How Multiple Imputation Makes a Difference

Ranjit Lall*

Forthcoming, *Political Analysis*

Abstract

Political scientists increasingly recognize that multiple imputation represents a superior strategy for analyzing missing data to the widely used method of listwise deletion. However, there has been little systematic investigation of *how* multiple imputation affects existing empirical knowledge in the discipline. This article presents the first large-scale examination of the empirical effects of substituting multiple imputation for listwise deletion in political science. The examination focuses on research in the major subfield of comparative and international political economy (CIPE) as an illustrative example. Specifically, I use multiple imputation to reanalyze the results of almost every quantitative CIPE study published during a recent five-year period in *International Organization* and *World Politics*, two of the leading subfield journals in CIPE. The outcome is striking: in almost half of the studies, key results "disappear" (by conventional statistical standards) when reanalyzed.

*Department of Government, Harvard University. I am grateful to Anthony Atkinson, Jeffry Frieden, Adam Glynn, James Honaker, Gary King, Walter Mattli, Margaret Roberts, Beth Simmons, Arthur Spirling, and the editors and anonymous reviewers of *Political Analysis* for helpful comments and suggestions. I also thank Olivier Accominotti, Todd Allee, Ben Ansell, Lucio Baccaro, Carles Boix, Sarah Brooks, Asif Efrat, Sean Ehrlich, Lawrence Ezrow, Marc Flandreau, Alexandra Guisinger, Caroline Hartzell, Philip Keefer, Jeffrey Kucik, Marcus Kurtz, Christopher Meissner, Sonal Pandya, Clint Peinhardt, Krzysztof Pelc, Kristopher Ramsay, Diego Rei, David Rueda, David Singer, and Hugh Ward for generously sharing data with me.

1 Introduction

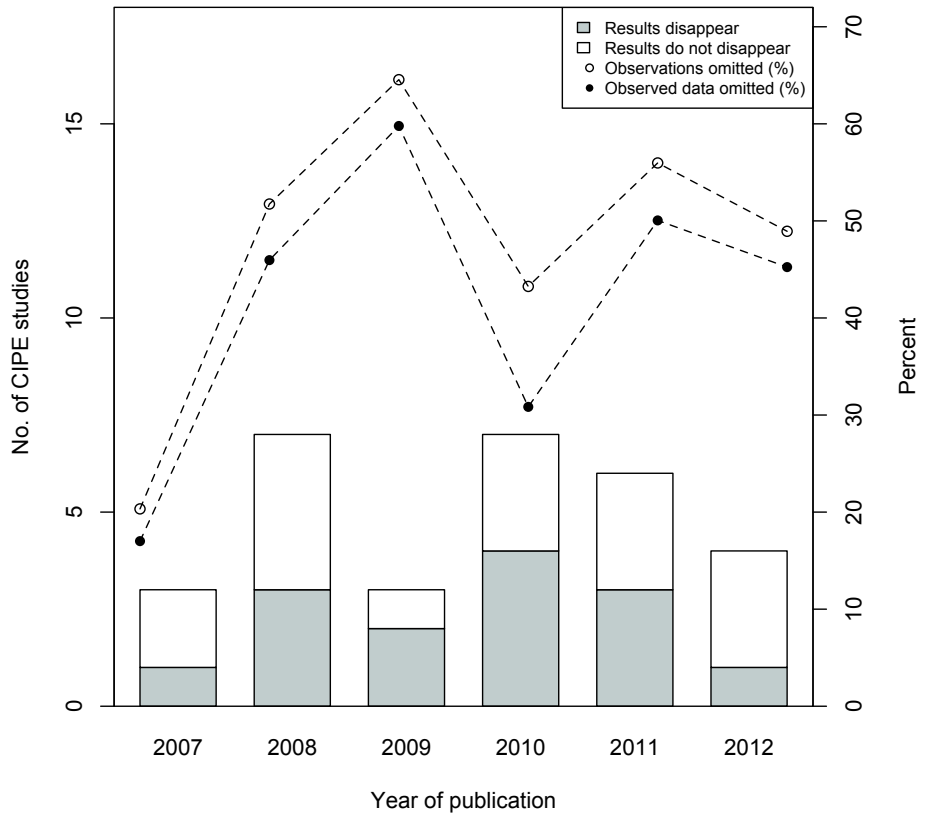
Political scientists increasingly recognize that multiple imputation represents a superior strategy for analyzing missing data to the widely used method of listwise deletion. The case for multiple imputation is clear. Listwise deletion, which involves omitting observations with missing values on any variable, produces inefficient inferences and is unbiased only in the unlikely situation that the pattern of missing data is completely random.[1] Multiple imputation, which involves replacing each missing cell with multiple values based on information in the observed portion of the dataset, not only generates considerably more efficient inferences than listwise deletion but also is unbiased under more realistic distributions of missing data.[2] While these advantages are now widely acknowledged in the discipline, however, there has been little systematic investigation of *how* multiple imputation affects existing empirical knowledge. Does employing the technique reaffirm or challenge established statistical results in political science?

