline three possible assumptions about the process by which data can become missing. To define these mechanisms, first let D denote the $\underline{D}$ata matrix, which includes the dependent variable Y and explanatory variables X, and with rows as observations: $D = \{Y, X\}$. If D were entirely observed,

[3] Public domain software accompanying Schafer's (1997) superb book implements monotone data augmentation (Rubin and Schafer, 1990; Liu, Wong, and Kong, 1994), the best available approach presently. The commercial programs Solas and SPlus have also promised implementations. SPSS recently released a missing data module that allows several types of imputation, but none of the options properly represent uncertainty.

Assumption	Acronym	You can predict M with:
Missing Completely At Random	MCAR	—
Missing At Random	MAR	D_{obs}
Nonignorable	NI	D_{obs} & D_{mis}

Table 1: *Three Possible Missingness Assumptions*

we would use some standard statistical method to analyze it and could ignore this paper. In practice, of course, some elements of D are missing. We define M as a missingness indicator matrix with the same dimensions as D but with a 1 in each entry for which the corresponding entry in D is observed and a 0 when the corresponding element of D is missing. Elements of D for which the corresponding entry in M is 0 are unobserved but do "exist" in a specific metaphysical sense. For example, everyone has a (positive or negative) income even if some respondents prefer not to share it with a survey researcher. However, "I don't know" given in response to questions about the national helium reserves or the job performance of the Secretary of Interior probably does not mean the respondent is hiding something! We focus on missing data for which actual data exist but are unobserved, although imputing values that the respondent really does not know can be of interest in specific applications, such as finding out how people would vote if they were more informed (Bartels, 1996). Finally, let D_{obs} and D_{mis} denote elements of D that are observed and missing, respectively, so $D = \{D_{\text{obs}}, D_{\text{mis}}\}$.

Unfortunately, standard terminology describing possible missingness assumptions is unintuitive to say the least. We try to clarify with Table 1. The key is that each missingness process can be characterized according to our ability to predict the values of M (i.e., which values of D will be missing). For example, the missing values in processes that are "missing completely at random" (MCAR) cannot be predicted with any information in D, observed or not: $P(M|D) = P(M)$. An example of an MCAR process is one in which respondents decide whether to answer survey questions on the basis of coin flips. Of course, the MCAR assumption rarely applies: if independents are more likely to decline to answer a vote preference or partisan identification question, then the data are not MCAR.

For processes that are "missing at random" (MAR), the probability that a cell value is missing may depend on the observed data, but it may not depend on values of D that are unobserved: $P(M|D) = P(M|D_{\text{obs}})$. For example, if Democratic party identifiers are more likely to refuse to answer the vote choice question, then the process is MAR so long as party identification is a question in the survey at least some answered. Similarly, if those planning to vote for Democrats do not answer the vote choice question as frequently as those planning to vote for Republicans, the process is not MCAR but it would be MAR if this difference could be predicted with any other variables in the data set (such as ideology, issue positions, income, education, etc.). (The prediction required is not causal, and so for example, the vote could be used whether or not the vote causes or is caused by party ID.) To an extent then, the analyst, rather than the world that generates the data, controls the degree to which the MAR assumption fits. For example, MAR assumptions can be made more applicable and more powerful by including more variables in the imputation process to help predict the pattern of missingness. Finally, in practice all statistical methods that implement versions of MAR are conditional on the specific model chosen to make the imputations, and so will only approximate the MAR ideal of using "all" information in the observed data.

Finally, if the probability that a cell is missing depends in part on the unobserved value of the missing response, the process is said to be nonignorable (NI): $P(M|D)$ does not simplify. An example of such a process is when high-income people are more likely to refuse to answer survey questions about income *and* when other variables in the data set cannot predict which respondents have high income.

3 How Bad is Listwise Deletion?

At best (when MCAR holds) listwise deletion discards considerable information for respondents who answered some but not all questions in a survey. At worst, the practice introduces severe bias into substantive results. An intermediate case is where MCAR holds only for Y given X in a regression-type procedure (i.e., where the analysis model is conditional on X), in which case listwise deletion is inefficient but can be unbiased when the functional form is known to hold exactly. That is, so long as missingness is not a function of $Y|X$, even with MAR or NI missingness for X (independent of Y), listwise deletion can produce consistent results (see Winship and Radbill, 1994). The problem is that much of the characteristic robustness of regression analysis to small amounts of measurement error, nonlinearity, etc. are lost in this instance and, with typical social science data, bias becomes likely.

For most applications, the worst case would seem to apply. That is, whenever the probability that a cell in a data matrix is missing can be predicted, the MCAR assumption, on which listwise deletion is based, is violated. So listwise deletion can bias our conclusions if those who think of themselves as "Independents" are less likely to respond to a party ID question, or if more educated people are more likely to answer issue opinion questions, or if less knowledgeable voters are less likely to reveal their voting preferences, or if wealthy people are more reticent about discussing their income, or when any relationship exists between the probability of missingness and anything else. These patterns might each be MAR or nonignorable, but they are not MCAR. Listwise deletion can result in drastically changed magnitudes or incorrect signs of the estimates of causal effects or descriptive inferences (Anderson et al., 1983). Listwise deletion will not always have such harmful effects; sometimes the fraction of missing observations will be small, and sometimes the assumptions will hold sufficiently well so that the bias is not large. Examples can easily be generated when MCAR is violated with bias of any size or direction. There is little doubt that the entire range could be found in the existing stock of political science publications.

In the remainder of this section, we quantify how harmful listwise deletion is *at best*, that is assuming MCAR holds and no bias exists. As we demonstrate here, even this best-case scenario is problematic.

Suppose we were interested in estimating the causal effect of X_1 on Y, which we label β_1, and for simplicity suppose that neither variable has any missing data. A naive approach in this situation might be to regress Y on X_1, but most scholars in this situation also plan to control for a list of potential confounding influences, which is a set of variables we label X_2. Scholars who are very careful, hard-working, and understand the process of getting articles accepted at journals typically collect a long list of variables to include in X_2. That is, as critics, we use omitted variables as the first line of attack and as authors we know that controlling for more variables helps protect ourselves from potential criticisms.

The goal then is to estimate β_1 in the least squares regression $E(Y) = X_1\beta_1 + X_2\beta_2$. If X_2 contains no missing data, then even if X_2 meets the rules for causing omitted variable bias (i.e., if the variables in X_2 are correlated with and causally prior to X_1 and affect Y), omitting it is still sometimes best. That is, including these variables will reduce bias, but they can also increase the variance of the estimate of β_1 (since by estimating additional parameters, we put more demands on the data). Thus, as is well known, the mean square error (a combination of bias and variance) may in some cases increase by including a control variable (see Goldberger, 1991: 256). Fortunately, since we typically have a large number of observations, adding an extra variable does not usually do much harm so long as it does not introduce substantial collinearity. As a result, we often make the reasonable decision to ignore this effect and include X_2 in the regression.

However, the same tradeoff between bias and variance looms much larger in the presence of missing data. Missing data will normally be present in Y, X_1 and X_2, but suppose for now that there is MCAR item nonresponse only in λ fraction of the n observations in X_2. Ideally, we would observe all of X_2 (i.e., $\lambda = 0$) and estimate β_1 with the complete data regression:

Infeasible Estimator Regress Y on X_1 and a fully observed X_2, and use the coefficient on X_1, which we denote b_1^I.

In contrast, when missing data exists ($0 < \lambda < 1$), most political scientists have only two estimators in their tool-box:

Omitted Variable Estimator Omit X_2 and estimate β_1 by regressing Y on X_1, which we denote b_1^O.

Listwise Deletion Estimator Perform listwise deletion on Y, X_1, and X_2, and then estimate the vector β_1 as the coefficient on X_1 when regressing Y on X_1 and X_2, which we denote b_1^L.

The omitted variable estimator risks bias and the listwise deletion estimator risks inefficiency. (We have ruled out by assumption the possibility that the listwise deletion estimator also introduces bias. In most cases, the MCAR assumption does not hold and this estimator is potentially even more problematic.)

Presumably because the risks of omitted variable bias are better known than the risks of listwise deletion, virtually every political scientist when confronted with this choice opts for the listwise deletion estimator. We quantify these risks with a formal proof in Appendix A, and discuss the results here. We first derive the difference in the mean square error between the two estimators, averaging over both the usual sampling uncertainty and also over the sampling randomness due to the λ fraction of data being MCAR. If $\mathrm{MSE}(a)$ is the mean square error for estimator a, then the difference $\mathrm{MSE}(b_1^L) - \mathrm{MSE}(b_1^O)$ is how we assess which method is better. When this difference is positive, the omitted variable estimator (b_1^O) has lower mean square error and is therefore better than the listwise deletion estimator (b_1^L); when it is negative, the listwise deletion estimator is better. The problem for how political science data analysis is practiced is that this difference is often positive and large.

We need to understand when this mean square error difference will take on varying signs and magnitudes. The actual difference is a somewhat complicated expression that turns out to have a very intuitive meaning:

$$\mathrm{MSE}(b_1^L) - \mathrm{MSE}(b_1^O) = \frac{\lambda}{1-\lambda}\mathrm{V}(b_1^I) + F[\mathrm{V}(b_2^I) - \beta_2\beta_2']F' \tag{1}$$

The second term on the right side of Equation 1 is the well-known tradeoff between bias and variance when no data are missing (where F are regression coefficients of X_2 on X_1, and b_2^I is the coefficient on X_2 in the infeasible estimator). The new result here is thus the first term, which is the extra mean square error due to listwise deletion. This first term is always positive and thus causes the comparison between the two estimators to tilt further away from listwise deletion. As we would expect, the degree of tilt gets larger as the fraction of missing data (λ) grows.

For a more intuitive understanding of the first term, we can estimate the average λ value in political science with the data from our content analysis. This calculation indicates that slightly under one-third of the observations are lost when listwise deletion is used to cope with item nonresponse in political science articles (this loss occurs after making up values for some variables). Because of the tradeoff between bias and variance, those who work harder to fend off more possible alternative explanations will have more control variables and consequently larger fractions of observations lost; those who are lucky to find data with few missing values will have lower values of λ. The average fraction of observations lost in the papers and posters at the 1997 annual meeting of the Society for Political Methodology was well over 50%, and in some cases over 90%. Since in practice scholars frequently drop some variables to avoid extreme cases of missingness, the "right" value of λ for our purposes (the fraction of observations deleted listwise) is larger than the observed fraction. To understand the result in Equation 1, we let $\lambda = 1/2$, and (although the square root of a sum is not the sum of square roots) take the square root of the first term to put it the interpretable units of the average degree of error. Thus,

$$\sqrt{\frac{\lambda}{1-\lambda}V(b_1^I)} = \sqrt{\frac{0.5}{1-0.5}}\mathrm{SE}(b_1^I) = \mathrm{SE}(b_1^I) \tag{2}$$

where SE stands for <u>S</u>tandard <u>E</u>rror.

