Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation

GARY KING *Harvard University*
JAMES HONAKER *Harvard University*
ANNE JOSEPH *Harvard University*
KENNETH SCHEVE *Yale University*

We propose a remedy for the discrepancy between the way political scientists analyze data with missing values and the recommendations of the statistics community. Methodologists and statisticians agree that "multiple imputation" is a superior approach to the problem of missing data scattered through one's explanatory and dependent variables than the methods currently used in applied data analysis. The discrepancy occurs because 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 have demanded considerable expertise. We adapt an algorithm and use it to implement a general-purpose, multiple imputation model for missing data. This algorithm is considerably faster and easier to use than the leading method recommended in the statistics literature. We also quantify the risks of current missing data practices, illustrate how to use the new procedure, and evaluate this alternative through simulated data as well as actual empirical examples. Finally, we offer easy-to-use software that implements all methods discussed.

On average, about half the respondents to surveys do not answer one or more questions analyzed in the average survey-based political science article. Almost all analysts contaminate their data at least partially by filling in educated guesses for some of these items (such as coding "don't know" on party identification questions as "independent"). Our review of a large part of the recent literature suggests that 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 filling in guesses for some.[1] Of course, similar problems with missing data occur in nonsurvey research as well.

This article addresses the discrepancy between the treatment of missing data by political scientists and the well-developed body of statistical theory that recommends against the procedures we routinely follow.[2] Even if the missing answers we guess for nonrespondents are right on average, the procedure overestimates the certainty with which we know those answers. Consequently, standard errors will be too small. Listwise deletion discards one-third of cases on average, which deletes both the few nonresponses and the many responses in those cases. The result is a loss of valuable information at best and severe selection bias at worst.

Gary King (King@harvard.edu, http://GKing.Harvard.Edu) is Professor of Government, Harvard University, and Senior Advisor, Global Programme on Evidence for Health Policy, World Health Organization, Center for Basic Research in the Social Sciences, Harvard University, Cambridge, MA 02138. James Honaker (tercer@latte.harvard.edu, http://www.gov.harvard.edu/graduate/tercer/) is a Ph.D. candidate, Department of Government, Harvard University, Center for Basic Research in the Social Sciences, and Anne Joseph (ajoseph@fas.harvard.edu) is a Ph.D. candidate in Political Economy and Government, Harvard University, Cambridge, MA 02138. Kenneth Scheve (kenneth.scheve@yale.edu, http://pantheon.yale.edu/~ks298/) is Assistant Professor, Department of Political Science, Institution for Social and Policy Studies, Yale University, New Haven, CT 06520.

The authors 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, Larry Bartels, Neal Beck, Adam Berinsky, Fred Boehmke, Ted Brader, Charles Franklin, Rob Van Houweling, Jas Sekhon, Brian Silver, Ted Thompson, and Chris Winship for helpful discussions; Joe Schafer for a prepublication copy of his extremely useful book; Mike Alvarez, Paul Beck, John Brehm, Tim Colton, Russ Dalton, Jorge Domínguez, Bob Huckfeldt, Jay McCann, and the Survey Research Center at the University of California, Berkeley, for data; and the National Science Foundation (SBR-9729884), the Centers for Disease Control and Prevention (Division of Diabetes Translation), the National Institutes on Aging (P01 AG17625-01), and the World Health Organization for research support. Our software is available at http://GKing.Harvard.Edu.

[1] These data come from our content analysis of five years (1993–97) of the *American Political Science Review*, the *American Journal of Political Science*, and the *British Journal of Political Science*. Among these articles, 203—24% of the total and about half the quantitative articles—used some form of survey analysis, and 176 of these were mass rather than elite surveys. In only 19% of the articles were authors explicit about how they dealt with missing values. By also asking investigators, looking up codebooks, checking computer programs, or estimating based on partial information provided, we were able to gather sufficient information regarding treatment of missing values for a total of 77% of the articles. Because the situation is probably not better in the other 23% of the articles without adequate reporting, both missing data practices and reporting problems need to be addressed. Our more casual examinations of other journals in political science and other social sciences suggest similar conclusions.

[2] This article is about item nonresponse, that is, respondents answer some questions but not others (or, in general, scattered cells in a data matrix are missing). A related issue is unit nonresponse: 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 into analyses. Globetti (1997) and Sherman (2000) 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), and panel attrition.

Some researchers avoid the problems missing data can cause by using sophisticated statistical models optimized for their particular applications (such as censoring or truncation models; see Appendix A). When possible, it is best to adapt one's statistical model specially to deal with missing data in this way. Unfortunately, doing so may put heavy burdens on the investigator, since optimal models for missing data differ with each application, are not programmed in currently available standard statistical software, and do not exist for many applications (especially when missingness is scattered throughout a data matrix).

Our complementary approach is to find a better choice in the class of widely applicable and easy-to-use methods for missing data. Instead of the default method for coping with the issue—guessing answers in combination with listwise deletion—we favor a procedure based on the concept of "multiple imputation" that is nearly as easy to use but avoids the problems of current practices (Rubin 1977).[3] Multiple imputation methods have been around for about two decades and are now the choice of most statisticians in principle, but they have not made it into the toolbox of more than a few applied statisticians or social scientists. In fact, aside from the experts, "the method has remained largely unknown and unused" (Schafer and Olsen 1998). The problem is only in part a lack of information and training. A bigger issue is that although this method is easy to use in theory, in practice it requires computational algorithms that can take many hours or days to run and cannot be fully automated. Because these algorithms rely on concepts of stochastic (rather than deterministic) convergence, knowing when the iterations are complete and the program should be stopped requires much expert judgment, but unfortunately, there is little consensus about this even among the experts.[4] In part for these reasons, no commercial software includes a correct implementation of multiple imputation.[5]

We begin with a review of three types of assumptions one can make about missing data. Then we demonstrate analytically the disadvantages of listwise deletion. Next, we introduce multiple imputation and our alternative algorithm. We discuss what can go wrong and provide Monte Carlo evidence that shows how our method compares with existing practice and how it is equivalent to the standard approach recommended in the statistics literature, except that it runs much faster. We then present two examples of applied research to illustrate how assumptions about and methods for missing data can affect our conclusions about government and politics.

TABLE 1. Three Missingness Assumptions

Assumption	Acronym	You Can Predict M with:
Missing completely at random	MCAR	—
Missing at random	MAR	D_{obs}
Nonignorable	NI	D_{obs} and D_{mis}

ASSUMPTIONS ABOUT MISSINGNESS

We now introduce three assumptions about the process by which data become missing. Briefly in the conclusion to this section and more extensively in subsequent sections, we will discuss how the various methods crucially depend upon them (Little 1992).

First, let D denote the data matrix, which includes the dependent Y and explanatory X variables: $D = \{Y, X\}$. If D were fully observed, a standard statistical method could be used to analyze it, but in practice, some elements of D are missing. Define M as a missingness indicator matrix with the same dimensions as D, but there is a 1 in each entry for which the corresponding entry in D is observed, or a 0 when 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 prefer not to reveal it in an interview. In some cases, however, "I don't know" given in response to questions about the national helium reserve or the job performance of the Secretary of Interior probably does not mean the respondent is hiding something, and it should be treated as a legitimate answer to be modeled rather than a missing value to be imputed. 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 predicting how people would vote if they were more informed (Bartels 1996). Finally, let D_{obs} and D_{mis} denote *obs*erved and *mis*sing portions of D, respectively, so $D = \{D_{obs}, D_{mis}\}$.

Standard terminology describing possible missingness assumptions is unintuitive (for historical reasons). In Table 1 we try to clarify the assumptions according to our ability to predict the values of M (i.e., which values of D will be missing) (Rubin 1976). For example, missing values in processes that are *missing completely at random* (MCAR) cannot be predicted any better with information in D, observed or not. More formally, M is independent of D: $P(M|D) = P(M)$. An example of an MCAR process is one in which respondents

[3] The most useful modern work on the subject related to our approach is Schafer (1997), which we rely on frequently. Schafer provides a detailed guide to the analysis of incomplete multivariate data in a Bayesian framework. He presents a thorough explanation of the use of the IP algorithm. Little and Rubin (1987), Rubin (1987a), and Rubin (1996) provide the theoretical foundations for multiple imputation approaches to missing data problems.

[4] Although software exists to check convergences, there is significant debate on the adequacy of these methods (see Cowles and Carlin 1996; Kass et al. 1998).

[5] The public domain software that accompanies Schafer's (1997) superb book implements monotone data augmentation by the IP algorithm, the best currently available approach (Liu, Wong, and Kong 1994; Rubin and Schafer 1990). The commercial programs Solas and SPlus have promised implementations. SPSS has released a missing data module, but the program only produces sufficient statistics under a multivariate normality model (means, variances, and covariates), so data analysis methods that require raw data cannot be used. Furthermore, it adds no uncertainty component, which produces standard errors biased toward zero.

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 *missing at random* (MAR) processes, the probability that a cell value is missing may depend on D_{obs} but (after controlling for D_{obs}) must be independent of D_{mis}. Formally, M is independent of D_{mis}: $P(M|D) = P(M|D_{obs})$. 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 to which at least some people respond. 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 can be predicted with any other variables in the data set (such as ideology, issue positions, income, and education). The prediction required is not causal; for example, the vote data could be used whether or not the vote causes or is caused by party identification. To an extent then, the analyst, rather than the world that generates the data, controls the degree to which the MAR assumption fits. It can be made to fit the data by including more variables in the imputation process to predict the pattern of missingness.

Finally, if the probability that a cell is missing depends on the unobserved value of the missing response, the process is *nonignorable* (NI). Formally, M is not independent of D: $P(M|D)$ does not simplify. An example occurs 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.[6]

The performance of different methods of analyzing incomplete data under MCAR, MAR, or NI depends upon the ultimate goals of the analysis. We consider various situations in some detail in subsequent sections, but a few general statements are possible at this stage. First, inferences from analyses using listwise deletion are relatively inefficient, no matter which assumption characterizes the missingness, and they are also biased unless MCAR holds. Inferences based on multiple imputation are more efficient than listwise deletion (since no observed data are discarded), and they are not biased under MCAR or MAR (Little and Rubin 1989; Little and Schenker 1995). Both listwise deletion and basic multiple imputation approaches can be biased under NI, in which case additional steps must be taken, or different models must be chosen, to ensure valid inferences. Thus, multiple imputation will normally be better than, and almost always not worse than, listwise deletion. We discuss below the unusual configuration of assumptions, methods, and analysis models for which listwise deletion can outperform multiple imputation.

In many situations, MCAR can be rejected empirically in favor of MAR. By definition, however, the presence or absence of NI can never be demonstrated using only the observed data. Thus, in most circumstances, it is possible to verify whether multiple imputation will outperform listwise deletion, but it is not possible to verify absolutely the validity of any multiple imputation model (or, of course, any statistical model). In sum, these methods, like all others, depend on assumptions that, if wrong, can lead the analyst astray, so careful thought should always go into the application of these assumptions.

DISADVANTAGES OF LISTWISE DELETION

Whenever it is possible to predict the probability that a cell in a data matrix is missing (using D_{obs} or D_{mis}), the MCAR assumption is violated, and listwise deletion may generate biased parameter estimates. For example, listwise deletion can bias conclusions if those who think of themselves as independents are less likely to respond to a party identification question, or if better educated people tend to answer issue opinion questions, or if less knowledgeable voters are less likely to reveal their voting preferences. These patterns might each be MAR or NI, but they are not MCAR. Listwise deletion can result in different magnitudes or signs of causal or descriptive inferences (Anderson, Basilevsky, and Hum 1983). It does not always have such harmful effects; sometimes the fraction of missing observations is small or the assumptions hold sufficiently well so that the bias is not large.