This article presents the first large-scale examination of the empirical effects of substituting multiple imputation for listwise deletion in political science. The examination focuses on research in the major subfield of comparative and international political economy (CIPE) as an illustrative example. I argue that, in addition to being highly inefficient, listwise deletion tends to produce biased statistical inferences in CIPE because the pattern of missing values is not completely random. Most notably, poorer and less democratic countries are more likely to have missing data, causing listwise deletion to give rise to a particular selection problem that I call *advanced democracy bias*. Despite these problems, however, use of listwise deletion remains widespread in CIPE. A review

[1] Listwise deletion is the default option for dealing with missing data in most statistical software programs used by political scientists (including Stata, R, SAS, and SPSS).

[2] Multiple imputation is emerging as the principal alternative to listwise deletion in many areas of the social and natural sciences. Van Buuren goes as far as to suggest that multiple imputation is "now accepted as the best general method to deal with incomplete data in many fields" (2012, 25). For statistics on the rapid growth of the applied literature on multiple imputation in recent decades, see 27-28.

Figure 1 Preview of Reanalysis



Notes: Bars correspond to the left y-axis and dashed lines to the right y-axis. The circular points connected by the lines represent averages for all articles published in a given year.

of almost 100 CIPE studies recently published in five leading political science journals indicates that 90 percent continue to employ listwise deletion as their primary missing-data method, while only five percent have switched to multiple imputation.[3]

Specifically, I use multiple imputation to reanalyze the results of almost every quantitative CIPE study published during a recent five-year period in *International Organization*

[3]The review covers all CIPE studies published in the *American Political Science Review*, the *American Journal of Political Science*, the *British Journal of Political Science*, *International Organization*, and *World Politics* between July 2007 and July 2012. The remaining five percent of studies employ another ad-hoc technique, such as averaging observed data or substituting zero for missing values. Worryingly, more than three-quarters of studies — all of which used listwise deletion — were not explicit about how they dealt with missing data.

and *World Politics*, two of the leading subfield journals in CIPE.[4] The outcome of the reanalysis, previewed in Figure 1, is striking. In almost half of the studies, key results "disappear" when the main statistical analysis is re-estimated using multiply imputed data (shaded portion of bars, corresponding to left y-axis). That is, at least half of the regression coefficients on the key explanatory variable(s) that were previously statistically significant at the 10 percent level either cease to be significant or experience a change in sign; alternatively, in the case of "negative" findings, at least half of the coefficients on the key explanatory variable(s) that were previously nonsignificant become significant (regardless of sign).[5] The reanalysis also sheds light on the considerable scale of the missing-data problem in CIPE: an average of 48 percent of eligible observations are excluded from the main analysis due to listwise deletion (hollow circles, corresponding to right y-axis), resulting in the loss of 43 percent of available observed data (solid circles).

In addition to challenging the results of a number of prominent recent studies in CIPE, the article's findings have important implications for quantitative work in other areas of political science, many of which are likely to be similarly ill-suited to listwise deletion and have paid equally little attention to missing-data issues. In the concluding section, I offer some brief speculations on whether and how substituting multiple imputation for listwise deletion might affect empirical knowledge in different subfields.

2 The Missing-Data Problem in CIPE

This section provides a brief overview of the missing-data problem in CIPE. The first part discusses the methodological issues that arise when listwise deletion is used to analyze missing values in CIPE datasets. The second part explains how and under what

[4]According to the Thomson Reuters *Journal Performance Indicators* database, *International Organization* and *World Politics* had the highest cumulative impact factors of all journals in the subject category of "International Relations" over the period 1980-2013 (see http://researchanalytics.thomsonreuters.com/).

[5]Section 3 provides more detailed information on the size of these changes, including percentage differences in coefficient estimates and t-ratios.

conditions multiple imputation can improve the quality of inferences in CIPE research. Throughout the section, I highlight points that can be generalized to other areas of political science.

2.1 Income, Institutions, and Advanced Democracy Bias

Sources of cross-national data on economic activity — such as the Penn World Table, the World Bank's World Development Indicators, and the International Monetary Fund's (IMF) World Economic Outlook — tend to contain a high proportion of missing values. It is thus surprising that CIPE scholars have not paid more attention to the potential methodological pitfalls of using listwise deletion to analyze such values. Generally speaking, the performance of listwise deletion can be evaluated in terms of three criteria: bias, efficiency, and the ability to yield reasonable estimates of uncertainty (Graham 2009). With respect to efficiency, listwise deletion is always wanting: by discarding information in incomplete observations, it results in higher standard errors and reduced statistical power. Although it fares better on the third criterion — estimated standard errors are generally valid — this advantage is offset by losses in efficiency (Allison 2002).