What the result in Equation 2 means is that *the point estimate in the average political science article is about a standard error farther away from the truth because of listwise deletion* (as compared to omitting X_2 entirely). The point estimates in some articles will be too high, and in others too low, but "a standard error farther from the truth" gives us a sense of how much farther off our estimates are on average, given MCAR. This is a remarkable amount of error, as it is half of the distance from no effect to what we often refer to as a "statistically significant" coefficient (i.e., two standard errors from zero).[4] Of course, we use the standard error here as a metric to abstract across data sets with very different substantive meanings, and so in any one application the meaning of the expression depends on how large the standard error is relative to meaningful changes in the variables. And this relative size in large part depends on the original sample size and cases lost to listwise deletion. Omitted variable bias, in contrast, does not diminish with a larger sample size. Nonetheless, although political scientists rarely choose it, except in extreme cases, omitted variable bias will often be a preferable fate, if only it or the evils of listwise deletion are the options. In practice, of course, one cannot avoid missing value problems since it affects all our data to some degree rather than only potential control variables.

Even if random, this is a lot of unnecessary error being added to the estimates of our statistical quantities of interest. Of course, because this result relies on the optimistic MCAR assumption, the degree of error will be more than a standard error in most real analyses, and it will not be in random directions (Globetti, 1997; Sherman, 1998). The actual case, rather than this "best" case, would seem to be a surprisingly serious problem.

[4] This is one of the infeasible estimator's standard errors, which is equivalent to 71% of the listwise deletion estimator's standard error (or in general, $\sqrt{\lambda} \times \mathrm{SE}(b_1^L)$). The calculated standard errors will be correct under MCAR but of course are larger then those for better estimators given the same data, and wrong if MCAR doesn't hold.

If required to make the insidious choice between low bias and low variance, the right decision will often be the one rarely made in our discipline, omitting the control variable rather than including it and performing listwise deletion. However, with better methods this choice need not be made, and much of the inefficiency and bias can be avoided.

4 Existing Approaches We Do Not Use

Available methods for analyzing data with item nonresponse can be divided into *application-specific* approaches (which are statistically optimal, but hard to use) and *general purpose* approaches (which are easy to use and more widely applicable, but statistically inadequate); we discuss these in Sections 4.1 and 4.2, respectively. In Section 5, we consider *multiple imputation*, which is statistically valid and, with our new algorithm, easy to use.[5]

4.1 Application-Specific Approaches

Application-specific approaches often come from economics or biostatistics and usually assume MAR or NI. The most common examples are models for selection bias, such as truncation or censoring (Achen, 1986; Brehm, 1993; Heckman, 1976; Amemiya, 1985: chapter 10; King, 1989: chapter 7; Winship and Mare, 1992). This approach explicitly models missingness M simultaneously with the outcome D. Such models have the advantage of including the maximum information in the estimation process. Indeed, NI problems are almost exclusively modeled with application-specific approaches. For regression-type problems where the model is conditional on X, this includes methods to avoid bias due to selection on $Y|X$. Unfortunately, almost all application-specific models deal only with missingness in or related to Y, and cannot handle missingness scattered throughout D.

If the assumptions apply, application-specific approaches are maximally efficient. However, inferences about the quantities of interest from these models tend to be fairly sensitive to small changes in specification (Stolzenberg and Relles, 1990). Moreover, no single application-specific model works well across applications; instead, a different model must be used for each type of application. As a result, when applied to new types of data sets, application-specific methods are most likely to be used by those willing to devote more time to methodological matters.[6]

More formally, application-specific approaches begin by modeling D and M jointly, and then factor the joint distribution into the marginal and conditional densities. One way to do this produces selection models:

$$P(D, M|\theta, \gamma) = P(D|\theta)P(M|D, \gamma) \tag{3}$$

where $P(D|\theta)$ is the likelihood function we would use if no data were missing (a function of θ, the parameter of interest), and $P(M|D, \gamma)$ is the distribution of the process by which some data become missing (a function of γ, which is not normally of interest). Once both distributions are specified, as they must be for these models, we can integrate over the missing data in Equation 3 to yield the likelihood function:

$$P(D_{\text{obs}}, M|\theta, \gamma) = \int P(D|\theta)P(M|D, \gamma)dD_{\text{mis}} \tag{4}$$

[5] The most useful modern work on the subject related to our approach is Schafer (1997), which we rely on frequently. Some canonical references in this large literature are Little and Rubin (1987) and Rubin (1987). Landerman, Land, and Pieper (in press) is also very helpful. See also Rubin (1996).

[6] Some political science uses of application-specific methods for missing data include Achen (1986), Berinksy (1997), Brehm (1993), Herron (1998), Katz and King (1997), King et al. (1989), Skalaban (1992), and Timpone (1998).

where the integral is over all elements of D_{mis} and stands for summation for discrete distributions. If one is willing to assume MAR (i.e., D and M are stochastically independent), then the likelihood in Equation 4 simplifies to:

$$P(D_{\text{obs}}, M|\theta, \gamma) = P(D_{\text{obs}}|\theta)P(M|D_{\text{obs}}, \gamma) \tag{5}$$

which is easier to express and maximize directly. If, in addition, θ and γ are parametrically independent, the model is ignorable, in which case the likelihood factors and only $P(D_{\text{obs}}|\theta)$ need be computed. (Our presentation in this paragraph is closest to Schafer (1997); the original definitions come from Rubin (1977, 1987) and Little and Rubin (1987).)

Application-specific approaches have some obvious difficulties. They require specifying $P(M|D, \gamma)$, a distribution in which scholars often (although not always) have no special interest and about which they possess little knowledge. Even if this distribution can be reasonably specified, evaluating the integral in Equation 4 can be difficult or impossible. Even with MAR and ignorablity assumptions, maximizing $P(D_{\text{obs}}|\theta)$ can be computationally complicated given the non-rectangular nature of the data. Computational problems should not distract from the point that, when they can be overcome, application-specific models are normally optimal in theory, even though they do make data analyses that treat missing data responsibly somewhat difficult to pull off.

4.2 General Purpose Approaches

General purpose approaches are far easier to use. The basic idea is to impute ("fill in") or delete the missing values and then analyze the resulting data set with any of the standard methods that assume the absence of missing data. Missingness is treated as a problem that, once fixed, can be ignored, and then all the standard methods, and existing statistical software, can be used without additional mathematical derivation or computer programming. General purpose methods other than listwise deletion include mean substitution (imputing the univariate mean of the observed observations), best guess imputation (quite common in political science), imputing a zero and then adding an additional dummy variable to control for the imputed value, pairwise deletion (which really only applies to covariance-based models), and hot deck imputation (imputing a complete observation that is similar in as many observed ways as possible to the observation that has a missing value). Under MAR (or NI), all of these methods are biased or inefficient, except in special cases. Most of those which impute give standard errors that are too small because they essentially "lie" to the computer program, telling it that we know the imputed values with as much certainty as we do the observed values.

If only one variable has missing data, one possibility is to run a regression (with listwise deletion) to estimate the relationship among the variables and then to use the regression's predicted values, to impute the missing values. A more sophisticated version of this procedure can also be used iteratively to fill in datasets where many variables are missing. This procedure is not biased for some quantities of interest even assuming MAR, since it conditions on the observed data. However, since the missing data are imputed on the regression line as if there were no error, the method produces standard errors that are too small and generates biased estimates of quantities of interest that require more than the conditional mean (such as $\Pr(Y > 7)$). Assuming that a statistical relationship is imperfect when observed but perfect when unobserved is optimistic to say the least! A better method, but one which still gives the wrong standard errors, is to impute based on this regression procedure using the predicted value, but adding a draw from the error term (see Section 6.2). Except in special situations, these general purpose methods do not work.

5 Multiple Imputation Approaches

We now provide a general definition of multiple imputation (Section 5.1), our specific model for generating the imputations (Section 5.2), and a list of clarifications and common misconceptions (Section 5.3). We also elaborate on what can go wrong when using our procedure and how to avoid such problems (Section 5.4), and then conclude with the best case that can be made for listwise deletion in comparison to our method (Section 5.5).

5.1 A General Definition

Our recommended approach, a version of *multiple imputation*, involves imputing m values for each missing item and creating m completed data sets. Across these completed data sets, the observed values are the same, but the missing values are filled in with different imputations to reflect our uncertainty levels. That is, for missing cells our model predicts well, the variation across the m imputations will be small; for other cases, the variation may be larger, or even asymmetric or bimodal, in order to reflect whatever knowledge and level of certainty is available about the missing information. The analyst then applies whatever statistical method would normally be used when there are no missing values to each of the m data sets, and uses a simple procedure we now describe to combine the results. (As we explain below, m can be as small as 3 or 4.)

To average the results from the m completed data sets, we first decide on a quantity of interest we want to compute, such as a univariate mean, regression coefficient, predicted probability, or first difference. The overall point estimate $\bar{q}$ is the average of the m separate estimates, q_j ($j = 1, \dots, m$):

$$\bar{q} = \frac{1}{m}\sum_{j=1}^{m} q_j \tag{6}$$

The variance of the point estimate is the average of the estimated variances from *within* each completed data set, plus the sample variance in the point estimates *across* the data sets (multipled by a factor that corrects for bias because $m < \infty$). Let $\hat{V}_{q_j}$ denote the estimated variance (squared standard error) of q_j from data set j, and $S_q^2 = \sum_{j=1}^{m}(q_j - \bar{q})^2/(m-1)$ be the sample variance across the m point estimates. Then the standard error of the multiple imputation point estimate is the square root of

$$\mathrm{SE}(q)^2 = \frac{1}{m}\sum_{j=1}^{m} \hat{V}_{q_j} + S_q^2\,(1 + 1/m) \tag{7}$$

Once the imputed data sets are created, analysts can conveniently use their familiar statistical models and computer programs (run m times). If, instead of point estimates and standard errors, simulations of q are desired, we create $1/m$th the needed number of simulations from each completed data set (following the usual procedures; see King, Tomz, and Wittenberg, 1998) and combine them into one set of simulations.

Most of the statistical procedures used to create multiple imputations assume that the data are MAR, conditional on the imputation model. Proponents claim that in practice most data sets include sufficient information so that the additional outside information in an NI model would not add much, and may be outweighed by the costs of non-robustness and difficulty of use. Whether or not this is true in any application, the advantages in terms of ease of use makes multiple imputation methods an attractive option for a wide range of potential applications. The MAR assumption can also be made more realistic by including more informative variables and information in the imputation process.

5.2 A Model for Imputations

Implementing multiple imputation requires a statistical model from which to compute the m imputations for each missing value in a data set. The only purpose of this model is to create predictions for the distribution of each of the missing values. Thus, unlike many political science statistical applications, the imputation stage of the statistical analysis is only concerned with prediction, and not with causal explanation, parameter interpretation, or anything else.