In this section, we quantify the efficiency loss due to listwise deletion under the optimistic MCAR assumption, so that no bias exists. We consider 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. One approach might be to regress Y on X_1, but most scholars would control for a list of potential confounding influences, variables we label X_2. As critics we use omitted variables as the first line of attack, and as authors we know that controlling for more variables helps protect us from potential criticism; from this perspective, the more variables in X_2 the better.

The goal is to estimate β_1 in the 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, controlling will reduce bias but may increase the variance of β_1 (since estimating additional parameters puts more demands on the data). Thus, the mean square error (a combination of bias and variance) may in some cases increase by including a control variable (Goldberger 1991, 256). Fortunately, since we typically have a large number of observations, adding an extra variable does not do much harm so long as it does not introduce substantial colinearity, and we often include X_2.

[6] Missingness can also be NI if the parameters of the process that generate D are not distinct from those that generate M, even if it is otherwise MAR. In the text, for expository simplicity, we assume that if a data set meets the MAR assumption, it also meets the distinctness condition and is therefore ignorable.

The tradeoff between bias and variance looms larger when data are missing. Missing data will normally be present in Y, X_1, and X_2, but suppose for simplicity 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:

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 data are missing ($0 < \lambda < 1$), most analysts consider only two estimators:

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 (b_1^O) risks bias, and the listwise deletion estimator (b_1^L) risks inefficiency (and bias except in the "best" case in which MCAR holds). Presumably because the risks of omitted variable bias are better known than the risks of listwise deletion, when confronted with this choice most scholars opt for listwise deletion. We quantify these risks with a formal proof in Appendix B and discuss the results here. If MSE(a) is the mean square error for estimator a, then the difference $\text{MSE}(b_1^L) - \text{MSE}(b_1^O)$ is how we assess which method is better. When this difference is positive, b_1^O has lower mean square error and is therefore better than b_1^L; when it is negative, b_1^L is better. The problem 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:

$$\text{MSE}(b_1^L) - \text{MSE}(b_1^O) = \frac{\lambda}{1-\lambda}\,\text{V}(b_1^I) + F[\text{V}(b_2^I) - \beta_2\beta_2']F'. \quad \textbf{(1)}$$

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 key here is the first term, which is the extra mean square error due to listwise deletion. Because this first term is always positive, it causes the comparison between the two estimators to tilt farther away from listwise deletion as the fraction of missing data (λ) grows.

To better understand equation 1, we estimate the average λ value in political science articles. Because of the bias-variance tradeoff, those who try to fend off more possible alternative explanations have more control variables and thus larger fractions of observations lost. Although, on average, slightly less than one-third of observations are lost when listwise deletion is used,[7] the proportion can be much higher. In the papers and posters presented at the 1997 annual meeting of the Society for Political Methodology, for example, the figure exceeded 50% on average and in some cases was more than 90%.[8] Because scholars usually drop some variables to avoid extreme cases of missingness, the "right" value of λ for our purposes is larger than the observed fraction. We thus study the consequences of setting $\lambda = 1/2$, which means the first term in equation 1 reduces to $\text{V}(b_1^I)$. The MSE also depends on the second term, which can be positive or negative depending on the application. For simplicity, consider the case in which this second term is zero (such as when $V(b_2^I) = \beta_2\beta_2'$, or X_1 and X_2 are uncorrelated). Finally, we take the square root of the MSE difference to put it in the interpretable units of the average degree of error. The result is that the average error difference is $\text{SE}(b_1^I)$, the standard error of b_1^I.

If these assumptions are reasonable, then the point estimate in the average political science article is about one standard error farther away from the truth because of listwise deletion (as compared to omitting X_2 entirely). This is half the distance from no effect to what usually is termed "statistically significant" (i.e., two standard errors from zero).[9] Of course, this is the average absolute error: Point estimates in some articles will be too high, in others too low. In addition, we are using the standard error here as a metric to abstract across applications with different meanings, but in any one application the meaning of the expression depends on how large the standard error is relative to changes in the variables. 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 as the sample size increases.

Although social scientists rarely choose it, omitted variable bias is often preferable, if only it and listwise deletion are the options. One cannot avoid missing value problems since they usually affect all variables rather than only potential control variables. Moreover, because this result relies on the optimistic MCAR assumption, the degree of error will often be more than one standard error, and its direction will vary as a function of the application, pattern of missingness, and model estimated (Globetti 1997; Sherman 2000). Fortunately, better methods make this forced choice between suboptimal procedures unnecessary.

A METHOD FOR ANALYZING INCOMPLETE DATA

We now describe a general definition of multiple imputation, a specific model for generating the impu-

[7] This estimate is based on our content analysis of five years of the *American Political Science Review*, the *American Journal of Political Science*, and the *British Journal of Political Science*.

[8] This estimate is based on 13 presented papers and more than 20 posters.

[9] This is one of the infeasible estimator's standard errors, which is 71% of the listwise deletion estimator's standard error (or, in general, $\sqrt{\lambda} \times \text{SE}(b_1^L)$). Calculated standard errors are correct under MCAR but larger than those for better estimators given the same data, and they are wrong if MCAR does not hold.

tations, and the existing computational algorithms and our alternative. We also make several theoretical clarifications and consider potential problems.

Definition of Multiple Imputation

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 uncertainty levels. That is, for missing cells the model predicts well, variation across the imputations is small; for other cases, the variation may be larger, or asymmetric, to reflect whatever knowledge and level of certainty is available about the missing information. Analysts can then conveniently apply the statistical method they would have used if there were no missing values to each of the m data sets, and use a simple procedure that we now describe to combine the m results. As we explain below, m can be as small as 5 or 10.

First estimate some *Q*uantity of interest, Q, such as a univariate mean, regression coefficient, predicted probability, or first difference in each data set j ($j = 1, \ldots, m$). The overall point estimate $\bar{q}$ of Q is the average of the m separate estimates, q_j:

$$\bar{q} = \frac{1}{m}\sum_{j=1}^{m} q_j. \qquad (2)$$

Let $\mathrm{SE}(q_j)$ denote the estimated standard error of q_j from data set j, and let $S_q^2 = \Sigma_{j=1}^{m}(q_j - \bar{q})^2/(m-1)$ be the sample variance across the m point estimates. Then, as shown by Rubin (1987a), the variance of the multiple imputation 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 (multiplied by a factor that corrects for bias because $m < \infty$):

$$\mathrm{SE}(q)^2 = \frac{1}{m}\sum_{j=1}^{m}\mathrm{SE}(q_j)^2 + S_q^2(1 + 1/m). \qquad (3)$$

If, instead of point estimates and standard errors, simulations of q are desired, we create $1/m$th the needed number from each completed data set (following the usual procedures; see King, Tomz, and Wittenberg 2000) and combine them into one set of simulations.

An Imputation Model

Implementing multiple imputation requires a statistical model from which to compute the m imputations for each missing value in a data set. Our approach assumes that the data are MAR, conditional on the imputation model. The literature on multiple imputation suggests that in practice most data sets include sufficient information so that the additional outside information in an application-specific NI model (see Appendix A) will not add much and may be outweighed by the costs of nonrobustness and difficulty of use (Rubin 1996; Schafer 1997). Although this is surely not true in every application, the advantages make this approach an attractive option for a wide range of potential uses. The MAR assumption can also be made more realistic by including more informative variables and information in the imputation process, about which more below. Finally, note that the purpose of an imputation model is to create predictions for the distribution of each of the missing values, not causal explanation or parameter interpretation.

One model that has proven useful for missing data problems in a surprisingly wide variety of situations assumes that the variables are jointly multivariate normal. This model obviously is an approximation, as few data sets have variables that are all continuous and unbounded, much less multivariate normal. Yet, many researchers have found that it works as well as more complicated alternatives specially designed for categorical or mixed data (Ezzati-Rice et al. 1995; Graham and Schafer 1999; Rubin and Schenker 1986; Schafer 1997; Schafer and Olsen 1998). Transformations and other procedures can be used to improve the fit of the model.[10] For our purposes, if there exists information in the observed data that can be used to predict the missing data, then multiple imputations from this normal model will almost always dominate current practice. Therefore, we discuss only this model, although the algorithms we discuss might also work for some of the more specialized models as well.

For observation i ($i = 1, \ldots, n$), 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 off-diagonal elements of Σ allow variables within D to depend on one another. The likelihood function for complete data is:

$$L(\mu, \Sigma|D) \propto \prod_{i=1}^{n} N(D_i|\mu, \Sigma). \qquad (4)$$

By assuming the data are MAR, we form the observed data likelihood. The procedure is exactly as for application-specific methods (equations 12–13 in Appendix A, where with the addition of a prior this likelihood is proportional to $P(D_{\mathrm{obs}}|\theta)$). We denote $D_{i,\mathrm{obs}}$ as the observed elements of row i of D, and $\mu_{i,\mathrm{obs}}$ and $\Sigma_{i,\mathrm{obs}}$ as the corresponding subvector and submatrix of μ and Σ (which do not vary over i), respectively. Then, because the marginal densities are normal, the observed data likelihood is

$$L(\mu, \Sigma|D_{\mathrm{obs}}) \propto \prod_{i=1}^{n} N(D_{i,\mathrm{obs}}|\mu_{i,\mathrm{obs}}, \Sigma_{i,\mathrm{obs}}). \qquad (5)$$

The changing compositions of $D_{i,\mathrm{obs}}$, $\mu_{i,\mathrm{obs}}$, and $\Sigma_{i,\mathrm{obs}}$ over i make this a complicated expression to evaluate,

[10] Most variables in political science surveys are ordinal variables with four to seven values, which are reasonably well approximated by the normal model, at least for the purpose of making imputations.

although for clarity of presentation we have omitted several computational conveniences that can help (see Schafer 1997, 16).[11]

The multivariate normal specification implies that the missing values are imputed linearly. Thus, we create an imputed value the way we would usually simulate from a regression. For example, let $\tilde{D}_{ij}$ denote a simulated value for observation i and variable j, and let $D_{i,-j}$ denote the vector of values of all observed variables in row i, except variable j. The coefficient β from a regression of D_j on the variables in D_{-j} can be calculated directly from elements of μ and Σ, since they contain all available information in the data under this model. Then we use this equation to create an imputation:

$$\tilde{D}_{ij} = D_{i,-j}\tilde{\beta} + \tilde{\epsilon}_i, \quad \textbf{(6)}$$

where $\sim$ indicates a random draw from the appropriate posterior. Thus, random draws of $\tilde{D}_{ij}$ are linear functions of the other variables whenever they are observed $D_{i,-j}$, of estimation uncertainty due to not knowing β (i.e., μ and Σ) exactly, and of fundamental uncertainty $\tilde{\epsilon}_i$ (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. The computational difficulty is taking random draws from the posterior of μ and Σ.

Equation 6 can 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 draws from a multinominal or other appropriate discrete distribution with mean equal to the normal imputation. For example, to impute a 0/1 variable, take a Bernoulli draw 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, 0 otherwise.