The bias caused by listwise deletion is a more complex issue that rests on the *mechanism* by which data become missing. Scholars usually distinguish between three such mechanisms. Data are (1) *missing completely at random* (MCAR) if the probability that a given value is missing does not depend on any information in the dataset; (2) *missing at random* (MAR) if it depends on observed data only; and (3) *missing not at random* (MNAR) if it depends (at least in part) on missing data.[6] Listwise deletion is unbiased

[6]More formally, if Z denotes an $(n \times p)$ dataset with an observed portion Z_{obs} and a missing portion Z_{mis}, M denotes a matrix of the same dimensions as Z in which cells have a value of 1 if missing and 0 otherwise, and ϕ denotes parameters from the joint distribution function of Z, MCAR can be expressed as: $p(M|Z_{\text{obs}}, Z_{\text{mis}}) = p(M|\phi)$; MAR as $p(M|Z_{\text{obs}}, Z_{\text{mis}}) = p(M|Z_{\text{obs}}, \phi)$; and MNAR as $p(M|Z_{\text{obs}}, Z_{\text{mis}}) = p(M|Z_{\text{obs}}, Z_{\text{mis}}, \phi)$. These definitions are presented in greater detail in Little and Rubin (2002). Note that many studies refer to MNAR as NMAR ("not missing at random") or NI ("nonignorable").

only when the restrictive MCAR assumption holds — that is, when omitting incomplete observations leaves a random sample of the data. Under MAR or MNAR, deleting such observations produces samples that are skewed away from units with characteristics that increase their probability of having incomplete data.

How do data become missing in CIPE? A first, crucial point is that the MCAR assumption is unlikely to be satisfied in any area of CIPE or political science more generally. As Cranmer and Gill note, "It is difficult to think of a situation in political science, other than a computer malfunction, that would result in missing values being entirely unrelated to *any* attribute or political phenomena, observed or unobserved" (2013, 429). By contrast, situations in which some units are systematically more likely to have missing data than others are ubiquitous across the discipline. To offer a few examples: in electoral surveys in American politics, respondents who identify as "independents" are more likely to decline to answer questions about partisan identification and voting preferences; in studies of interstate conflict in international relations, dyads involving socialist and small powers are more likely to have incomplete dispute and alliance data; in subnational comparative politics datasets, rural areas are more likely to have missing bureaucratic, demographic, and political information. In general, therefore, listwise deletion can be expected to produce biased inferences in CIPE and other subfields.[7]

What are the determinants of missingness in CIPE? While the answer will vary from one study to another depending on the specific contents of its dataset, two factors tend to be important across a wide range of CIPE applications. The first is a state's level of *economic development*. Measuring, recording, and updating detailed information on multiple economic variables is a costly exercise. Many governments in developing countries either lack the financial resources to carry out these tasks or prefer to direct their limited

[7] I later show that the MCAR assumption — which, unlike the MAR and MNAR assumptions, can be tested in practice — is violated in every study included in my reanalysis.

budgets to more urgent developmental objectives. Moreover, they often lack the physical infrastructure, bureaucratic capacity, and technical expertise to meet the logistical challenges of data collection — challenges that are especially acute when a high proportion of economic activity occurs outside the formal sector and in hard-to-access rural areas (as suggested above). It should also be noted that developing nations are more likely to experience disruptions to data collection — often for several years at a time — due to internal political, economic, and social crises as well as wars, natural disasters, epidemics, and other adverse "shocks."

The second determinant is a state's *political institutions*. Empirical studies have found that democracies are more likely to release economic data to the public and to international organizations than autocracies (controlling for income and other variables) (Edwards, Coolidge and Preston 2011; Hollyer, Rosendorff and Vreeland 2011).[8] One potential explanation for this difference is that democratic leaders have stronger incentives to adhere to popular demands for transparency because their survival depends more strongly on voter welfare (Hollyer, Rosendorff and Vreeland 2011). Another possible theory is that democracies depend less on effective economic performance for their political legitimacy and are thus less concerned about revealing the true state of the economy. It is also conceivable that democratic institutions embody norms of transparency and accountability that politicians externalize in their interactions with the international community. Regardless of the exact causal mechanism, measures of democracy are likely to be strongly related to missingness in cross-national economic data.

The upshot is that, when applied to CIPE datasets, listwise deletion will often give rise to a form of selection bias that might be called *advanced democracy bias*. Since poorer

[8]These studies also find that richer countries are more transparent (though in the latter study only when country fixed-effects are included in the analysis). Ross (2006) makes a more nuanced argument about the relationship between income, democracy, and transparency, positing that high-income autocracies are *less* likely to release economic data than low-income ones. The evidence I present in Section 3.1 suggests that this is not a general trend across CIPE (see fn.25).

and less democratic countries are more likely to have missing data, listwise deletion will tend to produce samples that are skewed toward the richest and most democratic nations in the dataset. Needless to say, inferences based on such samples are likely to differ sharply from those based on a truly random sample of observations.

2.2 Improving Inferences with Multiple Imputation

How can multiple imputation address the problems caused by listwise deletion in CIPE? Multiple imputation involves three key stages.[9] First, m values are imputed for each missing cell, with variation across values reflecting uncertainty about the correct imputation model.[10] Imputed values are independent draws from a posterior distribution of the missing data conditional on the observed data. This is typically derived from a parametric model that assumes the complete data follow a joint probability distribution (with unknown parameters), which is most frequently a multivariate normal distribution. While real data obviously do not always conform to multivariate normality, this model has been found to perform well in the presence of violations (Rubin and Schenker 1986; Schafer 1997). It is important to note, however, that multiple imputation is still an evolving method, and there is no clear consensus about whether the multivariate normal approach is generally superior to, for instance, modeling each variable conditionally on all others (Kropko et al. 2014) or employing a nonparametric strategy such as replacing missing values with observed ones from similar units (Cranmer and Gill 2013).[11]

[9]Multiple imputation was first proposed by Rubin in the late 1970s and further developed with collaborators over the next decade (Rubin 1976, 1977, 1987; Rubin and Schenker 1986; Little and Rubin 1987).