One model that has proven to be useful for missing data problems in a surprisingly wide variety of data types assumes that the variables are jointly multivariate normal. This model is obviously an approximation, as few data sets have variables that are all continuous and unbounded, much less multivariate normal. Yet researchers have frequently found it to work as well as much more complicated alternatives specially designed for categorical or mixed data (Ezzati-Rice et al., 1995; Graham and Schafer, in press; Schafer and Olsen, 1998; Schafer, 1997; Rubin and Schenker, 1986; Schafer et al., 1996). For our purposes, if there exists information in observed data that can be used to predict the missing data, multiple imputations from this normal model will almost always dominate the current practice of making up values combined with listwise deletion.[7] We therefore only discuss this multivariate normal model, although the algorithms we discuss in Section 6 may also work for some of the more specialized models as well. Transformations and other procedures can be used to improve the fit of this model to the data (about which more in Section 5.4).

For observation i ($i = 1, \dots, n$), let D_i denote the vector of values of the p (dependent Y_i and explanatory X_i) variables which if all observed would be distributed normally, with mean vector μ and variance matrix Σ. The non-zero off-diagonal elements of Σ allow the variables within D to depend on one another. The likelihood function for complete data is then:

$$L(\mu, \Sigma | D) \propto \prod_{i=1}^{n} N(D_i | \mu, \Sigma) \tag{8}$$

By assuming the data are MAR, we form the observed data likelihood. The procedure is exactly as in Equations 4 and 5, where with the addition of a prior this likelihood is proportional to $P(D_{\text{obs}}|\theta)$. We first denote $D_{i,\text{obs}}$ as the observed elements of row i of D, and μ_{obs} and Σ_{obs} as the corresponding subvector and submatrix of μ and Σ, respectively. Then, because the marginal densities of the multivariate normal are also normal, the observed data likelihood is

$$L(\mu, \Sigma | D_{\text{obs}}) \propto \prod_{i=1}^{n} N(D_{i,\text{obs}} | \mu_{\text{obs}}, \Sigma_{\text{obs}}) \tag{9}$$

The changing compositions of $D_{i,\text{obs}}$, μ_{obs}, and Σ_{obs} over i make this a complicated expression to evaluate, although for clarity of presentation, we have omitted several computational conveniences that can help (see Schafer, 1997: 16).[8]

The multivariate normal specification implies that the missing values are imputed linearly. Thus, we create an imputed value the way we would normally simulate from a

[7] Most variables based on political science surveys are 4–7 category ordinal variables that are reasonably well approximated by the normal model, at least for the purpose of making imputations.

[8] Since the number of parameters, $p+p(p+1)/2$, increases rapidly with the number of variables, adding prior distributions avoids overfitting. Priors also help with convergence and numerical stability for the algorithms discussed in Section 6.

regression. For example let $\tilde{y}_{ij}$ denote a simulated value of observation i for variable j, and let x_i include all variables in D except D_j and those that are missing for observation i. The coefficient β from a regression of y_j on x can be computed directly from μ and Σ using the formulas for computing a conditional distribution from a joint normal and thus contains all available information in the data under this model. Then we use this equation to create an imputation:

$$\tilde{y}_{ij} = x_i\tilde{\beta} + \tilde{\epsilon}_i \tag{10}$$

where $\sim$ is used to indicate a random draw from the appropriate posterior distribution. This means that random draws of $\tilde{y}_{ij}$ are linear functions of the other variables x and include estimation uncertainty, due to not knowing β (i.e., μ and Σ) exactly, and fundamental uncertainty (i.e., since Σ is not a matrix of zeros). If we had an infinite sample, $\tilde{\beta}$ could be replaced with the fixed β, but there would still be uncertainty generated by the world, ϵ_i. In real finite samples, $\tilde{\beta}$ has a non-degenerate posterior distribution that must be estimated in some way. The difficulty in using this model is taking random draws from the posterior distribution of μ and Σ.

The continuous normal imputations from Equation 10 can also be used to generate imputations for categorical variables by rounding off to the nearest valid integer (as recommended by Schafer, 1997). A slightly better procedure is to draw from a multinomial or other appropriate discrete distribution with mean equal to the normal imputation. For example, to impute a 0/1 variable, we would take a single draw from a Bernoulli distribution with mean equal to the imputation (truncated to [0,1] if necessary). That is, we impute a 1 with probability equal to the continuous imputation and 0 otherwise.

5.3 Clarifications and Common Misconceptions

Multiple imputation inferences have been shown to be statistically valid from both a Bayesian and a frequentist perspective (Rubin 1987; Schenker and Welsh 1988; Brownstone 1991; Meng 1994; Rubin 1996; Schafer 1997). Since there is some controversy over the strength and applicability of the assumptions involved from a frequentist perspective, we focus on the far simpler Bayesian version. This version also encompasses the likelihood framework, which covers the vast majority of social science statistical models.

The fundamental (Bayesian or likelihood) result involves approximating the correct posterior distribution or likelihood function $P(Q|D_{\text{obs}})$, which we would get from an optimal application-specific method, with the posterior distribution based on the completed data $P(Q|D_{\text{obs}}, D_{\text{mis}})$, that is filled in with imputations drawn from the posterior of the missing data $P(D_{\text{mis}}|D_{\text{obs}})$. Under MAR, we know that averaging $P(Q|D_{\text{obs}}, D_{\text{mis}})$ over D_{mis} gives exactly $P(Q|D_{\text{obs}})$:

$$P(Q|D_{\text{obs}}) = \int P(Q|D_{\text{obs}}, D_{\text{mis}})P(D_{\text{mis}}|D_{\text{obs}})dD_{\text{mis}} \tag{11}$$

This integral can be approximated as always with simulation: To draw a random value of Q from its posterior $P(Q|D_{\text{obs}})$, draw independent random imputations of D_{mis} from $P(D_{\text{mis}}|D_{\text{obs}})$, and then draw Q conveniently from $P(Q|D_{\text{obs}}, D_{\text{mis}})$, given the imputed D_{mis}. Thus, we can approximate $P(Q|D_{\text{obs}})$ or any point estimate based on it to any degree of accuracy by merely drawing a large enough number of simulations. This shows that if the complete-data estimator is consistent and produces nominal interval coverage equal to actual interval coverage, then multiple imputation based on $m = \infty$ is consistent and its nominal interval coverage equals the actual coverage.

Multiple imputation then becomes feasible when we recognize that the efficiency of estimators based on the procedure increases very quickly in m (see Rubin, 1987 and the citations in Meng, 1994); indeed the relative efficiency of estimators with m as low as 3–10 is nearly the same as with $m = \infty$, unless missingness is exceptionally high. Confidence interval coverage remains correct because Equation 7, used to combine the multiply imputed estimates, includes a factor to adjust for the extra Monte Carlo simulation error due to m being finite.

Multiple imputation becomes much more widely applicable by Meng's (1994) reassuring results that show what happens when the imputation model differs from the analysis model. Indeed, so long as the imputation model includes all the variables (and information) used in the analysis model, no bias is introduced and nominal confidence interval coverage will be at least as great as actual coverage, and equal when the two stages coincide (Fay, 1992; Rubin, 1996). When the imputation model includes more information than the analysis model, multiple imputation is more efficient than even the corresponding "optimal" application-specific method.

Thus, even with a very small m and an imputation model that differs from the analysis model, the convenient multiple imputation procedure gives a good approximation to the optimal posterior distribution or likelihood function, $P(Q|D_{\text{obs}})$. This result alone guarantees valid inferences in theory from multiple imputation. Indeed, deviating from it to focus on partial calculations sometimes leads to misconceptions on the part of researchers. For example, since the imputation model is not a causal model, no assumptions about causal ordering are required in making imputations: using variables that may be designated "dependent" in the analysis phase to impute missing values in variables to be designated "explanatory" generates no endogeneity since the imputations do not change the joint distribution. Similarly, randomness in the missing values in the explanatory variable from the multiple imputations do not cause coefficients to be attenuated (as when induced by random measurement error) because the imputations are being drawn from their posterior and not independently; again, the joint distribution is not being changed. Since the multiple imputation procedure taken as a whole approximates $P(Q|D_{\text{obs}})$, these "intuitions" based on parts of the procedure are invalid (see Schafer, 1997: 105ff).

Because the imputation and analysis stages are separate, many proponents of multiple imputation have argued that the imputations for public use data sets could be created by a central organization, such as the data provider, so that analysts could ignore the missingness problem altogether. This strategy has proven convenient for analysts and can be especially advantageous if the data provider is able to use confidential information in making the imputations that could not otherwise be made available to the analyst. It has proved convenient for those able to hire consultants to make the imputations. Others have not been sold on this idea even if they had the funds because it can obscure data problems that overlap the two stages and can provide a comforting but false illusion to analysts that missingness problems were "solved" by the imputer (in ways analysts may not have access to or knowledge of). The approach is also not feasible for large data sets like the NES because existing computational algorithms cannot reliably handle so many variables, even in theory. Our complementary approach is to make the tools of imputation very easy to use and available directly to researchers to make their own decisions and control their own analyses.

5.4 What Can Go Wrong and What to Do About It

Multiple imputation provides model-based estimates and, like any statistical approach, if the model is wrong, there are circumstances where the procedure will lead one astray. At the most basic level, the point of inference is to learn something about facts we do not observe by using facts we do observe; if the latter have nothing to do with the former, then we can be mislead with any statistical method that assumes otherwise. In the present context, our method assumes that the information in the observed data can be used to predict at least some aspects of the missing data. For an extreme counterexample, consider an issue scale with integer responses 1–7, and what you think is a missing value code of -9. If unbeknownst to you, the -9 is actually not a missing value but rather an extreme point on the same scale, then imputing values for it based on the observed data, and rounding to 1–7, will obviously be biased. However, in these and other extreme examples where bringing more data to bear on the problem makes for worse inferences, listwise deletion will be at least as bad since it makes strong assumptions about the missing data and generally discards far more observed information than our approach has to impute.

The other advantage of multiple imputation is that it is often robust to errors in the imputation model since (as with the otherwise inferior single imputation models) separating the imputation and analysis stages means that errors in the missingness model can have no effect on observed parts of the data set, since they are the same for all m imputations. If a very large fraction of missingness exists in a data set, then multiple imputation will be less robust, but listwise deletion and other methods will normally be worse.

Beyond these general meta-concerns, a key point for practice is that the imputation model should contain at least as much information as the analysis model. The primary way to go wrong with multiple imputation is to include information in the analysis model that is omitted from the imputation model, and the primary fix is to include this information. We provide several examples of this.

The simplest example of the problem is where a variable is excluded from the imputation model but used in the analysis model. In this situation, estimates of the relationship between this variable and others will be biased towards zero. As a general rule, researchers should include in the imputation model all the variables from the analysis model. For additional efficiency, they should also add any other variables in the data set that would help predict the missing values in the original variables.