Computational Algorithms

Computing the observed data likelihood in equation 5, and taking random draws from it, is computationally infeasible with classical methods. Even maximizing the function takes inordinately long with standard optimization routines. In response to such difficulties, the *I*mputation-*P*osterior (IP) and *E*xpectation-*M*aximization (EM) algorithms were devised and subsequently applied to this problem.[12] From the perspective of statisticians, IP is now the gold standard of algorithms for multivariate normal multiple imputations, in large part because it can be adapted to numerous specialized models. Unfortunately, from the perspective of users, it is slow and hard to use. Because IP is based on Markov Chain Monte Carlo (MCMC) methods, considerable expertise is needed to judge convergence, and there is no agreement among experts about this except for special cases. IP has the additional problem of giving dependent draws, so we need adaptations because multiple imputation requires that draws be independent. In contrast, EM is a fast deterministic algorithm for finding the maximum of the likelihood function, but it does not yield the rest of the distribution. We outline these algorithms and refer the reader to Schafer (1997) for a clear presentation of the computational details and historical development.

We also will discuss two additional algorithms, which we call EMs (EM with *s*ampling) and EMis (EM with importance resampling), respectively. Our recommended procedure, EMis, is quite practical: It gives draws from the same posterior distribution as IP but is considerably faster, and, for this model, there appear to be no convergence or independence difficulties. Both EMs and EMis are made up of standard parts and have been applied to many problems outside the missing data context. For missing data problems, EMs has been used, and versions of EMis have been used for specialized applications (e.g., Clogg et al. 1991). EMis also may have been used for problems with general patterns of missingness, 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 widespread potential (see Appendix C for information about software we have developed).

IP. A version of the data augmentation algorithm of Tanner and Wong (1987), IP enables us to draw random simulations from the multivariate normal observed data posterior $P(D_{\text{mis}}|D_{\text{obs}})$ (see Schafer 1997, 72ff). The basic idea is that drawing directly from this distribution is difficult, but "augmenting" it by conditioning on additional information makes the problem 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:

$$\tilde{D}_{\text{mis}} \sim P(D_{\text{mis}}|D_{\text{obs}}, \tilde{\mu}, \tilde{\Sigma}). \quad \textbf{(7)}$$

On the first application of equation 7, 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 the present imputed values for the missing data. This is called the posterior step:

$$\tilde{\mu}, \tilde{\Sigma} \sim P(\mu, \Sigma|D_{\text{obs}}, \tilde{D}_{\text{mis}}). \quad \textbf{(8)}$$

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

The advantage of IP is that the distributions are exact, but convergence to these distributions is known to occur only after an infinite number of iterations. The belief is that after a suitably long "burn-in period" (a number of iterations that are performed and discarded before continuing), perhaps recognizable by various

[11] Since the number of parameters $p(p + 3)/2$ increases rapidly with the number of variables p, priors help avoid overfitting and numerical instability in all the algorithms discussed here.

[12] Gelman et al. (1995), Jackman (2000), McLachlan and Krishan (1997), and Tanner (1996) provide excellent introductions to the literature on these algorithms and on Bayesian methods more generally.

diagnostics, convergence will have occurred, after which additional draws will come from the posterior. Unfortunately, experts disagree about how to assess convergence of this and other MCMC methods (Cowles and Carlin 1996; Kass et al. 1998).

In order to use the relatively simple equations 2 and 3 in combining the separate multiply imputed analyses, imputations must be statistically independent, but this is not a characteristic of successive draws from Markov chain methods such as IP. Some scholars reduce dependence by using every *r*th 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 (1992), recommends addressing both problems by creating one independent chain for each of the *m* desired imputations, with starting values drawn randomly from an overdispersed approximation distribution. The difficulty with taking every *r*th draw from one chain is the interpretation of autocorrelation functions (which requires analysts of cross-sectional data to be familiar with time-series methods). The difficulty of running separate chains is that the increase in run time, due to the need to burn in iterations to ensure convergence for each chain, is typically greater than the *m* times *r* iterations saved by not needing multiple draws from any one chain.

EM. The EM algorithm (Dempster, Laird, and Rubin 1977; McLachlan and Krishnan 1996; Orchard and Woodbury 1972) works like IP except that random draws from the entire posterior are replaced with deterministic calculations of posterior means. The draw of $\tilde{D}_{\text{mis}}$ in equation 7 is replaced with each missing cell's predicted value. The random draw of $\tilde{\mu}$ and $\tilde{\Sigma}$ in equation 8 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, reestimating 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 (relative to other options), it converges deterministically, and the objective function increases with every iteration. Like every numerical optimization algorithm, EM can sometimes settle on a local maximum, and for some problems convergence is slow, although these do not seem to be insurmountable barriers in most political science data. The more serious disadvantage of EM is that it yields only maximum values, rather than the entire density. It is possible to use EM to produce multiple imputations by treating point estimates of μ and Σ as if they were known with certainty. This means that estimation uncertainty is ignored, but the fundamental variability is included in the imputations. EM for multiple imputation works reasonably well in some instances, but ignoring estimation uncertainty means its standard errors are generally biased downward, and point estimates for some quantities will be biased.

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 it is difficult to draw from the posterior of μ and Σ. We approach this problem in two different ways. In this section, we use the asymptotic approximation (e.g., Tanner 1996, 54–9), which we find works as expected—well in large data sets and poorly in small ones.

To create imputations with this method, which we denote EMs, we first run EM to find the maximum posterior estimates of the parameters, $\hat{\theta} = \text{vec}(\hat{\mu}, \hat{\Sigma})$ (where the vec($\cdot$) operator stacks the unique elements). Then we compute the variance matrix, $V(\hat{\theta})$.[13] Next we draw a simulated θ from a normal with mean $\hat{\theta}$ and variance $V(\hat{\theta})$. From this, we compute $\tilde{\beta}$ deterministically, simulate $\tilde{\epsilon}$ from the normal, and substitute these values into equation 6 to generate an imputation. The entire procedure after the EM step and variance computation is repeated *m* times for the necessary imputations.

EMs is very fast, produces independent imputations, converges nonstochastically, and works well in large samples. For small samples, for data with many variables relative to the number of observations, or for 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.

EMis. EM finds the mode well, and EMs works well for creating fast and independent imputations in large samples, but it performs poorly with small samples or many parameters. We can improve EMs with a round of importance resampling (or "sampling importance/resampling"), an iterative simulation technique not based on Markov chains, to enhance small sample performance (Gelfand and Smith 1990; Gelman et al. 1995; Rubin 1987a, 192–4, 1987b; Tanner 1996; Wei and Tanner 1990).

EMis follows the same steps as EMs except that draws of θ from its asymptotic distribution are treated only as first approximations to the true (finite sample) posterior. We also put the parameters on unbounded scales, using the log for the standard deviations and Fisher's *z* for the correlations, to make the normal approximation work better with smaller sample sizes. 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 discarding the rest. Without priors, the importance ratio is

[13] To compute the variance matrix, we generally use the outer product gradient because of its speed. Other options are the inverse of the negative 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 estimator, which is more robust but slower. We have also developed an iterative simulation-based method that seems advantageous in speed and numerical stability when *p* is large.

$$\text{IR} = \frac{L(\tilde{\theta}|D_{\text{obs}})}{N(\tilde{\theta}|\tilde{\theta}, V(\tilde{\theta}))}. \quad (9)$$

We find that the normal approximation is usually good enough even in small, nonnormal samples so that the algorithm operates quickly.[14] In the final step, these draws of $\tilde{\theta}$ are used with equation 6 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. It is fast, does not rely on Markov chains, and produces the required fully independent imputations. Importance resampling, on which EMis is based, does not work well for all likelihood functions, especially when the normal density is not a good first approximation; for the present likelihood, however, our extensive experimentation with a wide variety of data types has not revealed any systematic differences when compared to runs of IP with immense numbers of iterations (so that judging MCMC convergence of IP is not as much of an issue). Our software includes the full range of standard diagnostics in case a problem arises that we have not foreseen. It also includes other approaches (IP, EM, EMs, and others), since our suggestion for improving methodological practice in political science is *not* to rely exclusively on EMis. Rather, we argue that any appropriately applied multiple imputation algorithm will generally outperform current incomplete data analysis practices.

Theoretical Clarifications and Common Misconceptions

It has been shown that multiple imputation inferences are statistically valid from both Bayesian and frequentist perspectives (Brownstone 1991; Meng 1994a; Rubin 1987a, 1996; Schafer 1997; Schenker and Welsh 1988). 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 result, for some chosen quantity Q to be estimated, involves approximating the correct posterior $P(Q|D_{\text{obs}})$. We would get this from an optimal application-specific method, with an approach based on the "completed" data $P(Q|D_{\text{obs}}, \tilde{D}_{\text{mis}})$, that is filled in with imputations $\tilde{D}_{\text{mis}}$ drawn from the conditional predictive density of the missing data $P(D_{\text{mis}}|D_{\text{obs}})$. Under MAR, we know that averaging $P(Q|D_{\text{obs}}, \tilde{D}_{\text{mis}})$ over $\tilde{D}_{\text{mis}}$ gives exactly $P(Q|D_{\text{obs}})$:

$$P(Q|D_{\text{obs}}) = \int P(Q|D_{\text{obs}}, D_{\text{mis}})P(D_{\text{mis}}|D_{\text{obs}})dD_{\text{mis}}. \quad (10)$$

This integral can be approximated with simulation. To draw a random value of $\tilde{Q}$ from $P(Q|D_{\text{obs}})$, draw independent random imputations of $\tilde{D}_{\text{mis}}$ from $P(D_{\text{mis}}|D_{\text{obs}})$, and then draw Q conveniently from $P(Q|D_{\text{obs}}, \tilde{D}_{\text{mis}})$, given the imputed $\tilde{D}_{\text{mis}}$. We can approximate $P(Q|D_{\text{obs}})$ or any point estimate based on it to any degree of accuracy with a large enough number of simulations. This shows that if the complete-data estimator is consistent and produces accurate confidence interval coverage, then multiple imputation based on $m = \infty$ is consistent, and its confidence intervals are accurate.

Multiple imputation is feasible because the efficiency of estimators based on the procedure increases rapidly with m (see Rubin 1987a and the citations in Meng 1994a; and especially Wang and Robins 1998). Indeed, the relative efficiency of estimators with m as low as 5 or 10 is nearly the same as with $m = \infty$, unless missingness is exceptionally high.

Multiple imputation is made widely applicable by Meng's (1994a) results regarding an imputation model that differs from the analysis model used. He finds that so long as the imputation model includes all the variables (and information) in the analysis model, no bias is introduced; nominal confidence interval coverage will be at least as great as actual coverage and equal when the two models coincide (Fay 1992). Robins and Wang (2000) indicate, however, that multiple imputation confidence intervals are not always conservative when there is misspecification of either both the imputation and analysis model or just the latter. (The next section considers in greater depth what can go wrong with analyses using multiple imputation.)[15]

In summary, even with a very small m and an imputation model that differs from the analysis model, this convenient procedure gives a good approximation to the optimal posterior distribution, $P(Q|D_{\text{obs}})$. 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, no assumptions about causal ordering are required in making imputations: The use of 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. Simi-

[14] For difficult cases, our software allows the user to substitute the heavier tailed t for the approximating density. The normal or t with a larger variance matrix, scaled up by some additional factor (1.1–1.5 to work well), can also help.

[15] When the information content is greater in the imputation than analysis model, multiple imputation is more efficient than even the "optimal" application-specific method. This is the so-called super-efficiency property (Rubin 1996). For example, suppose we want to run 20 cross-sectional regressions with the same variables measured in different years, and we discover an additional control variable for each that strongly predicts the dependent variable but on average across the set correlates at zero with the key causal indicator. Excluding this control variable will only bias the causal estimate, on average, if it is a consequence of the causal variable, whereas including it will substantially increase the statistical efficiency of all the regressions. Unfortunately, an application-specific approach would need to exclude such a variable if it were a consequence of the key causal variable to avoid bias and would thus give up the potential efficiency gains. A multiple imputation analysis could include this variable no matter what its causal status, so statistical efficiency would increase beyond an application-specific approach.

larly, 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; again, the joint distribution is unchanged. Since the multiple imputation procedure taken as a whole approximates $P(Q|D_{obs})$, these "intuitions" based on parts of the procedure are invalid (see Schafer 1997, 105ff).[16]

WHAT CAN GO WRONG?