[10]Contrary to a common misconception, it is indeed appropriate to impute values for the dependent variable. Excluding this variable from the imputation model implies that it has zero correlation with the included variables and thus results in downward-biased coefficient estimates (Little and Rubin 2002; Graham 2009).

[11]In making this choice, analysts should carefully consider the structure of their data. For instance, when the dataset includes categorical variables it may be possible to obtain better results with the conditional or non-parametric approach. Note, in addition, that multiple imputation is not well established for certain data structures — including multilevel data, high-dimensional data, survival data, multinomial data, and spatially lagged data — and should thus be used with caution in such applications.

In the second stage, each of the m complete datasets are analyzed and quantities of interest are estimated. Due to the separation between the imputation and analysis stages, complete-data methods can be applied to each dataset, making this a relatively straightforward task. Finally, the m separate point estimates are combined into one using the so-called "Rubin combination rules" (Rubin 1987). These rules state that the pooled point estimate is equal to the average of the m separate estimates, while its variance is equal to a weighted sum of the estimated variances within and between the m datasets.[12]

Multiple imputation is substantially more efficient than listwise deletion because it (1) utilizes rather than discards data in incomplete observations and (2) allows analysts to incorporate extra information into the imputation model by including variables that are not in the analysis ("auxiliary variables"). Multiple imputation also performs at least as well as listwise deletion on the third criterion mentioned earlier as it reflects uncertainty about imputed values and thus yields valid estimates of standard errors. This is a major advantage over ad-hoc "single" imputation methods such as replacing missing values with observed variable means (mean substitution), zero (zero imputation), or predicted values based on linear polynomials (linear interpolation).[13] These methods produce downward-biased standard errors because they treat imputed values as "knowns" rather than probabilistic estimates.[14] They can thus be legitimately accused of "making up data" — a common misconception about multiple imputation. The goal of multiple imputation is in fact to *preserve* key features of the existing data (such as means, variances, and covariances) while capturing the uncertainty of missing-data prediction.

Can multiple imputation avoid selection problems such as advanced democracy bias?

[12]That is, for a given quantity of interest β (say, a regression coefficient), $\hat{\beta} = \frac{1}{m}\sum_{i=1}^{m}\hat{\beta}_i$ and $\text{var}(\hat{\beta}) = W + (1 + \frac{1}{m})B$, where $W = \frac{1}{m}\sum_{i=1}^{m}\text{var}(\hat{\beta}_i)$ and $B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\beta}_i - \hat{\beta})^2$.

[13]Another common ad-hoc strategy that does not involve imputation is to simply drop control variables that result in the loss of a sizable number of observations. This strategy creates a tradeoff between sample size and omitted variable bias that can be avoided with multiple imputation.

[14]They also frequently produce biased point estimates (for different reasons in each case) (Little and Rubin 2002).

Unlike listwise deletion, multiple imputation is unbiased when data are MAR as well as MCAR. Under MNAR, however, multiple imputation cannot avoid bias: since missingness depends (to some extent) on missing values, observed data alone do not provide the basis for a valid imputation process. Strictly speaking, real data are almost always MNAR, with missingness depending in part on observed data and in part on missing data (Graham 2009). Critically, however, multiple imputation is not seriously biased under MNAR if missingness is strongly related to observed data and thus approximates MAR (Collins, Schafer and Kam 2001; Graham, Hofer and MacKinnon 1996; Schafer 1997). Thus, the key question is not simply: Are data MAR or MNAR? Rather, it is: How *much* does missingness depend on observed data?[15] Obviously, this is not possible to directly measure because we do not actually have access to missing data. Nevertheless, if the dataset contains one or more variables that are highly correlated with missingness, it is reasonable to assume that multiple imputation will perform almost as well as under (pure) MAR. Contrary to another common misconception, therefore, it often *is* appropriate to employ multiple imputation when data are MNAR.

If no variables in the dataset are strongly associated with missingness, however, multiple imputation can result in substantial bias.[16] It is important to stress, however, that this bias will not exceed that produced by listwise deletion in most cases: missingness is

[15]As Graham argues, "Because all missingness is MNAR (i.e., not purely MAR), then whether it is MNAR or not should never be the issue. Rather than focusing on whether [multiple imputation's] assumptions are violated, we should answer the question of whether the violation is big enough to matter to any practical extent" (2009, 567).

[16]The only way to avoid bias in this situation is to employ an "MNAR-specific method," which involves explicitly specifying the joint distribution of Z and M (Little 1993; Little and Rubin 2002). The two most widely used MNAR-specific methods are sample selection models and pattern-mixture models. In principle, both types of models could be appropriate in CIPE, though they should be used with caution because they are highly sensitive to empirically unverifiable assumptions (for instance, regarding the population distribution in selection models and pattern-specific parameters in pattern-mixture models). To my knowledge, there are no examples of either model in CIPE, most likely due to the difficulty of implementing them using standard statistical software and the general lack of attention to missing-data issues in this area.

no closer to being completely random under MNAR than under MAR.[17] In other words, the same conditions that cause multiple imputation to be severely biased also cause listwise deletion to be severely biased. Since multiple imputation is always more efficient than listwise deletion, even in this worst-case scenario it is still the preferable strategy.