Indeed, the ability of multiple imputation to include extra variables in the imputation model that are not in the analysis model is a special advantage of this approach over listwise deletion. For example, suppose the chosen analysis model is a regression of Y on X, but the missingness in X depends on variables Z that also affect Y (even after controlling for X). In this case, listwise deletion regression is inconsistent. Including Z in the regression would make the estimates consistent in the very narrow sense of correctly estimating the corresponding population parameters, but these would be the wrong population parameters since we were effectively forced to control for Z. For example, suppose the purpose of the analysis model is to estimate the causal effect of partisan identification X on the vote Y. We would certainly not want to control for voting intention five minutes prior to walking into the voting booth Z, since it is a consequence of party ID and so would incorrectly drive party ID's estimated effect to zero. Yet, voting intention five minutes before the vote would be a powerful predictor of missingness in the vote variable, and so the ability to include it in the imputation stage of a multiple imputation model, and also to leave it out of the analysis model, is a great advantage. In fact, in many applications scholars apply several analysis models to the same data (such as estimating the effect of party ID, while excluding voting intentions, and estimating the effect of voting intentions while including party ID). Despite these different theoretical goals, using different missingness models for the same variables, as listwise deletion effectively requires, is rarely justified. For another example, scholars often choose for an analysis model only one of several very similar issue preference variables from a data set to measure ideology. This is fine for the analysis model, and indeed doing otherwise will often bias causal estimates, but for the imputation model the entire set of issue preference questions should be included since an observed value in one can be especially useful for predicting a missing value in another variable in the same set.

A similar problem with more information in the analysis model than the imputation model occurs if the analysis model specifies a nonlinear relationship, as our imputation model is linear (see Equation 10). There is little problem with the set of nonlinear functional forms typically used in the social sciences (logit, probit, exponential, etc.), since a linear approximation to these forms has been shown to perform very well during imputation, even if not for the analysis model. However, more severe nonlinearity, such as quadratic terms that are the central question being researched, can cause problems if ignored. Fortunately, this is also easy to address. A quadratic form is estimated in an analysis model by including an explanatory variable and its square as separate terms. Omitting the squared term from the imputation model causes the same problems as omitting any other potentially important variable. The solution is easy: include the squared term in the imputation model. The same problem and solution apply to interaction terms (although the imputation procedure would be slightly less efficient if one variable had much more missingness than another).

In addition to informational differences between the imputation and analysis models, researchers using the methods we recommend should also try to meet the distributional assumptions of the imputation model. For the imputation stage, variables should be transformed to make them unbounded and relatively symmetric. For example, budget figures, which are often restricted to be positive and are positively skewed, can be logged to make them approximately normal. Event counts can be made closer to normal by taking the square root, which stabilizes the variance and makes them approximately symmetric. The logistic transformation can be used to make proportions unbounded and symmetrically distributed.

Ordinal variables should be coded to as close to interval scales as information indicates. For example, if categories of a variable measuring the degree of intensity of a dispute are arguing, yelling, punching, and killing, a coding of 1, 2, 3, and 4 would not seem approximately interval. Perhaps 1, 2, 20, and 200 might be closer. Of course, including transformations to fit distributional assumptions, and making ordinal codings more reasonable like this, are called for in any linear model, even if multiple imputation were not used.

5.5 What is the Best Case for Listwise Deletion?

When a researcher follows the rules laid out in this paper, what known conditions would be necessary for listwise deletion to be preferred over our approach?

The answer is that all of the following conditions must hold. (1) Extra variables Z to add to the imputation (but not analysis) stage should be unavailable. Given the typically large number of variables available in most social science surveys, this condition is most likely to occur only with very large analysis models. (2) The analysis model is conditional on X (such as a regression model) and the functional form is known to be correctly specified (so that listwise deletion is consistent and robustness is not lost when applying listwise deletion to data with slight problems of measurement error, endogeneity, nonlinearity, etc.). (3) There is NI selection on X, so that multiple imputation can give incorrect answers, and no Z variables are available that could be used in an imputation stage to fix the problem. We must also know that (4) missingness in X is not a function of Y, and similarly unobserved omitted variables Z that affect Y do not exist. This ensures that the normally substantial advantages of our approach in this instance do not apply. (5) You have enough information about problems with your variables so that you do not trust them to impute the missing values in your X's — or you worry more about using available information to imput the X's than the existence of selection on X as a function of Y in 4, which our approach would correct. Despite 5, (6) you still trust your data enough to want to use them in an analysis model. That is, we somehow know the same variables cannot be used to predict D_{mis} but can be used to estimate quantities of interest based on D_{obs}. Finally, (7) the number of observations left after listwise deletion should be very large so that the normal efficiency loss of listwise deletion does not counter balance (in a mean square error sense, for example) the biases induced by the other conditions. This condition does not hold in most political science surveys given their modest sizes, except perhaps exit polls and some nonsurvey data.

If these seven conditions hold, or if one of the problems listed in Section 5.4 holds and we do nothing about it, our approach can perform poorly and in extreme cases can do even worse than listwise deletion. Researchers should consider whether these conditions might hold in their data. However, we feel this situation — where using more information is worse — is likely to be exceptionally rare. It is indeed difficult to think of a substantively realistic MAR data generation process describing a real data set where using more data and other information with our procedure would mislead but where listwise deletion would work better.

A more interesting and potentially productive question, and one to which methodologists will no doubt continue to devote themselves in coming years, is when selection on $Y|X$ makes it worth the effort to design an application-specific approach.

6 Computational Algorithms

Computing the observed data likelihood in Equation 9, or the corresponding posterior, is a computationally intensive task, and taking random draws from it is infeasible with classical methods. Even maximizing the function with respect to μ and Σ would take an inordinately long time with standard optimization routines. In response to computational difficulties like these, the IP and EM algorithms were devised and subsequently applied to this problem. From the perspective of statisticians, IP is now the gold standard of algorithms for multivariate normal multiple imputations, in large part because they have found it very flexible in its ability to adapt to numerous specialized models. Unfortunately, from the perspective of users, it is slow and hard to use. Since IP is based on Markov Chain Monte Carlo (MCMC) methods, it requires considerable expertise to judge convergence, and there is no firm agreement among experts about this outside of special cases. IP has the additional problem of giving dependent draws, and so we need adaptations because multiple imputation requires that draws be independent. In contrast, EM is a fast algorithm for finding the maximum of the likelihood function. It converges deterministically, but it alone does not solve the problem since we require the entire posterior distribution rather than only the maximum. We outline these algorithms in Sections 6.1 and 6.2, and refer the reader to Schafer (1997) for an extremely clear presentation of the computational details and historical development.

In Sections 6.3 and 6.4, we discuss two additional algorithms, which we call EMs and EMis, respectively. Our recommended procedure, EMis, is quite practical: It gives draws from the same posterior distribution as IP but is considerably faster and, since it does not rely on MCMC methods, there are no convergence or independence difficulties. Both EMs and EMis are made up of standard parts and have been applied to many problems outside of the missing data context. For missing data problems, EMs has been used, and versions of EMis have been used for specialized applications. EMis may also have been used for problems with general patterns of missingness like we are studying, although we have not yet located any (and it is not mentioned in the most recent exposition of practical computational algorithms, Schafer (1997)). In any event, we believe this procedure has the potential for widespread use.

6.1 IP

IP, which stands for Imputation-Posterior, is based on the "data augmentation" algorithm of Tanner and Wong (1987). IP enables us to draw random simulations from the multivariate normal observed data posterior $P(D_{\text{mis}} \mid D_{\text{obs}})$ (see Li, 1988, and Schafer, 1997: 72ff).

The basic idea is that drawing directly from this distribution is difficult, but "augmenting" it by conditioning on additional information becomes easier. Because this additional information must be estimated, the procedure has two steps that are carried out iteratively. First, imputations $\tilde{D}_{\text{mis}}$ are drawn from the conditional predictive distribution of the missing data in what is called the imputation step:

$$\tilde{D}_{\text{mis}} \sim P(D_{\text{mis}} \mid D_{\text{obs}}, \tilde{\mu}, \tilde{\Sigma}) \tag{12}$$

On the first application of Equation 12, guesses are used for the additional information, $\tilde{\mu}$ and $\tilde{\Sigma}$. Then, new values of the parameters μ and Σ are drawn from their posterior distribution, which depends on the observed data and, to make it easier, the present imputed values for the missing data. This is called the posterior step:

$$\tilde{\mu}, \tilde{\Sigma} \sim P(\mu, \Sigma \mid D_{\text{obs}}, \tilde{D}_{\text{mis}}) \tag{13}$$

This procedure is iterated, so that over time draws of D_{mis}, and $\tilde{\mu}$ and $\tilde{\Sigma}$, come increasingly from their actual distributions than from the starting values.

The advantage of IP is that the distributions are exact and so the method does not depend on approximations. However, convergence in distribution is only known to occur as the number of iterations increases asymptotically. The belief is that after a suitably long burn-in period, perhaps recognizable by consulting various diagnostics, convergence will have essentially occurred, after which additional draws can be assumed to come from the posterior distribution. Unfortunately, there is considerable disagreement within the statistics literature on how to assess convergence of this and other MCMC methods (Cowles and Carlin, 1996; Kass et al., 1997).

For multiple imputation problems, we have the additional requirement that the draws we use for imputations must be statistically independent, which is not a characteristic of successive draws from Markov chain methods like IP. Some scholars reduce dependence by using every rth random draw from IP (where r is determined by examining the autocorrelation function of each of the parameters), but Schafer (1997), following Gelman and Rubin (1996), recommends addressing both problems by creating one independent Markov chain for each of the m desired imputations, with starting values drawn randomly from an overdispsersed approximation distribution. The difficulty with taking every rth draw from one chain is the interpretation of autocorrelation functions (requiring analysts of cross-sectional data to be familiar with time series methods); whereas the difficulty of running separate chains is that the run time is increased by a factor of m.

6.2 EM

The Expectation-Maximization algorithm was developed long ago but was formalized and popularized, and convergence was proven by Dempster et al. (1977), who also thought of it in the context of missing data. EM works very much like IP except that random draws from an entire posterior distribution are replaced with deterministic calculations of means. The draw of $\tilde{D}_{\text{mis}}$ in Equation 12 is replaced with the expected (or predicted) value for each missing cell. The random draw of $\tilde{\mu}$ and $\tilde{\Sigma}$ in Equation 13 is replaced with the maximum posterior estimate. In simple cases, this involves running regressions to estimate β, imputing the missing values with a predicted value, restimated β, and iterating until convergence. The result is that both the imputations and the parameters computed are the single (maximum posterior) values, rather than a whole distribution.