We first discuss common fixable stumbling blocks in the application of EMis and multiple imputation. We then consider the one situation in which listwise deletion would be preferable to multiple imputation, as well as situations in which application-specific approaches would sufficiently outperform multiple imputation to be preferable.

Practical Suggestions

As with any statistical approach, if the model-based estimates of EMis are wrong, then there are circumstances in which 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 misled with any statistical method that assumes otherwise. In the present context, our method assumes that the observed data can be used to predict 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 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.[17] Of course, in this case listwise deletion will be at least as bad, since it generally discards more observed information than EMis has to impute, and it is biased unless strong assumptions about the missing data apply.

An advantage of our approach over application-specific methods (see Appendix A) is that it is often robust to errors in the imputation model, since (as with the otherwise inferior single imputation models; see Appendix A) separating the imputation and analysis stages means that errors in the missingness model can have no effect on observed parts of the data set, because 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 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 EMis is to include information in the analysis model and omit it from the imputation model. For example, if a variable is excluded from the imputation model but used in the analysis, estimates of the relationship between this variable and others will be biased toward zero. As a general rule, researchers should include in the imputation model all the variables from the analysis model. For greater efficiency, add any other variables in the data set that would help predict the missing values.[18]

The ability 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 because in effect we were 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 certainly would not want to control for voting intention five minutes before walking into the voting booth Z, since it is a consequence of party identification and so would incorrectly drive that variable's estimated effect to zero. Yet, Z would be a powerful predictor of the missing value of the vote variable, and the ability to include it in the imputation stage of a multiple imputation model and

[16] Because the imputation and analysis stages are separate, proponents of multiple imputation argue that 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 would be convenient for analysts and can be especially advantageous if the data provider can use confidential information in making the imputations that otherwise would not be available. The strategy is also convenient for those able to hire consultants to make the imputations for them. Others are not enthusiastic about this idea (even if they have 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 to which analysts may not even have access). The approach also is not feasible for large data sets, such as the National Election Studies, because existing computational algorithms cannot reliably handle so many variables, even in theory. Our alternative but 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.

[17] In this sense, the problem of missing data is theoretically more difficult than ecological inference, for example, since both involve filling in missing cells, but in missing data problems deterministic bounds on the unknown quantities cannot be computed. In practice, dealing with the missing data problem may be relatively easier since its assumption (that observed data will not drastically mislead in predicting the missing data) is very plausible in most applications.

[18] If the data are generated using a complex or multistage survey design, then information about the design should be included in the imputation model. For example, to account for stratified sampling, the imputation model should include the strata coded as dummy variables. Our software allows one to include these directly or to condition on them. The former requires no special programming. The latter, which we do by letting μ be a linear function of the dummy variables, is easy to implement because the dummies are fully observed, and many fewer parameters need to be estimated. Other possibilities for dealing with complex sampling designs include hierarchical Bayesian models, the general location model, and other fixed effects designs.

also omit it from 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 identification 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, but for the imputation model the entire set of issue preference questions should be included, because an observed value in one can be especially useful for predicting a missing value in another.

A similar information discrepancy occurs if the analysis model specifies a nonlinear relationship, since the imputation model is linear (see equation 6). There is little problem with the set of nonlinear functional forms typically used in the social sciences (logit, probit, exponential, and so on), because a linear approximation to these forms has been shown to perform very well during imputation, even if not for the analysis model. Yet, more severe nonlinearity, such as quadratic terms that are the central question being researched, can cause problems if ignored. 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 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 will be less efficient if one variable has much more missingness than another).

Researchers also should try to meet the distributional assumptions of the imputation model. For the imputation stage, variables should be transformed to be unbounded and relatively symmetric. For example, budget figures, which are positive and often positively skewed, can be logged. 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 symmetric.

Ordinal variables should be coded to be as close to an interval scaling as information indicates. For example, if categories of a variable measuring the degree of intensity of international conflicts are diplomatic dispute, economic sanctions, military skirmish, and all out war, a coding of 1, 2, 3, and 4 is not 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 without missing data.[19]

[19] Researchers with especially difficult combinations of nominal and continuous variables may want to consider implementing the general location imputation model (Schafer 1997).

Finally, NI missingness is always a serious concern because, by definition, it cannot be verified in the observed data. We discuss this issue in different ways in the sections to follow.

When Listwise Deletion Is Preferable

For listwise deletion to be preferable to EMis, all four of the following (sufficient) conditions must hold. (1) 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 the characteristic robustness of regression is not lost when applied to data with measurement error, endogeneity, nonlinearity, and so on). (2) There is NI missingness in X, so that EMis can give incorrect answers, and no Z variables are available that could be used in an imputation stage to fix the problem. (3) Missingness in X is not a function of Y, and unobserved omitted variables that affect Y do not exist. This ensures that the normally substantial advantages of our approach in this instance do not apply. (4) The number of observations left after listwise deletion should be so large that the efficiency loss from listwise deletion does not counterbalance (e.g., in a mean square error sense) the biases induced by the other conditions. This last condition does not hold in most political science applications except perhaps for exit polls and some nonsurvey data.

In other words, in order to prefer listwise deletion, we must have enough information about problems with our variables so that we do not trust them to impute the missing values in the X's—or we worry more about using available information to impute the X's than the possibility of selection on X as a function of Y in (3), which our approach would correct. Despite this, to use listwise deletion we must still trust the data enough 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 based on D_{obs}. Furthermore, we must have no extra variables Z to predict X or Y, and many observations must be left after listwise deletion.

If all of these conditions hold, listwise deletion can outperform EMis, and researchers should consider whether these might hold in their data. However, we feel this situation—using more information is worse—is likely to be rare. It is indeed difficult to think of a real research project that fits these conditions sufficiently so that listwise deletion would be knowingly preferable to EMis. Probably the best case that can be made for listwise deletion is convenience, although our software should help close the gap.

When Application-Specific Approaches Are Worth the Trouble

Although proponents of application-specific methods and of multiple imputation frequently debate the right approach to analyzing data with missing values, if a good application-specific approach is feasible, we believe it should be adopted. Such an approach not only

is better statistically but also by definition allows inclusion of more of the normally substantial qualitative knowledge available to social scientists but not recorded in the numerical data. It encourages researchers to explore features of their data suggested by this qualitative knowledge or revealed by preliminary data analyses, and more information is extracted. Unfortunately, these methods do not exist for all applications, are especially rare for missingness scattered throughout X and Y, can be technically demanding to create, and often are not robust when the chosen model does not fit the data well. The rich variety of methods now available should be studied by social scientists, and the literature should be followed for the many advances likely to come. But if no such method is available, when is a social scientist's effort best devoted to developing a new application-specific method? We identify four situations.

First, as discussed above, imputing values that do not exist makes little sense. Answers to survey questions that are "inconvenient" for the analyst, as when "no opinion" means that the respondent really has no opinion rather than prefers not to share information with the interviewer, should be treated seriously and modeled directly, like any other survey response. In this situation, virtually any general-purpose imputation method would bias the analysis model, and listwise deletion would be no better. An application-specific approach is necessary to model the specific process that generated the survey responses.

Second, when missingness is a function of $Y|X$ (even after controlling for extra variables in the imputation stage), the data are NI. For example, researchers should be suspicious that MAR might not hold in measures of the duration of parliamentary cabinets that are censored due to governments that are still in office at the time of data collection. If these missing values can be predicted from the remaining variables, then the data are still MAR, but this fact is unverifiable, and researchers should tread especially carefully in these circumstances. When NI is a strong possibility, substantial gains can sometimes be had with an application-specific approach. Even if the selection mechanism is not so severe, but is central to the research question, then development of an application-specific approach may be worth considering.

Third, whenever key information in the analysis model cannot be approximated within the imputation model, it may be desirable to develop an alternative. For example, if the analysis model contains severe nonlinearity or very complex interactions that cannot be incorporated into our linear imputation model, then it may be worth developing an application-specific approach. Neural network models provide one such example that cannot be handled easily within the EMis imputation stage (Bishop 1995).

Finally, extreme distributional divergences from multivariate normal can be a good reason to consider an alternative approach. Ordinal and dichotomous variables will often do well under EMis, but variables that are highly skewed (even after transformation) or a variable of primary interest that is mixed continuous and discrete may make it worth the trouble to develop an alternative.

MONTE CARLO EVIDENCE

In this section, we provide analyses based on simulated data: a timing test that reveals EMis is much faster 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 demonstrating that IP and EMis give the same answers, and these results are only slightly worse than if no data were missing and normally are far better than listwise deletion. (We have run many other Monte Carlo experiments to verify that the reported standard errors and confidence intervals, as well as estimates for other quantities of interest and different analysis models, are correct, but we omit these here.)

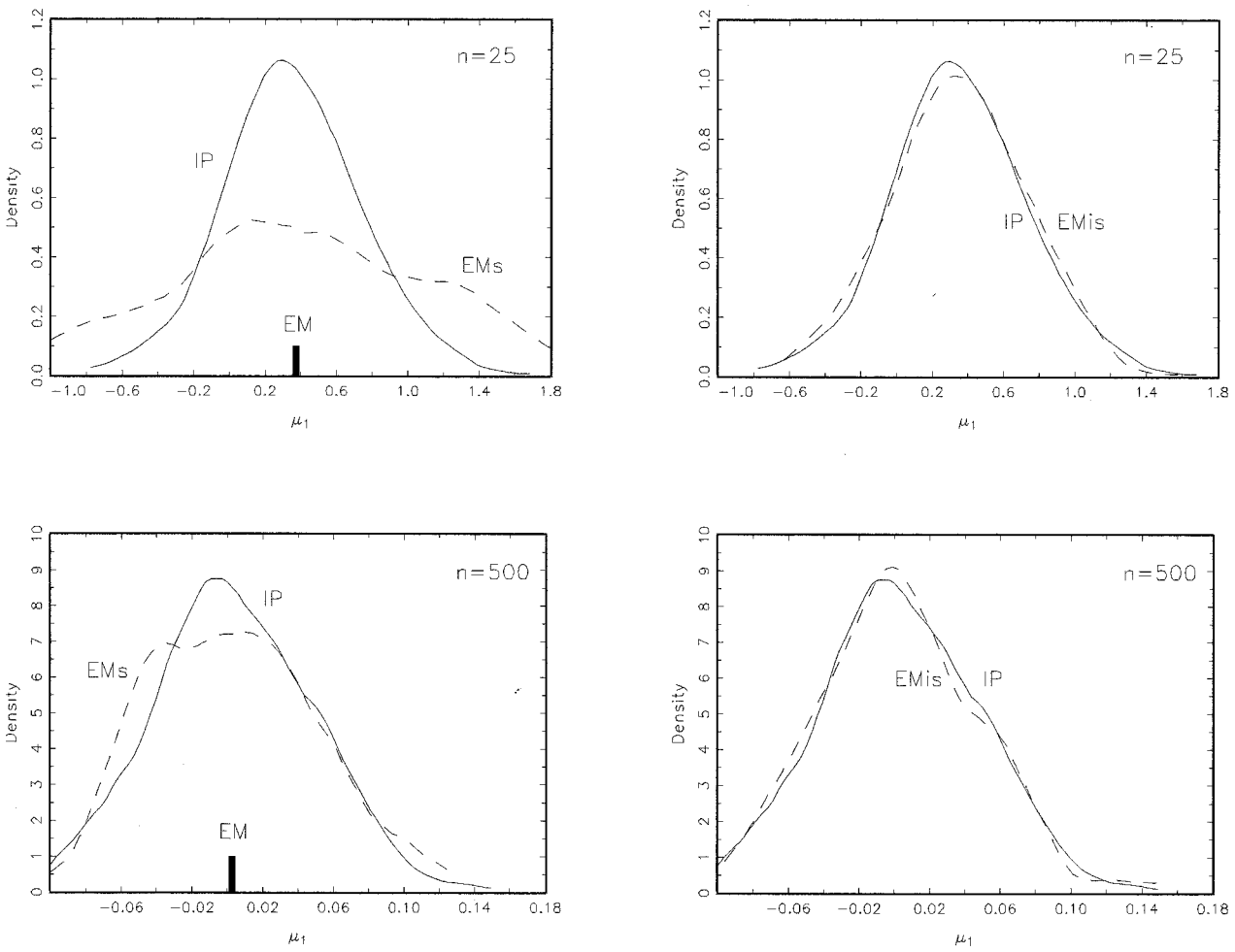
First, we compare the time it takes to run IP and EMis. Since imputation models are generally run once, followed by numerous analysis runs, imputation methods that take time are still useful. Runs of many hours, however, make productive analysis much less likely, especially if several data sets must be analyzed.