The implication is that multiple imputation can help to mitigate selection problems such as advanced democracy bias so long as variables that measure or are correlated with determinants of missingness — in this case income and democracy — are included in the dataset. In Section 3.1, I show that every dataset in my reanalysis contains either a direct proxy for income and democracy or a set of associated variables that are also strongly related to missingness. This suggests that in CIPE multiple imputation will often yield substantial gains in terms of reduced bias relative to listwise deletion.

In CIPE and elsewhere, such gains will be largest under three conditions (see Table 1). First, variables of interest have high levels of missing data. The higher the proportion of incomplete observations in CIPE datasets, for instance, the greater the extent to which richer and more democratic countries will tend to be overrepresented in samples produced by listwise deletion. Second, hypotheses are tested on a heterogeneous sample in terms of correlates of missingness. As variation in income and democracy in CIPE datasets increases, so too does the extent to which missingness depends on these variables and thus on observed data (given the typical composition of CIPE datasets). The greater, in turn, the reduction in bias achieved by multiple imputation compared with listwise deletion. Third, the dataset contains a large number of variables that are related to missingness. While most CIPE datasets contain at least a few such variables, as noted above, higher numbers increase the degree to which missingness is related to observed

[17] If the analysis model is a (correctly specified) regression of Y on X, data for X are MNAR, and missingness does not depend on Y, it is possible for listwise deletion to be less biased than multiple imputation. These conditions, however, are rarely satisfied in the real world (King et al. 2001; van Buuren 2012).

Table 1 Performance of Multiple Imputation

	Large Gains in Bias Reduction	**Small Gains in Bias Reduction**
Variables in dataset	Many variables highly correlated with missingness (e.g., income and democracy in CIPE)	No variables highly correlated with missingness
Analysis sample	Heterogeneous in terms of missingness correlates	Homogeneous in terms of missingness correlates
Mechanism of missingness	Missingness depends to a large extent on observed data (approximating MAR)	Missingness depends primarily on missing data (extreme MNAR)
Issue area in CIPE (likely)	Economic performance, political regimes, trade, foreign aid, governance, public goods	Inequality, redistribution, welfare regimes, economic integration, policy diffusion

data and thus lower the bias caused by multiple imputation (much like political and economic heterogeneity).

Which CIPE studies are most likely to satisfy these conditions? The obvious candidates are studies of economic performance and political regimes. These studies are, by their very nature, concerned with a diverse set of countries in terms of income and democracy. Moreover, their datasets tend to have a high proportion of missing values — precisely *because* relatively poor and autocratic countries are more likely to have incomplete economic data — and include multiple alternative measures of income or democracy. Yet while the three conditions are most clearly fulfilled in these studies, they can also be met in other issue areas of CIPE, particularly those in which propositions are typically global in scope and variables of theoretical interest are highly correlated with income or democracy. We should therefore expect sizable gains in bias reduction in issue areas ranging from trade and foreign aid to governance and public goods.

Conversely, multiple imputation will offer small gains in bias reduction when (1) variables of theoretical interest have a low proportion of missing values; (2) hypotheses are

tested on a homogeneous sample in terms of missingness correlates; and (3) the dataset contains few or no variables that are related to missingness. Such situations are most likely to arise in two types of CIPE datasets. The first are small, issue-specific datasets that contain no variables that measure or are correlated with income and democracy. The second are datasets in which such variables are included but exhibit little variation across countries. Here, missingness will depend mostly on idiosyncratic factors that are unlikely to be measured, such as the mandate of data-gathering agencies and the occurrence of natural disasters. This type of dataset is common in issue areas where studies tend to focus on advanced democracies, such as inequality, redistribution, and welfare regimes. It can also be found in studies that focus on a single region, which are conducted across all of CIPE but are particularly common in the issue areas of economic integration and policy diffusion.

3 Reanalysis

The preceding discussion suggests that in CIPE multiple imputation typically offers major gains in efficiency and bias reduction over listwise deletion (and almost never performs worse than it). This section investigates the empirical effects of substituting multiple imputation for listwise deletion by presenting my reanalysis of published CIPE studies.[18] The first part describes the scope of the reanalysis and provides an overview of missing-data patterns in the studies. The second part discusses the specific steps by which multiple imputation was implemented. The third part sets out the main findings.

[18]For replication materials, see Lall (2016).

3.1 Scope

The reanalysis includes almost all CIPE studies (articles and research notes) containing some form of statistical analysis published in *International Organization* and *World Politics* between July 2007 and July 2012. A study is classified as an instance of CIPE research if it fulfills the following two criteria: (1) it seeks to either explain or understand the effects of variation in an "economic" variable (broadly defined); and (2) its empirical analysis is not limited in scope to a single country or territory.