The advantages of EM are that it is fast, it converges deterministically, and the objective function increases with every iteration. Like every numerical optimization algorithm, EM can sometimes settle on a local rather than global maximum, and for some problems convergence is slow, although these do not seem like insurmountable problems in the kinds of data we have in political science. The bigger disadvantage of EM is that it only yields maximum values of the parameters, rather than draws from the entire distribution. Schafer (1997) uses EM to produce multiple imputations by acting as if the maximum likelihood estimates of μ and Σ are known with certainty. This means that estimation uncertainty is ignored but the fundamental variability is included in the imputations (random draws of $\tilde{\beta}$ in Equation 10 are replaced by the maximum posterior estimate). EM for multiple imputation works reasonably well in some instances, but ignoring estimation uncertainty means its standard errors are generally biased downwards.

6.3 EMs

Our strategy is to begin with EM and to add back in estimation uncertainty so we get draws from the correct posterior distribution of D_{mis}. The problem is that the posterior distribution of μ and Σ is not easy to draw from. We solve this problem in two different ways. In this section, we use the asymptotic approximation (e.g., Tanner, 1996: 54–59), which we find works as expected — well in large data sets due to the central limit theorem and poorly in small ones.

To create multiple imputations with this method, which we denote EMs (EM with sampling), we first run EM to find the maximum posterior estimates of the parameters, $\hat{\theta} = \text{vec}(\hat{\mu}, \hat{\Sigma})$ (where the $\text{vec}(\cdot)$ operator stacks the unique elements of its argument). Then we compute the variance matrix of the parameters, $\text{V}(\hat{\theta})$.[9] Then we draw a simulated θ from a normal distribution with mean $\hat{\theta}$ and variance $\text{V}(\hat{\theta})$. From this, we compute $\tilde{\beta}$ deterministically, simulate $\tilde{\epsilon}$ from the normal distribution, and substitute these values into Equation 10 to generate an imputation. The entire procedure after the EM step is repeated m times to produce the necessary imputations.

[9] There are several methods of computing the variance matrix. We tried several but generally use the outer product gradient method for speed. Other options are the hessian, which is asymptotically the same and supposedly somewhat more robust in real problems; "supplemented EM" which is somewhat more numerically stable but not faster; and White's "sandwich" estimator which is more robust but slower.

The advantage of this method is that it is very fast, produces independent imputations, does not require stochastic convergence techniques, and works well in large samples. In small samples, data with many variables relative to the number of observations, or highly skewed categorical data, EMs can be misleading in the shape or variance of the distribution. As a result, the standard errors of the multiple imputations, and ultimately of the quantities of interest, may be biased.

6.4 EMis

EM works well for finding the mode, and EMs works well in large samples for creating fast and independent imputations, but not well for smaller samples. We now correct the problem with EMs with a round of importance sampling (or "sampling importance/resampling"), an iterative simulation technique not based on Markov chains, to get the best of both worlds (Rubin, 1987: 192–4; Tanner, 1996; Gelman et al., 1996; Wei and Tanner, 1990).

EMis (EM with importance sampling) follows the same steps as EMs except that draws of θ from its asymptotic distribution are treated only as approximations to the true (finite sample) posterior distribution. We also put the parameters on unbounded scales to make the normal approximation work better with smaller sample sizes. As in King (1997: 136), we take the natural logarithm of the standard deviation terms and the Fisher z transformation of the correlation parameters, leaving the means alone. We then use an acceptance-rejection algorithm by keeping draws of $\tilde{\theta}$ with probability proportional to the "importance ratio" — the ratio of the actual posterior to the asymptotic normal approximation, both evaluated at $\tilde{\theta}$ — and rejecting the rest. Without prior distributions, the importance ratio is

$$\mathrm{IR} = \frac{L\left(\tilde{\theta} \mid D_{\mathrm{obs}}\right)}{N\left(\tilde{\theta} \mid \hat{\theta}, V(\hat{\theta})\right)} \tag{14}$$

We find that the normal approximation is good enough even in small, nonnormal samples so that the rate of acceptance is high enough to keep the algorithm operating quickly. In the final step, these draws of $\tilde{\theta}$ are used with Equation 10 to produce the desired m imputations.

EMis has all the advantages of IP, since it produces multiple imputations from the exact, finite sample posterior distribution. In addition, it is very fast and does not rely on stochastic convergence criteria. The resulting imputations are fully independent, as required.

7 Monte Carlo Evidence

In this section, we provide several analyses based on simulated data: a timing test that shows how much faster EMis is than IP under different conditions; an illustration of how EMis corrects the problems in EMs and EM in order to match IP's (correct) posterior distribution; and more extensive Monte Carlo evidence that demonstrates that IP and EMis are giving the same answers, and that these results are only slightly worse than if no data were missing and normally far better than listwise deletion.

First, we compare the time it takes to run IP and EMis. Since imputation models are generally run once, followed by numerous analysis runs, imputation methods that take a while are still useful. When runs start taking many hours, however, they make productive analysis much less likely, especially if one has several data sets to analyze.

We made numerous IP and EMis runs, but timing IP precisely is not obvious due to the absence of clear rules for knowing when the stochastic convergence algorithm has finished. As is, we made educated guesses about convergence, based on our experiments where we knew the distribution to which IP was converging, profile plots of the likelihood function, and, when possible, using Schafer's (1997) recommended defaults. On the basis of this experience, we chose $\max(1000, 100p)$ iterations to generate the explicit timing numbers below. We used a computer with average speed, which would be roughly what most users have access to in 1998 (a 200Mhz Pentium with 96M of memory). We then created a data set with 1000 observations, of which 50 observations and one variable were fully observed. Every remaining cell was missing with 5% probability, which is not far from most political science survey data.

For five variables, IP takes 14 minutes, whereas EMis finishes in 13 seconds. For 10 variables, IP takes 1 hour 25 minutes and EMis runs for 48 seconds. With 20 variables, IP takes 21 hours (depending on convergence criterion), whereas EMis takes 3.6 minutes. With 40 variables, IP takes 34.3 days, which is probably beyond the range of what is feasible, while EMis runs for 68 minutes. Overall, EMis ranges from 65 to 726 times faster, with the advantage increasing with the number of variables. Counting the analyst's time that is necessary to evaluate $p+p(p+1)/2$ convergence plots for each of the parameters in IP (since convergence should generally be evaluated by the worst converging parameters, you need to look at them all) would make these comparisons more dramatic. Running one IP chain would be about twice as fast as the recommended approach of separate chains, but that would require evaluating an additional $p+p(p+1)/2$ autocorrelation function plots.[10]

Second, we plot smooth histograms (density estimates of 200 simulations) of one mean parameter from a Monte Carlo run to illustrate how EM, EMs, and EMis approximate the posterior computed by IP and known to be correct. Figure 1 gives these results. The first row of graphs are for $n = 25$ and the second row are for $n = 500$. The first column compares EMs and EM to IP and the second EMis to IP. In all four figures, the correct posterior, computed by IP, is a solid line. The first point emphasized by these figures is that the maximum likelihood point estimate found by EM is not an adequate approximation to the entire posterior distribution. As a result, multiple imputation analyses that use EM ignore estimation variability and thus underestimate the standard errors and confidence intervals of their quantities of interest.

The figure also enables us to evalute EMs and EMis. For example, the dashed line in the top left graph shows how, with a small sample, EMs produces a poor approximation to the true IP posterior. The bottom left graph shows how EMs improves with a larger sample, courtesy of the central limit theorem. In this example, more than 500 observations are apparently required to have a close match between the two, but EMs does not perform badly with $n = 500$. In contrast, EMis closely approximates the true IP posterior when the sample is as small as 25 (in the top right) and not noticeably different when $n = 500$. (The small differences remaining between the lines in the two right graphs are attributable to approximation error in drawing the graphs based on only 200 simulations.)

[10] We programmed IP and EMis using the same compiler (Gauss) to keep the two comparable. Additional vectorization will speed up both algorithms. For example, Schafer's (1997) FORTRAN implementation of IP (which should be approximately as fast as vectorized code in a modern vectorized language) is about 40 times as fast as our Gauss implementation of IP following Schafer's pseudocode (although our incompletely vectorized Gauss implemention of EMis is still at least 10 times faster). We also expect substantial gains in speed for EMis when it too is fully vectorized.

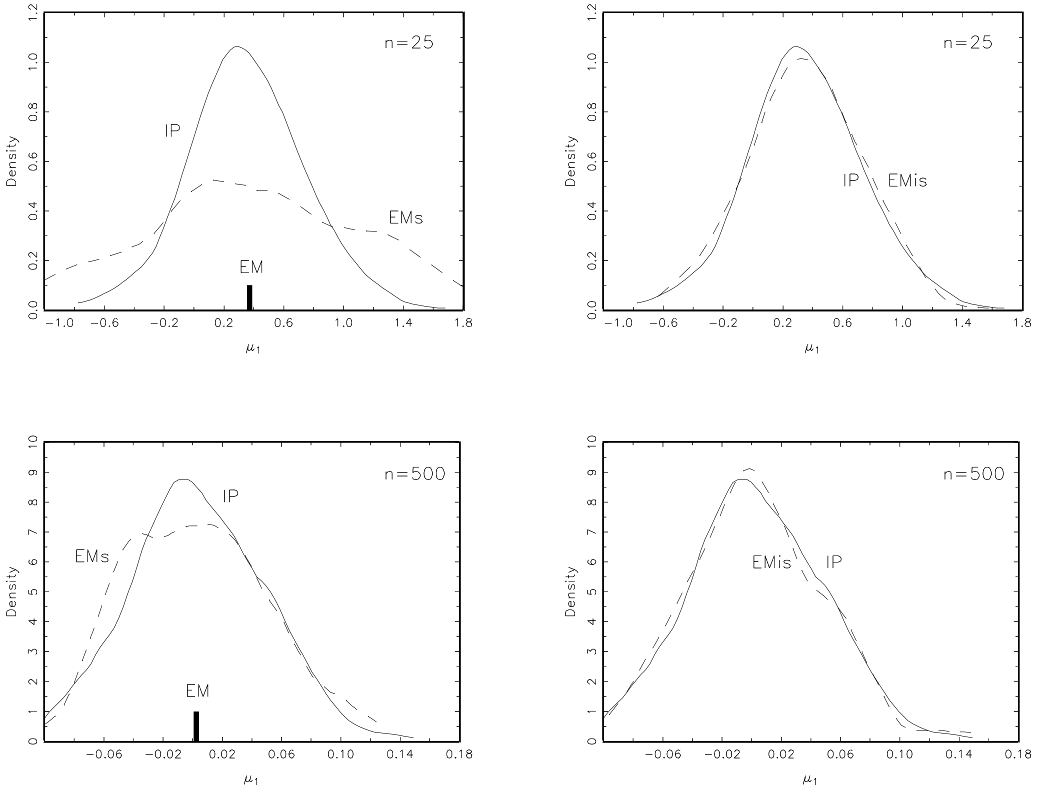

Figure 1: *Comparing Posterior Distributions*

Finally, we provide Monte Carlo evidence by generating data sets and missingness with different characteristics and compare their mean square errors. Since a Monte Carlo experiment is always a test of a discrete point in a continuous parameter space, there is no end to the possible data generation mechanisms one can analyze. The ones we present here were representative of the many others we tried and were consistent with others in the literature. We first generated 100 data sets randomly from each of five data generation processes, each with five variables, $Y, X_1, \dots, X_4$. We defined our quantities of interest as β_1 and β_2 in the regression $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. Including variables in the imputation model even if they do not appear in the analysis model (such as X_3 and X_4) is generally a good idea since the extra variables can be used to help predict the missing values. (Doing the reverse is not recommended; see Meng, 1994.)[11]

MCAR-1 Y, X_1, X_2, X_4 are MCAR; X_3 is completely observed. About 83% of the variables used in the regression are fully observed.