We made numerous IP and EMis runs, but it is not obvious how IP should be timed because there are no clear rules for judging convergence. We made educated guesses, ran experiments in which we knew the distribution to which IP was converging, studied profile plots of the likelihood function, and otherwise used Schafer's (1997) recommended defaults. On the basis of this experience, we chose $\max(1000, 100p)$ iterations to generate the timing numbers below, where p is the number of variables. For the EMis algorithm we chose a very conservative 1/50 ratio of draws to imputations. With each algorithm we created ten imputed data sets. We used a computer with average speed for 1999 (450Mhz with 128M of RAM). We then created a data set with 1,000 observations, of which 50 of these and one variable were fully observed. Every remaining cell was missing with 5% probability, which is not unusual for most social science survey data.

For 5 variables, IP takes 4.8 minutes, whereas EMis finishes in 3 seconds. For 10 variables, IP takes 28 minutes, and EMis runs for 14 seconds. With 20 variables, IP takes 6.2 hours, and EMis takes 2 minutes. With 40 variables, IP takes 3.5 days, whereas EMis runs for 36 minutes. Overall, EMis ranges from 96 to 185 times faster. Counting the analyst's time that is necessary to evaluate convergence plots would make these comparisons more dramatic.[20] Running one IP chain would be 2–3 times as fast as the recommended approach of separate chains, but that would require evaluating an additional $p(p + 3)/2$ autocorrelation

[20] Since convergence is determined by the worst converging parameter, one typically needs to monitor $p(p + 3)/2$ convergence plots. For applications in which the posterior is nearly normal, evaluating the worst linear function of the parameters can sometimes reduce the number of plots monitored. We also did not include the time it would take to create an overdispersed set of starting values for the IP chains.

FIGURE 1. Comparison of Posterior Distributions



Note: These graphs show, for one mean parameter, how the correct posterior (marked IP) is approximated poorly by EM, which only matches the mode, and EMs when n is small (top left). IP is approximated well by EMs for a larger n (bottom left) and by EMis for both sample sizes (right top and bottom).

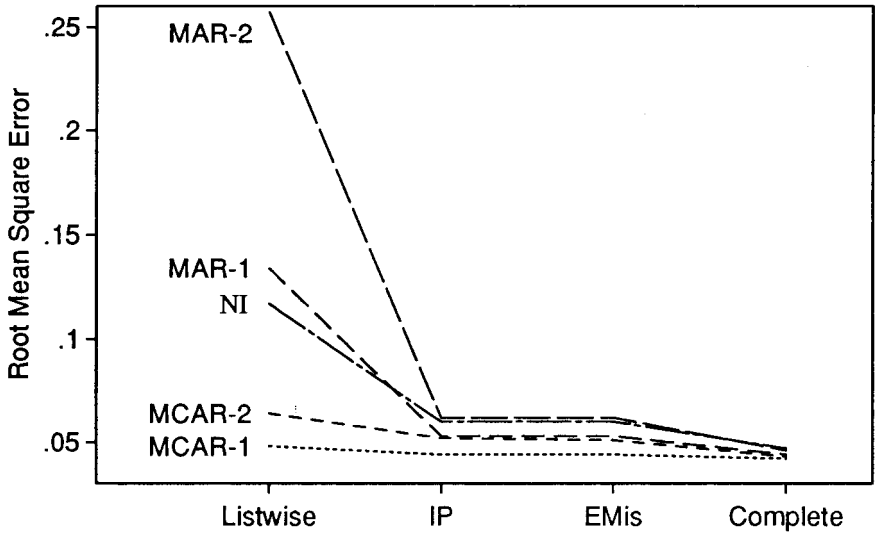
function plots to avoid creating dependent imputations.[21]

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 (see Figure 1). The first row of graphs is for $n = 25$, and the second row is for $n = 500$. The first column compares EMs and EM to IP, and the second compares EMis to IP. In all four graphs, the correct posterior, computed by IP, is a solid line. Clearly, the maximum likelihood point estimate found by EM (and marked by a small vertical bar on the left graphs) is not an adequate approximation to the entire posterior. By ignoring estimation variability, EM underestimates standard errors and confidence intervals.

The figure also enables us to evaluate 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 is not noticeably different when $n = 500$. (The

[21] We programmed both IP and EMis in the same language (GAUSS), which keeps them comparable to a degree. Our algorithm is more suited to the strengths of the GAUSS language. Additional vectorization will speed up both algorithms, but not necessarily in the same ratio. 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. Schafer's FORTRAN implementation of EM is about 25 times as fast as the EM portion of EMis. Similarly, the speed of our variance calculation could be substantially improved with complete vectorization. We use a FORTRAN implementation, as part of our GAUSS code, for calculating the likelihood in the importance sampling portion of the EMis algorithm, making the calculation of the likelihood fully vectorized. We do this because it is a calculation not well suited to GAUSS. Without this, our algorithm in GAUSS runs for 5 seconds, 52 seconds, 25 minutes, and 25 hours, respectively, or from 4 to 58 times faster than IP.

FIGURE 2. Root Mean Square Error Comparisons



Note: 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 that IP and EMis have the same root mean square error, which is lower than listwise deletion and higher than the complete data.

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.)

Finally, we generate data sets with different missingness characteristics and compare the mean square errors of the estimators. The Monte Carlo experiments we analyze here were representative of the many others we tried and are consistent with others in the literature. We generate 100 data sets randomly from each of five data generation processes, each with five variables, Y, $X_1, \ldots, X_4$.

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

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

MAR-1: Y and X_4 are MCAR: X_1 and X_2 are MAR, with missingness a function of X_3, which 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: Missing values in Y and X_2 depend on their observed and unobserved values; X_1 depends on the observed and unobserved values of X_3; and X_3 and X_4 are generated as MCAR. About 50% of rows are fully observed.[22]

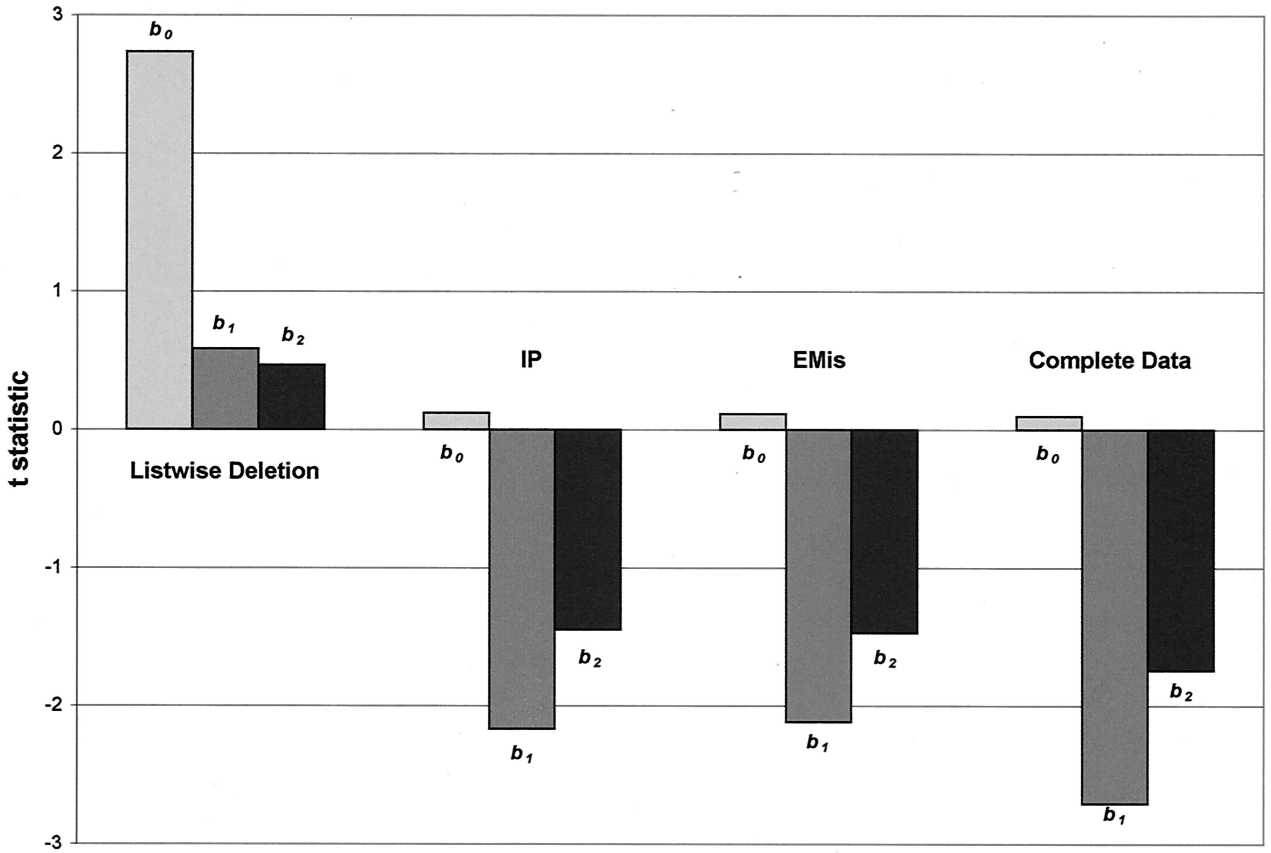
The quantities of interest are β_1 and β_2 in the regression $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$.[23] The Σ matrix is set so that β_1 and β_2 are each about 0.1. For each of the 100 data sets and five data-generation processes, we estimate these regression coefficients with imputation models based on listwise deletion, IP, and EMis as well as 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 above. We then computed the average root mean square error for the two coefficients in each run and 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 demonstrate three points. First, the root mean square error of EMis is virtually identical to that of IP for each data-generation process. This confirms again the equivalence of the two approaches.