A total of 42 publications satisfy these two criteria, a full list of which can be found in Section I of the online appendix. Three studies are automatically excluded from the reanalysis: two that already employ multiple imputation (Houle 2009; Scheve and Stasavage 2009) and one that contains no missing data (Obinger and Schmitt 2011). Of the remaining 39 studies, all of which use listwise deletion as their primary missing-data method, I managed to obtain the datasets for 30 through a combination of personal communications with authors and searches of institutional websites and online data repositories.[19] Only 10 of the 39 datasets could be acquired without a request to the study's author(s), and in almost one-third of the remaining cases such requests were not answered.

The reanalysis focuses on a study's main statistical analysis — that is, the set of estimation models in which its central theoretical or empirical proposition is tested (typically presented in the form of a single regression table).[20] For reasons of feasibility, I exclude analyses that test subsidiary propositions or merely examine the robustness of prior results.[21] The number of models comprising a study's main analysis varies considerably,

[19] If I was unable to find a study's dataset online, I contacted its author(s) via email to request access to it. I sent at least two follow-up emails to authors who did not respond to my initial request.

[20] In every study, this proposition is clearly stated in the abstract, introduction, or theory section. In the few instances where there are multiple propositions with no obvious ranking in terms of theoretical or empirical significance, I focus on the proposition that is tested first.

[21] Note, however, that I do reanalyze subsidiary propositions that are included in the main analysis.

ranging from one to 24 (with an average of 5.2). In total, the reanalysis encompasses 156 models across the 30 studies.

Summary statistics on missing-data patterns in the studies' main analyses are displayed in Table 2.[22] Three features of Table 2 are worth highlighting. The first is the substantial quantity of missing data in the studies.[23] On average, almost one-fifth of cells in their datasets are missing, with this figure exceeding 30 percent in around one-third of studies (and reaching as high as 73 percent). This alone is a cause for concern and gives us reason to view the results of some of the analyses with caution.

The second and most conspicuous feature is the remarkably high proportion of data excluded from the analyses as a result of listwise deletion. In almost half of the studies, over 50 percent of eligible observations in the dataset are excluded. Only in 7 studies is the rate of exclusion less than 25 percent. The upshot is that much of the observed data that could have been utilized in the analyses are discarded. In more than one-third of studies, over 50 percent of available observed values are lost; in the majority of such cases, the figure exceeds two-thirds. This is stark evidence of the inefficiency caused by listwise deletion in CIPE. By preserving information in incomplete observations, multiple imputation enables us to utilize an average of 77 percent more observed data.

Finally, a relatively large proportion of eligible countries in the studies' datasets — almost one-quarter on average — are not just underrepresented but entirely *omitted* from their analyses. In several cases, the majority of countries are left out, implying severe selection bias. It is also worth noting that in the 23 time-series cross-section (TSCS) studies a reasonably high proportion of eligible years are excluded (16 percent on average). This suggests that many of the analyses are likely to suffer from bias due to the

[22]For analyses that contain more than one estimation model, each statistic is averaged across all models.

[23]The five most commonly used data sources in the studies are (in order): (1) the World Bank's World Development Indicators; (2) the Polity data series; (3) the United Nations Conference on Trade and Development's (UNCTAD) Trade Analysis and Information System database; (4) the Penn World Tables; and (5) the International Monetary Fund's World Economic Outlook.

Table 2 Missing Data in Reanalyzed Studies

Study	Dataset missing (%)	Main analysis	*N* omitted (%)	Observed data omitted (%)	Data structure	Countries omitted (%)	Years omitted (%)
Allee and Scalera 2012	3.94	Table 4	46.95	44.60	TSCS	18.82	12.70
Dreher and Gassebner 2012	10.48	Table 1	60.97	54.19	TSCS	47.62	9.41
Caraway, Rickard, and Anner 2012	20.30	Table 2	72.63	67.67	TSCS	65.25	4.76
Brooks and Kurtz 2012	3.38	Table 1	15.16	14.49	TSCS	0.00	0.00
Pelc 2011a	37.66	Table 3	71.59	66.49	CS	5.56	
Pelc 2011b	34.15	Table 2	51.46	46.64	CS	18.47	
Allee and Peinhardt 2011	14.94	Table 2	73.48	63.02	TSCS	50.24	39.47
Ramsay 2011	2.97	Table 3	22.77	21.79	TSCS	5.73	0.00
Ward, Ezrow, and Dorussen 2011	12.62	Table 1	29.44	20.27	TSCS	0.00	0.00
Oatley 2011	25.65	Table 2	87.08	82.11	TSCS	45.81	36.67
Broz and Plouffe 2010	9.47	Table 4	29.77	25.25	CS	62.52	
Pandya 2010	3.60	Table 2	16.12	15.23	CS	0.00	
Cao and Prakash 2010	49.80	Table 2	84.41	74.13	TSCS	41.67	50.00
Winters 2010	7.78	Figure 4	42.51	39.86	TSCS	8.99	0.00
Guisinger and Singer 2010	33.30	Table 1	73.70	63.68	TSCS	47.37	14.29
Hartzell, Hoddie, and Bauer 2010	8.75	Table 2	43.39	40.37	TSCS	24.84	3.33
Efrat 2010	1.38	Table 1	11.02	9.86	CS	11.02	
Gawande, Krishna, and Olarreaga 2009	17.79	Table 1/4	73.27	72.91	TSCS	0.00	
Bueno de Mesquita and Smith 2009	18.11	Table 1	41.86	36.58	TSCS	27.22	24.16
Morrison 2009	32.65	Table 3	78.52	69.84	TSCS	44.57	32.56
Lopez-Cordova and Meissner 2008	17.59	Table 5	56.20	41.11	CS	56.18	
Kucik and Reinhardt 2008	43.47	Table 1	61.28	55.96	TSCS	30.10	47.73
Ansell 2008	42.42	Table 1	70.67	64.18	TSCS	24.60	42.64
Boix 2008	72.97	Table 1	32.34	23.39	TSCS	23.22	11.03
Rueda 2008	1.56	Figure 7	21.24	20.56	TSCS	4.17	2.17
Accominotti and Flandreau 2008	7.58	Table 4	80.22	78.32	TSCS	0.00	0.00
Kurtz and Brooks 2008	2.31	Table 3	40.25	38.20	TSCS	17.65	5.26
Baccaro and Rei 2007	4.73	Table 1	15.99	15.30	TSCS	0.00	9.76
Ehrlich 2007	4.87	Table 1	29.95	23.42	TSCS	4.55	20.83
Keefer 2007	2.43	Table 4	15.04	12.32	TSCS	15.24	4.17
Average	18.29		48.31	43.39		23.38	16.13