MCAR-2 The same as MCAR-1, with about 50% of rows fully observed.

[11] We chose regression as our analysis model for these experiments because it is probably still the most commonly used statistical method used in political science and most social sciences. Obviously any other analysis model could have been chosen instead, but much research has already demonstrated that multiple imputation works in a diverse variety of situations. For our own testing, we also did extensive runs with logit, linear probability, and several univariate statistics, as well as more limited testing with other more complicated models.

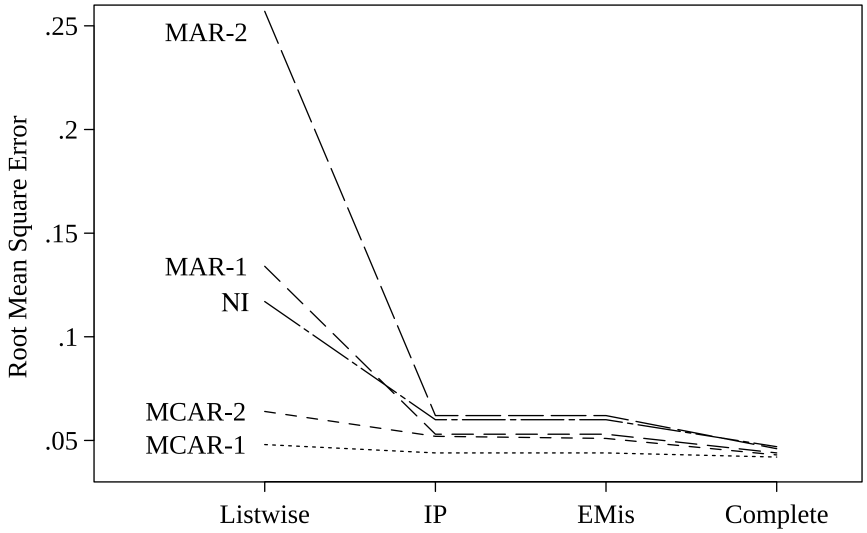

Figure 2: *Root Mean Square Error Comparisons: This figure plots the average root mean square error for four missing data procedures — listwise deletion, multiple imputation with IP and EMis, and the true complete data — and the five data generation processes described in the text. Each point in the graph represents the root mean square error averaged over two regression coefficients in each of 100 simulated data sets. Note how IP and EMis have the same root mean square error, which is lower than listwise deletion and higher than the complete data.*

MAR-1 Y, X_4 are MCAR; X_1, X_2 are MAR, with missingness a function of X_4. X_3 is completely observed. About 78% of rows are fully observed.

MAR-2 The same as MAR-1, with about 50% of rows fully observed.

NI A NonIgnorable missingness mechanism with missing values in Y, X_2 depending on their observed and unobserved values, X_1 depending on the observed and unobserved values of X_3, and with X_3, X_4 generated as MCAR. About 50% of rows are fully observed.

The Σ matrix was set so that the regression coefficients β_1 and β_2 would each be about 0.1. For each of the 100 data sets and five data generation processes, we estimated these regression coefficients using imputation models based on listwise deletion, IP, EMis, and with the true complete data set. For each application of IP and EMis, we multiply imputed ten data sets and averaged the results as described in Section 5. We then computed the average root mean square error for the two coefficients in each run, and then averaged these over the 100 simulations for each data type and statistical procedure.

The vertical axis in Figure 2 is this averaged root mean square error. Each line connects the four different estimations for a single data generation process. The graph helps us demonstrate three points. First, the root mean square error of EMis is virtually identical to that for IP, for each data generation process. This confirms again the equivalence of the two approaches. Second, the error for EMis and IP is not much higher than the complete (usually unobserved) data set, despite high levels of missingness. Finally, listwise deletion, the current practice in political science, ranges from slightly inferior to the two multiple

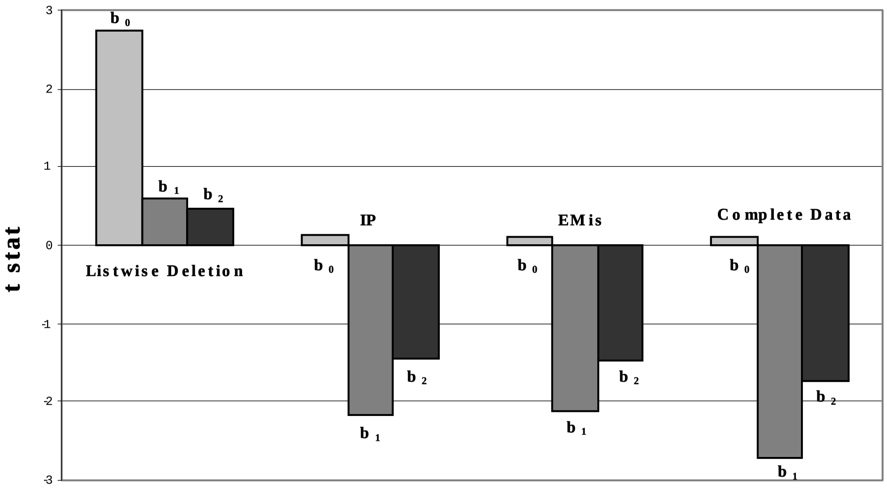

Figure 3: *Monte Carlo Comparison of t-Statistics: t statistics for the constant (b_0) and the two regression coefficients (b_1, b_2) for the MAR-1 run in Figure 2.*

imputation methods — in the MCAR cases when the assumptions of listwise deletion hold — to a disaster — in the MAR and NI cases. Since the true value of the coefficients being estimated is about 0.1, root mean square errors this large can bias results by flipping signs or greatly changing magnitude. Which articles in political science have immense mean square errors like that for MAR-2? Undoubtedly, some do and some don't, but we cannot tell which until political scientists start using more appropriate methods.

To illustrate the results of our Monte Carlo study further, Figure 3 gives a different view of the results of the MAR-1 run in Figure 2. MAR-1 was the case of low missingness, where the root mean square error for listwise deletion was higher than for the other methods but not as high as for MAR-2. Figure 3 graphs the t statistic for the constant term and each of the two regression coefficients, averaged over the 100 runs for each of the four imputation procedures. For the two regression coefficients, the sign is negative (and "significant" for b_1) when estimated by the true complete data, IP, and EMis, but the opposite for listwise deletion. In the listwise deletion run, both coefficients have point estimates that are positive, but statistically indistinguishable from zero. Most of the action in the listwise case is generated in the substantively uninteresting constant term.

Figure 3 is a clear example of the dangers political scientists face in continuing to use listwise deletion as our primary method of coping with missing data problems. Only 22% of the observations were lost by listwise deletion in this case, and yet what would be the key substantive conclusions are reversed by choosing an inferior method. We can easily generate hypothetical data with effects that are of far larger magnitudes, but we feel this one is probably representative of much work in political science and of the risks we face.

8 Replications

In this section we report on the replication of two scholarly analyses. One is a "replication before publication," which we conducted in order to help a colleague preparing a book manuscript. The other replicates a published article. Both analyses demonstrate how markedly different substantive conclusions can come from switching from listwise deletion to our approach to multiple imputation.

Different conclusions will not always occur from switching between these methods. In fact, we replicated two other studies and found more modest effects than we show below. We examined Domínguez and McCann's (1996) study of Mexican elections and found that the multiple imputation results somewhat strengthened the authors' conclusions. Their main argument, that voters focused primarily on the potential of the ruling party rather than specific issues, came through stronger under multiple imputation. In addition, several of the specific issue positions that Domínguez and McCann were forced to justify ignoring or explaining away, turned out to be an artifact of listwise deletion.

We also replicated Dalton et al.'s (1998) analysis of partisan cues from newspaper editorials, in which they analyzed a merged data set of editorial content analyses and individual level survey responses. Most missing data in this study resulted from the authors' inability to content analyze the numerous newspapers reported having been read by individual respondents. Because the other survey variables contained little information useful for predicting content analyses that were not completed, an MCAR missingness mechanism could not be rejected from the data, resulting in no substantial change in the point estimates with multiple imputation. Of course, the confidence intervals and standard errors did get smaller. Since Dalton et al.'s analysis was at the county-level, it would be possible to gather additional variables from census data and add them to the imputation stage. If this were done, our approach would likely have a larger effect.

8.1 Voting Behavior in Russian Elections

Our first example explores missing data problems with Timothy Colton's (1998) research on voting behavior in recent Russian elections. Colton proposes and tests an extensive model of electoral choice in Russia's 1995 parliamentary election and 1996 presidential election. He finds, among many other things, the emergence of systematic patterns in the choices of Russian voters. This finding contradicts many contemporary accounts of voting behavior in emerging democracies which portray electoral choices as random at worst and personalistic at best.

Colton's data are taken from the 1995–1996 Russian Election Study. We focus our attention on only a very small portion of Colton's study, and simplify his analysis for our purposes. Specifically, we estimate a logit model with the dependent variable defined as a one if the voter casts his or her ballot for the Communist Party of the Russian Federation (KPRF) and zero otherwise. With over 22% percent of the popular vote, the KPRF was the clear winner in the 1995 parliamentary elections and thus understanding voter support for this party is essential to understanding Russian voting behavior. The explanatory variables for the model vary depending on which stage of the voter's decision making process is being tested, in order to avoid controlling for the consequences of key causal variables. Listwise deletion loses 36%, 56%, and 58% respectively in the three stages from which we use data. The stages and specific measures are consistent with previous voting studies and we refer the interested reader to Colton (1998) for details.

In Table 2, we present estimates of three quantities of interest derived from our logit regressions for listwise deletion and multiple imputation. First, we estimate the effect

	Listwise	Multiple Imputation
Satisfaction with Democracy	−.06	−.10
	(.06)	(.04)
Opposition to the Market Economy	.08	.12
	(.08)	(.05)
Trust in the Russian Government	−.06	−.12
	(.08)	(.04)

Table 2: *First Difference Effects on Voting in Russia: entries are changes in the probability of voting for the Communist party in the 1995 parliamentary election as a function of changes in the explanatory variable (listed on the left), with standard errors in parentheses.*

of a voter's satisfaction with democracy on the probability of supporting the KPRF. In Colton's model, satisfaction with democracy is one measure of voters' assessments of current economic and political conditions in Russia. He hypothesizes that voters more satisfied with democracy are less likely to support the KPRF than those who are dissatisfied. The quantity of interest is the difference between the predicted probability for a voter who is completely dissatisfied with how democracy is developing in Russia and the predicted probability for a voter who is completely satisfied, holding all other values of the explanatory variables constant at their means. The listwise deletion estimate of this parameter is -0.06 with a relatively large standard error of 0.06 — for all practical purposes no finding. In contrast the multiple imputation estimate, is -0.10 with a standard error of 0.04. The unbiased and more efficient multiple imputation estimate is nearly twice as large and estimated much more precisely. Thus, with our better procedure we can be relatively confident that individuals highly satisfied with Russian democracy were about 10% less likely to support the KPRF, a fact not ascertainable with existing methods.