[22] We drew $n = 500$ observations from a multivariate normal with means 0, variances 1, and correlation matrix {1 −.12 −.1 .5 .1, −.12 1 .1 −.6 .1, −.1 .1 1 −.5 .1, .5 −.6 −.5 1 .1, .1 .1 .1 .1 1}, where commas separate rows. For each missingness process, we created M as follows. Let row i and column j of M be denoted M_{ij}, and let u be a uniform random number. Recall that columns of M correspond to columns of $D = \{Y, X_1, \ldots, X_4\}$. For MCAR-1, if $u < 0.6$, then $M_{ij} = 1$, 0 otherwise. For MCAR-2, if $u < 0.19$, then $M_{ij} = 1$, 0 otherwise. For MAR-1, M_{i1} and M_{i5} were created as in MCAR-1; $M_{i4} = 0 \forall\ i$; and if $X_{i3} < -1$ and $u < 0.9$, then $M_{i2} = 1$ and (with a separate value of u) $M_{i3} = 1$, 0 otherwise. For MAR-2, M_{i1} and M_{i5} equal 0 if $u < 0.12$, 1 otherwise; $M_{i4} = 0 \forall\ i$; and if $X_{i3} < -0.4$ and $u < 0.9$, then $M_{i2} = 1$ and (with a separate value of u) $M_{i3} = 1$. For NI, $M_{i1} = 1$ if $Y_i < -0.95$; $M_{i2} = 1$ if $X_{i3} < -0.52$; $M_{i3} = 1$ if $X_{i2} > 0.48$; and M_{i4} and M_{i5} were created as in MCAR-1. In other runs, not reported, we changed every parameter, the generating density, and the analysis model, and our conclusions were very similar.

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

FIGURE 3. Monte Carlo Comparison of *t* Statistics



Note: *T* statistics are given for the constant (b_0) and the two regression coefficients (b_1, b_2) for the MAR-1 run in Figure 2. Listwise deletion gives the wrong results, whereas EMis and IP recover the relationships accurately.

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 ranges from slightly inferior to the two multiple 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. An open question is which articles in political science have large mean square errors like that for MAR-2 due to listwise deletion.

A further illustration of the results of our Monte Carlo study is provided in Figure 3, which gives a different view of the MAR-1 run in Figure 2. For MAR-1, the case of low missingness, 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 is the case 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. Only 22% of the observations were lost in this case, yet the key substantive conclusions are reversed by choosing an inferior method. It is easy to generate hypothetical data with larger effects, but this instance is probably closer to the risks we face.

EXAMPLES

We present two examples that demonstrate how switching from listwise deletion can markedly change substantive conclusions.

Voting Behavior in Russian Elections

The first example is vote choice in Russia's 1995 parliamentary election. Analyses of elections in Russia and emerging democracies generally present conflicting descriptions of individual voting behavior. In one view, electoral choice in these elections is thought to be chaotic at worst and personalistic at best. The alterna-

TABLE 2. First-Difference Effects on Voting in Russia

	Listwise Deletion	Multiple Imputation
Satisfaction with democracy	−.06 (.06)	−.10 (.04)
Opposition to the market economy	.08 (.08)	.12 (.05)
Trust in the Russian government	−.06 (.08)	−.12 (.04)

Source: Authors' reanalysis of data from Colton 2000.
Note: 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.

tive perspective is that voting decisions are based in predictable ways on measurable social, demographic, attitudinal, and economic variables (not unlike voters in more established democracies). Our analysis illustrates how inferences can be substantially improved by implementing the EMis algorithm.

We present only a simplified voting model, but detailed accounts of behavior in recent Russian elections are available (Brader and Tucker 2001; Colton 2000; Fish 1995; Miller, Reisinger, and Hesli 1998; White, Rose, and McAllister 1997; Whitefield and Evans 1996).[24] Using data from the Russian Election Study (Colton n.d.), we estimate a logit model with the dependent variable defined as 1 if the voter casts a ballot for the Communist Party of the Russian Federation (KPRF), 0 otherwise. With more than 22% of the popular vote, the KPRF was the plurality winner in the 1995 parliamentary elections, which makes understanding this vote essential to a correct interpretation of the election. The explanatory variables for our simple model vary according to the stage of the voter's decision-making process being tested, in order to avoid controlling for the consequences of key causal variables. Listwise deletion loses 36%, 56%, and 58% of the observations, respectively, in the three stages from which we use data.

Table 2 presents estimates of three first differences derived from our logit regressions for listwise deletion and EMis. First, we estimate the effect of a voter's satisfaction with democracy on the probability of supporting the KPRF. This is one measure of voters' assessments of current economic and political conditions in Russia. Voters more satisfied with democracy may be 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 is −0.06 with a relatively large standard error of 0.06, which for all practical purposes is no finding. In contrast, the EMis estimate is −0.10 with a standard error of 0.04. The unbiased and more efficient EMis estimate is nearly twice as large and is estimated much more precisely. As such, we can be relatively confident that voters highly satisfied with Russian democracy were about 10% less likely to support the KPRF, a finding not ascertainable with existing methods.

Issue opinions are another likely determinant of vote choice. In particular, are voters who oppose the transition to a market economy more likely than others to support the Communist Party? The answer seems obvious, but listwise deletion reveals little support for this hypothesis; again, the first-difference estimate is in the hypothesized direction but is only as large as its standard error (and thus not "significant" by any relevant standard). In contrast, the EMis estimate suggests that voters opposed to the transition were about 12% more likely to vote for the KPRF, with a very small standard error.

The final comparison that we report is the voting effect of trust in the Russian government. Positive evaluations should have had a negative influence on KPRF support in the 1995 Duma election. Again, listwise deletion detects no effect, but multiple imputation finds a precisely estimated twelve percentage point difference.

Table 2 presents only these three of the forty-six effects we estimated. Overall, we found substantively important changes in fully one-third of the estimates. Ten changed in importance as judged by traditional standards (from "statistically significant" to not, or the reverse, plus some substantively meaningful difference), and roughly five others increased or decreased sufficiently to alter the substantive interpretation of their effects.

Public Opinion about Racial Policies

The second example replicates the analysis by Alvarez and Brehm (1997) of the factors that explain Americans' racial policy preferences and the variance in those preferences. They use a heteroskedastic probit to model citizens preferences about racial policies in fair-housing laws, government set asides, taxes to benefit minority educational opportunities, and affirmative action in university admissions. Their explanatory variables are scales constructed to measure individual's core values or beliefs, such as individualism, authoritarianism, egalitarianism, and ideology. They also include scales measuring antiblack stereotypes, generic out-group dislike (proxied by anti-Semitism), and modern racism. The latter term is a subject of debate in the literature (Kinder 1986; Kinder and Sears 1981; McConahay 1986); proponents argue that there is "a form of racism that has replaced overt expressions of racial superiority" (Alvarez and Brehm 1997, 347), and it defines attitudes to racial policies and questions. This "symbolic or modern racism denotes a conjunction of antiblack affect with traditional American values, tak-

[24] We were alerted to the potential importance of missing data problems in this literature by Timothy Colton as he experimented with alternative strategies for his study, *Transitional Citizens: Voters and What Influences Them in the New Russia* (2000).

ing form in the sense that blacks are receiving more attention from government or other advantages than they deserve" (p. 350).[25]

Alvarez and Brehm employ a statistical model that explains with these variables not only the racial policy preferences of individuals but also the individual variability in responses. When variability is explained by the respondent's lack of political information, then it is considered to be caused by uncertainty, whereas if variability is explained by a conflict between "competing core values" or "incommensurable choices," then it is caused by ambivalence. They find that these preferences are not motivated by core values such as individualism, and so on, but are solely determined by a person's level of modern racism. The authors are more interested substantively in understanding what causes variability in response. They find that the "individual variability in attitudes toward racial policy stems from uncertainty" (Alvarez and Brehm 1997, 369) derived from a "lack of political information" (p. 370), not from a conflict of core values, such as individualism with egalitarianism. The same model shows variability in abortion policy preferences to be due to a conflict of core values (Alvarez and Brehm 1995), but variability in response on racial policy is due to a lack of political information. Therefore, better informed individuals might change their responses, which offers encouragement to advocates of education and debate about racial policy.

To tap core values, Alvarez and Brehm constructed "core belief scales" from responses to related feeling thermometers and agree/disagree measures. A missing value in any of the individual scale items caused the entire scale value for that observation to be treated as missing. This problem was severe, since listwise deletion would have eliminated more than half the observations.

For one of the scales—ideology—Alvarez and Brehm dealt with the missingness problem by replacing the scale (based on a question using the terms "liberal-conservative") with an alternate question if respondents refused to answer or did not know their ideology in the terms of the original question. The alternate question pressed the respondent to choose liberal or conservative, which Alvarez and Brehm coded as a neutral with a weak leaning to the side finally chosen. This is a clear case of unobserved data and the use of a reasonable but ad hoc imputation method.[26] If the question concerned party identification, a valid response might be "none," and this might not be a missing value, merely an awkward response for the analyst. Yet, although "ideological self-placement" may be legitimately missing, the self-placement question is considered to be at fault. The individual presumably has some ideological stance, no matter how uncertain, but is not willing or able to communicate it in the terminology of the survey question. Nevertheless, to press the respondent to choose and then guess how to code these values on the same scale as the original question risks attenuating the estimated relationships.[27]

Fortunately, use of the forcing question is unnecessary, since items on homelessness, poverty, taxes, and abortion can easily be used to predict the technical placement without shifting the responsibility to the respondent who does not understand, or has not thought about, our academic terminology. Indeed, bias seems to be a problem here, since in the Alvarez and Brehm analysis, ideology rarely has an effect. When we impute missing ideology scores from the other items, however, instead of using the alternate question, ideology becomes significant just over half the time, and the coefficients all increase in both the choice and the variance models (for all the dependent variables they used).

We apply EMis for the missing components of the scales to counter the problem of nonresponse with greater efficiency and less bias. We present first-difference results in the style of Alvarez and Brehm in Table 3. The first differences represent the change in probability of supporting an increase in taxation to provide educational opportunities to minorities when a particular variable is moved from its mean to its mean plus two standard deviations, as in Alvarez and Brehm.[28]

The main substantive finding, that variance in policy choice between respondents is driven by a lack of information rather than a conflict between the core values, still holds. In contrast, the secondary finding, which explains individual preferences and which contributes to the more mainstream and developed policy argument, is now reversed. Most important, individual racial policy choice now appears to be a broad function of many competing values, not just modern racism. An individual's level of authoritarianism, anti-Semitism, and egalitarianism as well as ideological position all strongly affect the probability that a person will support increased taxes for minority educational opportunities.

Finally, and quite important, the chi-square test reported at the bottom of Table 3 is insignificant under Alvarez and Brehm's original specification but is now

[25] Alvarez and Brehm measured modern racism with three questions relating to the amount of attention minorities are paid by government, anger that minorities are given special advantages in jobs and education, and anger about minority spokespersons complaining about discrimination.

[26] This procedure was made known to us, and other portions of the replication were made possible, when the authors provided us code from their original analysis, for which we are grateful.

[27] Consistent with the literature (e.g., Hinich and Munger 1994), we assume that ideology measures an individual's underlying policy preferences. If one assumes that people have at least some policy views, then they have an ideology, even if they are unwilling or unable to place themselves on an ideological scale. Alternative treatments, especially in the European context, view ideology as an exogenous orientation toward politics. Missingness in ideology in that framework might be treated very much like partisan identification.

[28] These results mirror those presented by Alvarez and Brehm (1997, 367) in their Table 3, column 3, rows 1–7. Similar effects are found in all the other rows and columns of their tables 3 and 4. Our replication using their methods on the original data does not match their results exactly, including the *N*, but the substantive findings of our replication of their methods and their results are almost entirely the same throughout tables 1–4 of the original work. We also include standard errors in the reporting of first differences in our presentation (King, Tomz, and Wittenberg 2000).