Notes: For analyses that contain more than one estimation model, I take the average value across all models. In column 6, "TSCS" = time-series cross-sectional; "CS" = cross-sectional. Figures for Gawande, Krishna, and Olarreaga 2009 refer to Table 1 in the article, which provides the basis for the results of its main empirical analysis, Table 4.

underrepresentation of particular time periods as well as particular countries.

An examination of the composition of the 30 datasets suggests that they are considerably better suited to multiple imputation than listwise deletion. First, the MCAR assumption is not satisfied in a single case. I checked this assumption using the standard "Little's MCAR test," which evaluates a null MCAR hypothesis that observed variable means for subgroups of observations sharing the same missing-data pattern do not differ from expected population means based on maximum likelihood (ML) estimates (Little 1988). As shown in Table A2 in the online appendix, in every case the χ^2 test statistic — a weighted sum of the standardized differences between the subgroup and expected means — was statistically significant at the one percent level, resulting in a rejection of the null hypothesis.[24]

Second, missingness is strongly related to observed data. Four-fifths of the datasets include a measure of GDP per capita, while almost two-thirds contain a variable recording Polity scores. Table 3 displays means for each variable in the sample included in a study's main analysis and the sample excluded from it. In almost all studies, means in the included sample are higher than those in the excluded sample. While we cannot compare absolute levels of the variables due to differences in calibration and data sources, the average included observation has a GDP per capita and Polity score 42 percent and 131 percent higher, respectively, than the average excluded observation. Importantly, a Student's *t*-test reveals that the difference between the included and excluded means is

[24]The test statistic is defined as:

$$d^2 = \sum_{j=1}^{J} m_j (\bar{z}_{obs,j} - \hat{\mu}_{obs,j}) \widetilde{\Sigma}_{obs,j}^{-1} (\bar{z}_{obs,j} - \hat{\mu}_{obs,j})^T \quad (1)$$

where z_i is a $(1 \times p)$ vector of values for observation i (assumed to follow a multivariate normal distribution), m_j is the number of observations with missing-data pattern j, $\bar{z}_{obs,j}$ is the observed sample average for j, $\hat{\mu}$ is the ML estimate of the $(1 \times p)$ population mean vector (μ), and $\widetilde{\Sigma}$ is the ML estimate of the $(p \times p)$ covariance matrix of z_i (Σ). The test was implemented using the *mcartest* command in Stata (version 13.1), which in most instances required removing highly collinear variables from the dataset.

statistically significant at the five percent level in 23 out of 24 studies in the case of GDP per capita and 17 out of 19 studies in the case of Polity scores.[25]

To more rigorously assess the extent to which missingness is related to income and democracy, for each study I estimated a logit model in which the dependent variable is a dummy variable indicating whether or not a given observation is included in the main analysis and the regressors are GDP per capita and/or Polity scores (depending on which variables are in the dataset). As shown in Table 3, the coefficients on GDP per capita are positive in 17 out of 24 studies and significant at the five percent level in 22; the coefficients on Polity scores are positive in all 19 studies and significant in 18.[26] For the five datasets that contain neither variable, I estimated a similar model in which the regressors are variables in the dataset that tend to be highly correlated with income and democracy, such as trade, financial openness, and public spending. In every model, at least one — and in most cases several — of the coefficients on the regressors were significant, indicating that missingness is strongly associated with observed data.

[25]For analyses containing more than one estimation model, I calculate the difference for each model separately and combine the p-values from the multiple t-tests using Fisher's method, which yields a single test statistic:

$$X^2_{2k} = -2\sum_{i=1}^{k} ln(p_i) \tag{2}$$

where p_i is the p-value for the ith hypothesis test and k is the number of tests being combined. If dyads rather than countries are the unit of analysis, I average GDP per capita and Polity scores across the two countries.

[26]When the sample is restricted to autocracies, the coefficient on GDP per capita remains positive (and significant) in the majority of studies. Thus, there is no general tendency for high-income autocracies to disclose less data than low-income ones, as suggested by Ross (2006). Nevertheless, the fact that the coefficient is negative in several of the restricted and unrestricted models suggests that the positive effect of income on missingness may be nonlinear or conditional on another variable. This is an interesting avenue for further research.