Colton is also interested in examining the effect of issue opinions on vote choice. For example, are voters opposed to the transition to a market economy more likely to support the communist party? Using the listwise deletion estimator, we find little support for this hypothesis as again the first difference estimate is in the hypothesized direction but is estimated imprecisely. The multiple imputation estimate, however, suggests that voters opposed to the transition were about 12% more likely to vote with the KPRF, with a small standard error. The final comparison that we report is on the effect of an individual's trust in the Russian Government on vote choice. Positive evaluations should have had a negative impact on KPRF voting at the time of this Duma election. Again, listwise deletion detects no effect, while multiple imputation finds a precisely estimated twelve percentage point difference.

The first differences in Table 2 represent only three of the logit effects estimated. Overall, this analysis included 46 coefficients, of which 10 changed in importance judging by traditional standards (from "statistically significant" to not or the reverse, plus some substantively meaningful difference). In addition, roughly 5 other coefficients increased or decreased in magnitude sufficiently to alter the substantive interpretation of their effects.

8.2 Public Opinion About Racial Policies

We also replicate Alvarez and Brehm's (1997) analysis of the factors explaining Americans' racial policy preferences as well as the variance in those preferences. To explain the variance they use a heteroskedastic probit to model respondent preferences over racial policies in fair-housing laws, government set asides, taxes to benefit minority educational opportunities, and affirmative action in university admissions. They find that the "individual variability in attitudes toward racial policy stems from uncertainty" derived from a "lack of political information" and not from a conflict of core values, such as individualism or egalitarianism.

To tap Americans' core values and predict individual policy preferences, Alvarez and Brehm construct "core belief scales" from responses to related feeling thermometers and agree/disagree measures. Contrary to the interpretation that modern racism is simply a proxy for antiblack stereotypes, authoritarianism, and egalitarianism about which people have preferences, they find that only modern racism, of all the scales, has consistent power to explain policy choice.

Constructing the scale variables exacerbates missing data problems since a missing value in any of the items in the scale causes the entire scale value for that observation to be missing. Thus, a deceptively small number of explanatory variables, from which we might not usually have large missingness problems in a well designed survey, actually contains the missing values and missing mechanisms of all their many components. This problem of missing observations was severe, since listwise deletion would have resulted in over half of the observations being lost.

Alvarez and Brehm responded to this problem by replacing the ideology scale with an alternate if the respondent had refused to answer or did not know their ideology in the liberal-conservative terms used. The alternate question pressed the respondent to choose liberal or conservative, which Alvarez and Brehm coded as a neutral with a weak leaning. This is a clear case of unobserved data, with a reasonable but ad hoc imputation method. If the question concerned party identification, a valid response might be "none" and this might not be a missing value, but merely an awkward response for the analyst. However, while "ideological self-placement" might be legitimately missing, it is the self-placement which is to blame. The individual presumably has some ideological standing, no matter how uncertain, but is not able to communicate it to us with our terminology in our survey question. To press the respondent to guess and for the analyst to guess how to code these values on the same scale as the original question risks attenuating the estimated relationships.

Fortunately, using the forced question is unnecessary since from the questions on homelessness, poverty, taxes, and abortion, we can easily predict the technical placement we are looking for without shifting the responsibility to the respondent who does not understand, or has not thought about our academic jargon. Indeed bias would seem to be a problem here, since in the Alvarez and Brehm analysis, ideology rarely has an effect. However, if we impute instead of guess the ideological scale, it becomes significant just over half the time, and the coefficients all increase in both the choice and the variance models (of all the dependent variables estimated).

We use multiple imputation for the missing components of the scales to counter the problem of non-response with greater efficiency and less bias. We present first difference results in the style of Alvarez and Brehm in Table 3.

While the main substantive finding of the effect of modern racism still holds (and is in fact strengthened), the secondary finding explaining individual preferences, which contributes to the more mainstream and developed policy argument, is now reversed. The

	Replication	Multiple Imputation
Modern racism	−.195*	−.220*
	(.045)	(.043)
Individualism	.016	.001
	(.019)	(.019)
Anti-black	−.019	−.001
	(.044)	(.042)
Authoritarianism	.025	.040*
	(.020)	(.020)
Antisemitism	−.074	−.096*
	(.045)	(.042)
Egalitarianism	.151*	.143*
	(.039)	(.032)
Ideology	−.076	−.119*
	(.051)	(.051)
N	1574	2009
χ^2	7.95	11.50*
$p(\chi^2)$	.09	.02

Table 3: *Estimated First Differences of Core Beliefs: The first column of numbers replicates Alvarez and Brehm's (1997) calculation of first difference effects for the substantive variables in the mean function, with the addition of standard errors. The second column is derived from our multiple imputation reanalysis. Asterisks in the table indicate $p < 0.05$, as in the original article. The χ^2 test indicates whether the heteroskedastic probit model is distinguishable from the simpler probit model.*

act of individual racial policy choice now appears to be a broad function of many competing values, no longer driven only by modern racism. An individual's level of authoritarianism, antisemitism, and egalitarianism, as well as their ideological position, all strongly affect the probability of supporting an increase in taxes for minority educational opportunities. Alvarez and Brehm were thus correct to hedge their opposite conclusion on this point.[12]

Finally, and quite importantly, the chi-square test reported at the bottom of Table 3 is insignificant under Alvarez and Brehm's original specification, but is now significant. This test measures whether their sophisticated analysis model is statistically superior to a simple probit choice model, and thus whether the terms in the variance model warrant our attention. Under their treatment of missing values, the variance component of the model does not explain the between-respondent variances, meaning that the test indicates that their methodological complications are superfluous. However, our approach rejects the simpler probit in favor of their more complicated model.

[12] The variance model in the heteroskedastic probit is still dominated by the chronic information term. This affirms Alvarez and Brehm's conclusion that shows that variance in policy choice between respondents is driven by a lack of information, and not a conflict between the core values of egalitarianism and individualism.

9 Concluding Remarks

For political scientists conducting substantive research, most any disciplined statistical model of multiple imputation would do better than our current practices. The threats to the validity of our inferences stemming from listwise deletion are of roughly the same magnitude as those resulting from the much better known problems of omitted variable bias. Our proposed new "default" method is much faster and far easier to use than existing multiple imputation methods, and amounts to a way of using about 50% more information in our data than we now use. This method will surpass listwise deletion in most cases when there exists information in the data with which to predict the missing values. Political scientists can also easily jettison the nearly universal but biased practice of making up the answers for some missing values. Although it is of course possible to fool any statistical method including this one, and although we generally prefer application-specific methods when they are available, multiple imputation with our algorithm will normally do better than current practices.

The idea of multiple imputation seems well-suited to make statistical analysis easier for applied researchers, but the methods of imputation were so difficult to use that in the twenty years since the idea was put forward it has been used by only a few of the most sophisticated statistical researchers. Outside of these few experts, "the method has remained largely unknown and unused" (Schafer and Olsen, 1998). We hope the suggestions we provide herein will bring this powerful idea to some of those who can put it to best use.

Indeed, we believe the method offered here may make a material difference in the life of and research produced by many social scientists. For example, consider a graduate student writing a dissertation who needs to collect about eight months worth of complete data in uncomfortable circumstances far from home. Ideally every datum collected would be complete, but even the best researchers lose about one-third of their cases to item nonresponse and listwise deletion. So nonresponse must be anticipated as part of any realistic research plan. However, instead of booking a trip for 12 months and planning to lose a third of the data, and four months of his or her life, it probably makes more sense to collect data for 8 months and take a few days to learn and implement our methodology.

Or consider the community of researchers using the National Election Studies and other large public use data sets. These researchers have made numerous important findings, but inevitably others remain ambiguous: confidence intervals are too wide and item nonresponse bias looms large. Learning the methods offered here seems vastly easier and more justifiable than trying to convince the National Science Foundation or other funding agencies of the need for additional expensive data collection. Indeed, since using these methods will make federal funds go about 50% farther, at essentially no cost, it may even be our obligation to use them.

Finally, as an analogy to the 1970s, imagine carrying your carefully key-punched cards to the computer center to do a run, and accidentally dropping one-third of them into the street. Do you bother to pick them up and try very hard to put them back in order, or do you keep walking?

A Proof of Mean Square Error Comparisons

Model Let $E(Y) = X\beta = X_1\beta_1 + X_2\beta_2$ and $V(Y) = \sigma^2 I$, where $X = (X_1', X_2')'$, $\beta = (\beta_1', \beta_2')'$, and λ is the fraction of rows of X_2 missing completely at random (other rows of X_2 and all of Y and X_1 are observed). The ultimate goal is to find the best estimator for β_1; the specific goal is to derive Equation 1.

We evaluate the three estimators of β_1 in Section 3 by comparing their mean square errors (MSE). The MSE of an estimator $\hat{\theta}$ with respect to θ is, roughly speaking, how close the distribution of $\hat{\theta}$ is concentrated around θ. More formally, $\text{MSE}(\hat{\theta}, \theta) = E[(\hat{\theta} - \theta)^2] = V(\hat{\theta}) + E(\hat{\theta} - \theta)E(\hat{\theta} - \theta)' = \text{variance} + \text{bias}^2$. We begin with a formal definition of the three estimators and then compute, bias, variance, and MSE of each. We then derive the result in Equation 1.

Estimators We consider three estimators (with underlined letters indicating our mnemonic notational device): Let $b^I = AY = (b_1^{I'}, b_2^{I'})'$, where $A = (X'X)^{-1}X'$. Then b_1^I denotes the Infeasible estimator of β_1. Let $b_1^O = A_1 Y$ be the Omitted variable bias estimator of β_1, where $A_1 = (X_1'X_1)^{-1}X_1'$. Finally, let $b^L = A^L Y^L = (b_1^{L'}, b_2^{L'})'$, where $A^L = (X^{L'}X^L)^{-1}X^{L'}$, and where the superscript L denotes listwise deletion applied to X and Y (i.e., deleting rows of all three variables when rows of X_2 are missing). So b_1^L is the Listwise deletion estimator of β_1.