TABLE 3. Estimated First Differences of Core Beliefs

	Listwise Deletion	Multiple Imputation
Modern racism	−.495*	−.248*
	(.047)	(.046)
Individualism	.041	.005
	(.045)	(.047)
Antiblack	−.026	−.011
	(.047)	(.042)
Authoritarianism	.050	.068*
	(.045)	(.035)
Anti-Semitism	−.097	−.115*
	(.047)	(.045)
Egalitarianism	.201*	.236*
	(.049)	(.053)
Ideology	−.076	−.133*
	(.054)	(.063)
N	1,575	2,009
χ^2	8.46	11.21*
$p(\chi^2)$	.08	.02

Note: The dependent variable is support for an increase in taxation to support educational opportunities for minorities. The first column reports our calculation of first difference effects and standard errors for the substantive variables in the mean function, using the same data set (the 1991 Race and Politics Survey, collected by the Survey Research Center, University of California, Berkeley) used by Alvarez and Brehm (1997). (For details on the survey and availability information, see their note 1.) Although we followed the coding rules and other procedures given in their article as closely as possible, our analysis did not yield the same values reported by Alvarez and Brehm for the first difference effects. Even so, our listwise deletion results confirm the substantive conclusions they arrived at using this method of dealing with missing data. The second column is our reanalysis using EMis. Asterisks 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.

significant.[29] This test measures whether their sophisticated analysis model is superior to a simple probit 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, which implies that their methodological complications were superfluous. Our approach, however, rejects the simpler probit in favor of the more sophisticated model and explanation.[30]

[29] See Meng (1994b) and Meng and Rubin (1992) for procedures and theory for p values in multiply imputed data sets. We ran the entire multiple imputation analysis of $m = 10$ data sets 100 times, and this value never exceeded 0.038.

[30] Sometimes, of course, our approach will strengthen rather than reverse existing results. For example, we also reanalyzed Domínguez and McCann's (1996) study of Mexican elections and found that their main argument (voters focus primarily on the potential of the ruling party and viability of the opposition rather than specific issues) came through stronger under multiple imputation. We also found that several of the results on issue positions that Domínguez and McCann were forced to justify ignoring or attempting to explain away turned out to be artifacts of listwise deletion.

We also replicated Dalton, Beck, and Huckfeldt's (1998) analysis of partisan cues from newspaper editorials, which examined a merged data set of editorial content analyses and survey responses. Most missing data resulted from the authors' inability to content analyze the numerous newspapers that respondents reported reading. Because the survey variables contained little information useful for predicting content analyses that were not completed, an MCAR missingness mechanism could not be rejected, and the point estimates did not substantially change under EMis, although confidence intervals and standard errors were reduced. Since Dalton, Beck, and Huckfeldt'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, which likely would substantially improve the analysis.

CONCLUSION

For political scientists, almost any disciplined statistical model of multiple imputation would serve better than current practices. The threats to the validity of inferences from listwise deletion are of roughly the same magnitude as those from the much better known problems of omitted variable bias. We have emphasized the use of EMis for missing data problems in a survey context, but it is no less appropriate and needed in fields that are not survey based, such as international relations. Our method is much faster and far easier to use than existing multiple imputation methods, and it allows the usage of about 50% more information than is currently possible. Political scientists also can jettison the nearly universal but biased practice of making up the answers for some missing values. Although any statistical method can be fooled, including this one, and although we generally prefer application-specific methods when available, EMis normally will outperform current practices. Multiple imputation was designed to make statistical analysis easier for applied researchers, but the methods are so difficult to use that in the twenty years since the idea was put forward it has been applied by only a few of the most sophisticated statistical researchers. We hope EMis will bring this powerful idea to those who can put it to best use.

APPENDIX A. CURRENT APPROACHES

Available methods for analyzing data sets with item nonresponse can be divided into two approaches: application specific (statistically optimal but hard to use) and general purpose (easy to use and more widely applicable but statistically inadequate).

Application-Specific Approaches

Application-specific approaches usually assume MAR or NI. The most common examples are models for selection bias, such as truncation or censoring (Achen 1986; Amemiya 1985, chap. 10; Brehm 1993; Heckman 1976; King 1989, chap. 7; Winship and Mare 1992). Such models have the advantage of including all information in the estimation, but almost all allow missingness only in or related to Y rather than scattered throughout D.

When the assumptions hold, application-specific approaches are consistent and maximally efficient. In some cases, however, inferences from these models tend to be sensitive to small changes in specification (Stolzenberg and Relles 1990). Moreover, different models must be used for each type of application. As a result, with new types of data, application-specific approaches are most likely to be used by

those willing to devote more time to methodological matters.[31]

More formally, these approaches model D and M jointly and then factor the joint density into the marginal and conditional. One way to do this produces selection models, $P(D, M|\theta, \gamma) = P(D|\theta)P(M|D, \gamma)$, where $P(D|\theta)$ is the likelihood function when no data are missing (a function of θ, the parameter of interest), and $P(M|D, \gamma)$ is 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, averaging over the missing data yields the following likelihood:

$$P(D_{\text{obs}}, M|\theta, \gamma) = \int P(D|\theta)P(M|D, \gamma)dD_{\text{mis}}, \qquad \textbf{(11)}$$

where the integral is over elements of D_{mis} and is summation when discrete. If MAR is appropriate (i.e., D and M are stochastically independent), then equation 11 simplifies:

$$P(D_{\text{obs}}, M|\theta, \gamma) = P(D_{\text{obs}}|\theta)P(M|D_{\text{obs}}, \gamma). \qquad \textbf{(12)}$$

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.

Unlike multiple imputation models, application-specific approaches require specifying $P(M|D, \gamma)$, about which scholars often have no special interest or knowledge. Evaluating the integral in equation 11 can be difficult or impossible. Even with MAR and ignorability assumptions, maximizing $P(D_{\text{obs}}|\theta)$ can be computationally demanding, given its nonrectangular structure. When these problems are overcome, application-specific models are theoretically optimal, even though they can make data analyses difficult in practice. (Software that makes this easier includes Amos and Mx, but only for linear models and only assuming MAR.)

General Purpose Methods

General purpose approaches are 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 standard treatment that assumes the absence of missing data. General purpose methods other than listwise deletion include mean substitution (imputing the univariate mean of the observed observations), best guess imputation (common in political science), imputing a zero and then adding a dummy variable to control for the imputed value, pairwise deletion (which really only applies to covariance-based models), and hot deck imputation (imputing from 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 these techniques 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. It is worth noting that listwise deletion, despite the problems discussed above, does generate valid standard errors, which makes it preferable in an important way to approaches such as mean substitution and best guess imputation.

When 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 use the predicted values to impute the missing values. A more sophisticated version of this procedure can be used iteratively to fill in data sets with many variables missing. This procedure is not biased for certain quantities of interest, even assuming MAR, since it conditions on the observed data. Since the missing data are imputed on the regression line as if there were no error, however, 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 $\Pr(Y > 7)$). To assume that a statistical relationship is imperfect when observed but perfect when unobserved is optimistic, to say the least.

Finally, one general purpose approach developed recently is an imputation method that combines elements of the multiple imputation procedures presented in this article and the application-specific methods discussed above. Analysts generate one or more imputed data sets in the first step and then calculate estimates of the relevant quantity of interest and its variance using alternative formulas to equations 2 and 3 (Robins and Wang 2000; Wang and Robins 1998). Like application-specific methods, this approach is theoretically preferred to multiple imputation but requires different adjustments for each analysis model, and it is not currently available in commercial software packages. Since this approach can be more efficient than multiple imputation, and the computed variances are correct under several forms of misspecification, there is much to recommend it.

[31] For application-specific methods in political science, see Achen 1986; Berinsky 1997; Brehm 1993; Herron 1998; Katz and King 1999; King et al. 1990; Skalaban 1992; and Timpone 1998.

APPENDIX B. 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 by comparing their mean square errors (MSE). MSE is a measure of how close the distribution of the estimator $\hat{\theta}$ is concentrated around θ. More formally, $\text{MSE}(\hat{\theta}, \theta) = E[(\hat{\theta} - \theta)^2] = V(\hat{\theta}) + E(\hat{\theta} - \theta)E(\hat{\theta} - \theta)' = \text{variance} + \text{bias}^2$.

Estimators

Let $b^I = AY = (b_1^{I'}, b_2^{I'})'$, where $A = (X'X)^{-1}X'$. Then b_1^I is the *I*nfeasible estimator of β_1. Let $b_1^O = A_1Y$ be the *O*mitted variable bias estimator of β_1, where $A_1(X_1'X_1)^{-1}X_1'$. Finally, let $b^L = A^LY^L = (b_1^{L'}, b_2^{L'})'$, where $A^L = (X^{L'}X^L)^{-1}X^{L'}$, and where the superscript L denotes listwise deletion applied to X and Y. So b_1^L is the *L*istwise 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 factor of coefficients from a regression of a column of X_2 on X_1 so bias $(b_1^O) = F\beta_2$. If MCAR holds, then listwise deletion is also unbiased, $E(b^L) = E(A^LY^L) = A^LX^L\beta = \beta$, 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 $V(b_1^O) = V(b_1^I) - FV(b_2^I)F'$. Because $V(b^L) = V(A^L Y^L) = A^L\sigma^2 I A^{L'} = \sigma^2(X^{L'}X^L)^{-1}$, 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: $\text{MSE}(b_1^O) = V(b_1^I) + F[\beta_2\beta_2' - V(b_2^I)]F'$, and $\text{MSE}(b_1^L) = \sigma^2(Q^L)^{11}$.

Comparison

In order to evaluate when listwise deletion outperforms the omitted variable bias estimator, we compute the *d*ifference d in MSE:

$$d = \text{MSE}(b_1^L) - \text{MSE}(b_1^O) = [V(b_1^L) - V(b_1^I)] + F[V(b_2^I) - \beta_2\beta_2']F'. \quad \textbf{(13)}$$

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 13 is the usual bias-variance tradeoff, so our primary concern is with the first term. $V(b^I)[V(b^L)]^{-1} = \sigma^2(X^{L'}X^L + X'_{\text{mis}}X_{\text{mis}})^{-1}1/\sigma^2(X^{L'}X^L) = I - (X^{L'}X^L + X'_{\text{mis}}X_{\text{mis}})^{-1}(X'_{\text{mis}}X_{\text{mis}})$, where X_{mis} includes the rows of X deleted by listwise deletion (so that $X = \{X^L, X_{\text{mis}}\}$). Since exchangability among rows of X is implied by the MCAR assumption (or, equivalently, takes the expected value over sampling permutations), we write $(X^{L'}X^L + X'_{\text{mis}}X_{\text{mis}})^{-1}(X'_{\text{mis}}X_{\text{mis}}) = \lambda$, which implies $V(b_1^L) = V(b^I)/(1 - \lambda)$. This, by substitution into equation 13, yields and thus completes the proof of equation 1.

APPENDIX C. SOFTWARE

To implement our approach, we have written easy-to-use software, *Amelia: A Program for Missing Data* (Honaker et al. 1999). It has many features that extend the methods discussed here, such as special modules for high levels of missingness, small n's, high correlations, discrete variables, data sets with some fully observed covariates, compositional data (such as for multiparty voting), time-series data, time-series cross-sectional data, t distributed data (such as data with many outliers), and data with logical constraints. We intend to add other modules, and the code is open so that others can add modules themselves.

The program comes in two versions: for Windows and for GAUSS. Both implement the same key procedures. The Windows version requires a Windows-based operating system and no other commercial software, is menu oriented and thus has few startup costs, and includes some data input procedures not in the GAUSS version. The GAUSS version requires the commercial program (GAUSS for Unix 3.2.39 or later, or GAUSS for Windows NT/95 3.2.33 or later), runs on any computer hardware and operating system that runs the most recent version of GAUSS, is command oriented, and has some statistical options not in the Windows version. The software and detailed documentation are freely available at http://GKing.Harvard.Edu.

REFERENCES

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

Alvarez, R. Michael, and John Brehm. 1995. "American Ambivalence Towards Abortion Policy: A Heteroskedastic Probit Method for Assessing Conflicting Values." *American Journal of Political Science* 39 (November): 1055–82.

Alvarez, R. Michael, and John Brehm. 1997. "Are Americans Ambivalent Towards Racial Policies?" *American Journal of Political Science* 41 (April): 345–74.

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

Anderson, Andy B., Alexander Basilevsky, and Derek P.J. Hum. 1983. "Missing Data: A Review of the Literature." In *Handbook of Survey Research*, ed. Peter H. Rossi, James D. Wright, and Andy B. Anderson. New York: Academic Press. Pp. 415–94.

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

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

Berinsky, Adam. 1997. "Heterogeneity and Bias in Models of Vote Choice." Paper presented at the annual meetings of the Midwest Political Science Association, Chicago.

Bishop, C. M. 1995. *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

Brader, Ted, and Joshua Tucker. 2001. "The Emergence of Mass Partisanship in Russia, 1993–96." *American Journal of Political Science* 45 (1): 69–83.

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

Brownstone, David. 1991. "Multiple Imputations for Linear Regression Models." Technical Report MBS 91-37, Department of Mathematical Behavior Sciences, University of California, Irvine.

Clogg, Clifford C., Donald B. Rubin, Nathaniel Schenker, Bradley Schultz, and Lynn Weidman. 1991. "Multiple Imputation of Industry and Occupation Codes in Census Public-Use Samples Using Bayesian Logistic Regression." *Journal of the American Statistical Association* 86 (March): 68–78.

Colton, Timothy. 2000. *Transitional Citizens: Voters and What Influences Them in the New Russia*. Cambridge, MA: Harvard University Press.

Cowles, Mary Kathryn, and Bradley P. Carlin. 1996. "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review." *Journal of the American Statistical Association* 91 (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 (March): 111–26.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1997. "Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society*, Methodological Series B, 39: 1–38.

Domínguez, Jorge, and James A. McCann. 1996. *Democratizing Mexico: Public Opinion and Electoral Choice*. Baltimore, MD: Johns Hopkins University Press.

Ezzati-Rice, T. M., W. Johnson, M. Khare, R. J. A. Little, D. B. Rubin, and J. L. Schafer. 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*. Washington, DC: Bureau of the Census. Pp. 257–66.

Fay, Robert E. 1992. "When Are Inferences from Multiple Imputation Valid?" *Proceedings of the Survey Research Methods Section of the American Statistical Association* 81 (1): 227–32.

Fish, M. Steven. 1995. "The Advent of Multipartism in Russia, 1993–95." *Post Soviet Affairs* 11 (4): 340–83.

Franklin, Charles H. 1989. "Estimation across Data Sets: Two-Stage Auxiliary Instrumental Variables Estimation (2SAIV)." *Political Analysis* 1: 1–24.

Gelfand, A. E., and A. F. M. Smith. 1990. "Sampling-Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association* 85 (June): 398–409.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 1995. *Bayesian Data Analysis*. New York: Chapman and Hall.
Gelman, Andrew, and Donald B. Rubin. 1992. "Inference from Iterative Simulation Using Multiple Sentences." *Statistical Science* 7 (November): 457–72.
Gelman, Andrew, Gary King, and Chuanhai Lin. 1999. "Not Asked and Not Answered: Multiple Imputation for Multiple Surveys." *Journal of the American Statistical Association* 93 (September): 846–57; with comments by John Brehm, David R. Judkins, Robert L. Santos, and Joseph B. Kadane, and rejoinder by Gelman, King, and Liu, pp. 869–74.
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.
Goldberger, Arthur S. 1991. *A Course in Econometrics*. Cambridge, MA: Harvard University Press.
Graham, J. W., and J. L. Schafer. 1999. "On the Performance of Multiple Imputation for Multivariate Data with Small Sample Size." In *Statistical Strategies for Small Sample Research*, ed. Rick Hoyle. Thousand Oaks, CA: Sage.
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: 475–92.
Heitjan, Daniel. F. 1989. "Inference from Grouped Continuous Data: A Review." *Statistical Science* 4 (May): 164–79.
Herron, Michael C. 1998. "Voting, Abstention, and Individual Expectations in the 1992 Presidential Election." Paper presented at the annual meetings of the Midwest Political Science Association, Chicago.
Hinich, Melvin J., and Michael C. Munger. 1994. *Ideology and the Theory of Political Choice*. Ann Arbor: University of Michigan Press.
Honaker, James, Anne Joseph, Gary King, Kenneth Scheve, and Naunihal Singh. 1999. *Amelia: A Program for Missing Data*. Cambridge, MA: Harvard University. http://GKing.Harvard.edu (accessed December 11, 2000).
Jackman, Simon. 2000. "Estimation and Inference via Bayesian Simulation: An Introduction to Markov Chain Monte Carlo." *American Journal of Political Science* 44 (April): 375–404.
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* 52 (2): 93–100.
Katz, Jonathan, and Gary King. 1999. "A Statistical Model for Multiparty Electoral Data." *American Political Science Review* 93 (March): 15–32.
Kinder, Donald R. 1986. "The Continuing American Dilemma: White Resistance to Racial Change 40 Years after Myrdal." *Journal of Social Issues* 42 (2): 151–71.
Kinder, Donald R., and David O. Sears. 1981. "Prejudice and Politics: Symbolic Racism versus Racial Threats to the Good Life." *Journal of Personality and Social Psychology* 40 (3): 414–31.
King, Gary. 1989. *Unifying Political Methodology: The Likelihood Model of Statistical Inference*. 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 (August): 846–71.
King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44 (2): 341–55.
Li, K. H. 1988. "Imputation Using Markov Chains. *Journal of Statistical Computation and Simulation* 30 (1): 57–79.
Little, J. Rodrick. 1992. "Regression with Missing X's: A Review." *Journal of the American Statistical Association* 87 (December): 1227–37.
Little, J. Rodrick, and Donald Rubin. 1987. *Statistical Analysis with Missing Data*. New York: Wiley.
Little, J. Rodrick, and Donald Rubin. 1989. "The Analysis of Social Science Data with Missing Values." *Sociological Methods and Research* 18 (November): 292–326.
Little, J. Rodrick, and Nathaniel Schenker. 1995. "Missing Data." In *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, ed. Gerhard Arminger, Clifford C. Clogg, and Michael E. Sobel. New York: Plenum. Pp. 39–75.
Liu, Jun S., Wing Hung Wong, and Augustine Kong. 1994. "Covariance Structure of the Gibbs Sampler with Applications to the Comparisons of Estimators and Augmentation Schemes." *Biometrika* 81 (March): 27–40.
McConahay, John B. 1986. "Modern Racism, Ambivalence, and the Modern Racism Scale." In *Prejudice, Discrimination, and Racism: Theory and Research*, ed. John Dovidio and Samuel L. Gaertner. New York: Academic Press. Pp. 57–99.
McLachlan, Geoffrey J., and Thriyambakam Krishan. 1997. *The EM Algorithm and Extensions*. New York: Wiley.
Meng, X. L. 1994a. "Multiple-Imputation Inferences with Uncongenial Sources of Input." *Statistical Science* 9 (4): 538–73.
Meng, X. L. 1994b. "Posterior Predictive *p*-Values." *Annals of Statistics* 22 (September): 1142–60.
Meng, X. L., and Donald Rubin. 1992. "Performing Likelihood Ratio Tests with Multiply-Imputed Data Sets." *Biometrika* 79 (March): 103–11.
Miller, Arthur H., William M. Reisinger, and Vicki L. Hesli. 1998. "Leader Popularity and Party Development in Post-Soviet Russia." In *Elections and Voters in Post-Communist Russia*, ed. Matthew Wyman, Stephen White, and Sarah Oates. London: Edward Elgar. Pp. 100–35.
Orchard, T., and Woodbury, M. A. 1972. "A Missing Information Principle: Theory and Applications." In *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press. Pp. 697–715.
Raghunathan, T. E., and J. E. Grizzle. 1995. "A Split Questionnaire Survey Design." *Journal of the American Statistical Association* 90 (March): 54–63.
Robins, James, and Naisyin Wang. 2000. "Inference for Imputation Estimators." *Biometrika* 87 (March): 113–24.
Rubin, Donald. 1976. "Inference and Missing Data." *Biometrika* 63 (3): 581–92.
Rubin, Donald. 1977. "Formalizing Subjective Notions about the Effect of Nonrespondents in Sample Surveys." *Journal of the American Statistical Association* 72 (September): 538–43.
Rubin, Donald. 1987a. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
Rubin, Donald. 1987b. "A Noniterative Sampling/Importance Resampling Alternative to the Data Augmentation Algorithm for Creating a Few Imputations When Fractions of Missing Information Are Modest: The SIR Algorithm. Discussion of Tanner and Wong." *Journal of the American Statistical Association* 82 (June): 543–6.
Rubin, Donald. 1996. "Multiple Imputation after 18+ Years." *Journal of the American Statistical Association* 91 (June): 473–89.
Rubin, Donald B., and J. L. Schafer. 1990. "Efficiently Creating Multiple Imputations for Incomplete Multivariate Normal Data." In *Proceedings of the Statistical Computing Section of the American Statistical Association*. Pp. 83–8.
Rubin, Donald, and Nathaniel Schenker. 1986. "Multiple Imputation for Interval Estimation from Single Random Samples with Ignorable Nonresponse." *Journal of the American Statistical Association* 81 (June): 366–74.
Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. London: 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*. Washington, DC: Bureau of the Census. Pp. 459–87.
Schafer, Joseph L., and Maren K. Olsen. 1998. "Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective." *Multivariate Behavioral Research* 33 (4): 545–71.
Schenker, Nathaniel, and A. H. Welsh. 1988. "Asymptotic Results for Multiple Imputation. *Annals of Statistics* 16 (December): 1550–66.
Sherman, Robert P. 2000. "Tests of Certain Types of Ignorable Nonresponse in Surveys Subject to Item Nonresponse or Attrition." *American Journal of Political Science* 44 (2): 356–68.
Skalaban, Andrew. 1992. "Interstate Competition and State Strategies to Deregulate Interstate Banking 1982–1988." *Journal of Politics* 54 (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 (May): 395–415.

Tanner, Martin A. 1996. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 3d ed. New York: Springer-Verlag.

Tanner, M. A., and W. H. Wong. 1987. "The Calculation of Posterior Distributions by Data Augmentation." *Journal of the American Statistical Association* 82 (June): 528–50.

Timpone, Richard J. 1998. "Structure, Behavior, and Voter Turnout in the United States." *American Political Science Review* 92 (March): 145–58.

Wang, Naisyin, and James Robins. 1998. "Large-Sample Theory for Parametric Multiple Imputation Procedures." *Biometrika* 85 (December): 935–48.

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 (September): 699–704.

White, Stephen, Richard Rose, and Ian McAllister. 1997. *How Russia Votes*. Chatham, NJ: Chatham House.

Whitefield, Stephen, and Geoffrey Evans. 1996. "Support for Democracy and Political Opposition in Russia, 1993–95." *Post Soviet Affairs* 12 (3): 218–52.

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