Bias The infeasible estimator is unbiased: $E(b^I) = E(AY) = AX\beta = \beta$, and thus $\text{bias}(b_1^I) = 0$. The omitted variable estimator is biased, as per the usual calculation, $E(b_1^O) = E(b_1^I + Fb_2^I) = \beta_1 + F\beta_2$, where each column of F is a vector of coefficients from a regression of a column of X_2 on all columns of X_1, and so $\text{bias}(b_1^O) = F\beta_2$. If MCAR holds, the listwise deletion estimator is also unbiased: $E(b^L) = E(A^L Y^L) = A^L X^L \beta = \beta$, and thus $\text{bias}(b_1^L) = 0$.

Variance The variance of the infeasible estimator is $V(b^I) = V(AY) = A\sigma^2 IA' = \sigma^2(X'X)^{-1}$. Since $V(b_1^I) = V(b_1^I - Fb_2^I) = V(b_1^O) - FV(b_2^I)F'$, the omitted variable bias variance is $V(b_1^O) = V(b_1^I) - FV(b_2^I)F'$. And since $V(b^L) = V(A^L Y^L) = A^L \sigma^2 I A^{L'} = \sigma^2(X^{L'}X^L)^{-1}$, the variance of the listwise deletion estimator is $V(b_1^L) = \sigma^2(Q^L)^{11}$, where $(Q^L)^{11}$ is the upper left portion of the $(X^{L'}X^L)^{-1}$ matrix corresponding to X_1^L.

MSE Putting together the (squared) bias and variance results gives MSE computations for the omitted variable bias and listwise deletion estimators: $\text{MSE}(b_1^O) = V(b_1^I) + F[\beta_2\beta_2' - V(b_2^I)]F'$, and $\text{MSE}(b_1^L) = \sigma^2(Q^L)^{11}$.

Comparison In order to evaluate when the listwise deletion estimator outperforms the omitted variable bias estimator, we can compute the difference in MSE, which we denote by d:

$$\begin{aligned} d &= \text{MSE}(b_1^L) - \text{MSE}(b_1^O) \\ &= [V(b_1^L) - V(b_1^I)] + F[V(b_2^I) - \beta_2\beta_2']F' \qquad (15) \end{aligned}$$

Listwise deletion is better than omitted variable bias when $d < 0$, worse when $d > 0$, and no different when $d = 0$. The second term in Equation 15 is the usual bias-variance tradeoff, and so our primary concern is with the first term. Since

$$\begin{aligned} V(b^I)[V(b^L)]^{-1} &= \sigma^2(X^{L'}X^L + X_{\text{mis}}'X_{\text{mis}})^{-1}\frac{1}{\sigma^2}(X^{L'}X^L) \\ &= (X^{L'}X^L + X_{\text{mis}}'X_{\text{mis}})^{-1}(X^{L'}X^L) \\ &= I - (X^{L'}X^L + X_{\text{mis}}'X_{\text{mis}})^{-1}(X_{\text{mis}}'X_{\text{mis}}) \end{aligned}$$

where X_{mis} includes the rows of X deleted by listwise deletion (so that X^L and X_{mis} comprise all the information in X). Since exchangability among rows of X is implied by the MCAR assumption (or equivalently taking the expected value over sampling permutations), we can write $(X^{L'}X^L + X'_{\text{mis}}X_{\text{mis}})^{-1}(X'_{\text{mis}}X_{\text{mis}}) = \lambda$, which implies $V(b_1^L) = V(b^I)/(1-\lambda)$, which by substitution into Equation 15 yields, and thus completes the proof of, Equation 1.

References

Achen, Christopher. 1986. *Statistical Analysis of Quasi-experiments*, Berkeley: University of California Press.

Alvarez, R. Michael and John Brehm. 1997. "Are Americans Ambivalent Towards Racial Policies?" *American Journal of Political Science*, 41, 2 (April): 345–374.

Amemiya, Takeshi. 1985. *Advanced Econometrics*, Cambridge: Harvard University Press.

Anderson, Andy B.; Alexander Basilevsky; and Derek P.J. Hum 1983. "Missing Data: A Review of the Literature," Pp. 415–494 in Peter H. Rossi, James D. Wright and Andy B. Anderson, eds., *Handbook of Survey Research* Academic Press, Inc.

Bartels, Larry. 1998. "Panel Attrition and Panel Conditioning in American National Election Studies" paper prepared for the 1998 meetings of the Society for Political Methodology, San Diego.

Bartels, Larry. 1996. "Uninformed Votes: Information Effects in Presidential Elections," *American Journal of Political Science*, 40: 194–230.

Brehm, John. 1993. *The Phantom Respondents: Opinion Surveys and Political Representation.* Ann Arbor: University of Michigan Press.

Colton, Timothy. 1998. "Transitional Citizenship: Voting in Post-Soviet Russia," book manuscript in progress.

Cowles, Mary Kathryn and Bradley P. Carlin. 1996. "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review," *Journal of the American Statistical Association*, 91, 434 (June): 883–904.

Dalton, Russell J.; Paul A. Beck; and Robert Huckfeldt. 1998. "Partisan Cues and the Media: Information Flows in the 1992 Presidential Election," *American Political Science Review*, 92: 111-126

Dempster, Arthur P. et al. 1977. "Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Association*, B, 39: 1–38.

Ezzati-Rice, T.M., et al., 1995. "A Simulation Study to Evaluate the Performance of Model-Based Multiple Imputations in NCHS Health Examination Surveys," in *Proceedings of the Annual research Conference*, pp. 257–266. Bureau of the Census, Washington, D.C.

Fay, Robert E. 1992. "When are Inferences from Multiple Imputation Valid?" *Proceedings fo the Survey Research methods Section of the American Statistical Association*, 81: 354–365.

Franklin, Charles. H. 1989. "Estimation across data sets: two-stage auxiliary instrumental variables estimation (2SAIV)," *Political Analysis* 1: 1–24.

Globetti, Suzanne. 1997. "What We Know About 'Don't Knows': An Analysis of Seven Point Issue Placements," paper presented at the annual meetings of the Political Methodology Society, Columbus, Ohio.

Graham, J.W. and J.L. Schafer. In press. "On the performance of Multiple Imputation for Multivariate Data with Small Sample Size," in Hoyle, R., ed., *Statistical Strategies for Small Sample Research*, Sage: Thousand Oaks.

Heckman, James. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables, and Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, 5: 475–492.

Herron, Michael C. 1998. "Voting, Abstention, and Individual Expectations in the 1992 Presidential Election," paper presented for the Midwest Political Science Association conference, Chicago.

Kass, Robert E.; Bradley P. Carlin; Andrew Gelman; and Radford M. Neal. 1998. "Markov Chain Monte Carlo in Practice: A Roundtable Discussion" *The American Statistician.*

Katz, Jonathan and Gary King. 1997. "A Statistical Model for Multiparty Electoral Data," paper presented at the annual meetings of the Midwest Political Science Association, Chicago.

King, Gary. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*, Princeton: Princeton University Press.

King, Gary. 1989. *Unifying Political Methodology: The Likelihood Model of Statistical Infernce*, Cambridge: Cambridge University Press.

King, Gary; James Alt; Nancy Burns; and Michael Laver. 1990. "A Unified Model of Cabinet Dissolution in Parliamentary Democracies," *American Journal of Political Science*, 34, 3 (August): 846–871.

King, Gary; Michael Tomz; and Jason Wittenberg. 1998. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation," paper prepared for the Annual Meetings of the American Political Science Association, Boston.

Landerman, Lawrence R.; Kenneth C. Land; and Carl F. Pieper. in press. "An Empirical Evaluation of the Predictive Mean Matching Method for Imputing Missing Values," *Sociological Methods and Research.*

Little, J. Rodrick and Donald Rubin. 1987. *Statistical Analysis with Missing Data*, New York: Wiley.

Liu, J., Wong, W. H., and Kong, A. 1994. "Covariance Structure of the Gibbs Sampler with Applications to the Comparisons of Estimators and Augmentation Schemes," *Biometrika*, 81, 27–40.

Meng, X.L. 1994. "Multiple-imputation Inferences with Uncongenial Sources of Input," *Statistical Science*, 9, 4: 538–573.

Raghunathan, T. E., and Grizzle, J. E. 1995. "A Split Questionnaire Survey Design," *Journal of the American Statistical Association*, 90: 54–63.

Rao, J.N.K. 1996. "On Variance Estimation with Imputed Survey Data," *Journal of the American Statistical Association*, 91: 499–506.

Rubin, Donald. 1996. "Multiple Imputation after 18+ Years," *Journal of the American Statistical Association*, 91: 473–89.

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

Rubin, Donald. 1977. "Formalizing Subjective Notions about the Effect of Nonrespondents in Sample Surveys," *Journal of the American Statistical Association*, 72, 538: 543.

Rubin, D. B., and Schafer, J. L. 1990. "Efficiently Creating Multiple Imputations for Incomplete Multivariate Normal Data," *Proceedings of the Statistical Computing Section of the American Statistical Association*, 83–88.

Rubin, Donald and Nathaniel Schenker. 1986. "Multiple Imputation for Interval Estimation From Single Random Samples eith Ignorable Nonresponse," *Journal of the American Statistical Association*, 81, 394: 366-374.

Schafer, Joseph L. 1997. *Analysis and Simulation of Incomplete Multivariate Data: Algorithms and Examples*, Chapman and Hall.

Schafer, Joseph L.; Meena Khare; and Trena M. Ezzati-Rice. 1993. "Multiple Imputation of Missing Data in NHANESIII Proceedings of the Annual Research Conference," Bureau of the Census, D.C., 459-487.

Schafer, Joseph L. and Maren K. Olsen. 1998. "Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective," *Multivariate Behavioral Research*, forthcoming.

Sherman, Robert P. 1998. "A Test of the Validity of Complete-Unit Analysis in Surveys Subject to Item Nonresponse or Attrition," manuscript, Caltech.

Skalaban, Andrew. 1992. "Interstate Competition and State Strategies to Deregulate Interstate Banking 1982-1988," *Journal of Politics*, 54, 3. (August): 793–809.

Stolzenberg, Ross M. and Daniel A. Relles. 1990. "Theory Testing in a World of Constrained Research Design: The Significance of Heckman's Censored Sampling Bias Correction for Nonexperimental Research," *Sociological Methods and Research*, 18, 4 (May): 395–415.

Tanner, Martin A. 1996. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, third edition, New York: Springer-Verlag.

Tanner, M. A., and Wong, W. H. 1987. "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82: 528–550.

Timpone, Richard J. 1998. "Structure, Behavior, and Voter Turnout in the United States," *American Political Science Review*, 92, 1: 145–158.

Wei, Greg C. G. and Martin A. Tanner. 1990. "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms," *Journal of the American Statistical Association*, 85: 699–704.

Winship, Christopher and Robert D. Mare. 1992. "Models for Sample Selection Bias," *Annual Review of Sociology*, 18: 327–50.

Winship, Christopher and Larry Radbill. 1994. "Sampling Weights and Regression Analysis," *Sociological Methods and Research*, 23, 2 (November): 230–257.