Table 3 GDP per Capita and Polity Scores in Included and Excluded Samples

	Average GDPPC		Average Polity Score		Effect of GDPPC on Inclusion		Effect of Polity on Inclusion	
Study	**Included**	**Excluded**	**Included**	**Excluded**	**Sign**	$p < 0.05$	**Sign**	$p < 0.05$
Allee and Scalera 2012	4,245.10	5,033.80	-3.14	-3.41	+	✔	+	✔
Dreher and Gassebner 2012			0.75	-2.68			+	✔
Caraway, Rickard, and Anner 2012	1,635.17	1,309.26	10.98	10.07	+	✔	+	✘
Brooks and Kurtz 2012	3,307.48	2,753.11			+	✔		
Pelc 2011a	7,272.27	3,350.32	11.70	10.58	+	✔	+	✔
Pelc 2011b	5,601.46	7,882.31	3.42	3.09	-	✔	+	✔
Allee and Peinhardt 2011	5,436.32	8,247.87	1.99	-0.28	-	✔	+	✔
Ramsay 2011	6,923.50	1,846.58	0.49	0.21	+	✔	+	✔
Oatley 2011	2,501.24	2,974.34	1.40	-2.96	+	✘	+	✘
Broz and Plouffe 2010	4,859.88	1,969.15	4.93	3.88	+	✔	+	✔
Cao and Prakash 2010	10,021.90	6,728.42	4.12	-1.01	+	✔	+	✔
Winters 2010	1544.78	1975.27	2.79	2.50	-	✔	+	✔
Guisinger and Singer 2010	5,910.49	3,815.71			+	✔		
Hartzell, Hoddie, and Bauer 2010	4,737.36	3,692.79	2.55	-4.15	-	✔	+	✔
Efrat 2010	10,736.49	5,026.16	4.94	2.52	+	✔	+	✔
Bueno de Mesquita and Smith 2009	11,234.61	9,392.65			+	✔		
Morrison 2009	8,514.80	4,122.82	3.09	-1.38	+	✔	+	✔
Lopez-Cordova and Meissner 2008	3,748.97	2,709.14	1.97	0.05	-	✔	+	✔
Kucik and Reinhardt 2008	3,985.11	5,510.40	1.03	-0.68	-	✔	+	✔
Ansell 2008	7,764.48	5,491.92	3.03	-0.71	+	✔	+	✔
Boix 2008	4,680.99	2,717.83	-7.50	-44.29	+	✔	+	✔
Accominotti and Flandreau 2008	584.44	577.59			+	✘		
Kurtz and Brooks 2008	5,853.41	5,081.93	7.49	5.79	+	✔	+	✔
Ehrlich 2007	9,797.02	6,272.99			+	✔		
Keefer 2007	7,570.05	10,815.20			-	✔		

Notes: Effects of GDP per capita (GDPPC) and Polity scores on inclusion represent coefficients on these variables from a logit regression in which the dependent variable is a dummy for whether a given observation is included in a study's main empirical analysis. For analyses containing more than one estimation model, separate regressions are conducted for each model, with p-values combined using Fisher's method. When dyads rather than countries are the unit of analysis, GDP per capita and Polity scores are averaged across the two countries.

3.2 Implementing Multiple Imputation

To implement multiple imputation, I use Honaker, King, and Blackwell's (2011) *Amelia II* program in R, the most widely used multiple imputation software in political science.[27] *Amelia II* is the successor to the original *Amelia* program developed by King et al. (2001). Both versions implement joint multivariate normal multiple imputation and employ a bootstrapping expectation-maximization (EM) algorithm to take draws from the posterior distribution. Yet unlike its predecessor, which was designed primarily for cross-sectional survey data, *Amelia II* includes special features for TSCS data, which are common in CIPE. As discussed shortly, while I take full advantage of these features, my implementation strategy differs slightly from that recommended by Honaker, King, and Blackwell in light of recent findings from the statistics literature.

Building an imputation model involves three key steps. The first is to identify which variables in the dataset to include in the model. To avoid bias, the model must contain every variable in the subsequent analysis (Meng 1994). This includes interaction terms and squares, as omitting such variables is equivalent to assuming that they have zero correlation with other analysis variables and thus biases regression estimates downward (von Hippel 2009). It also includes the main cross-section and time-series index variables (typically "Country" and "Year" in CIPE datasets), which must be declared to *Amelia II*. Of the auxiliary (non-analysis) variables, those in the following four categories can be safely excluded because they provide no extra information: (1) additional index variables; (2) individual items of composite variables; (3) dummies derived from other variables; and (4) variables measuring data parameters such as means and variances.

In principle, including all of the remaining auxiliary variables in the imputation

[27]The software manual for *Amelia II*, which was published in 2011, already has more than 950 citations on Google Scholar (search performed 4 May 2016). The program itself has been available since 2006 and thus could have been used by any of the studies in the reanalysis. Other notable multiple imputation software packages include *mice*, *Hmisc*, and *hot.deck* in R, *ice* and the *mi* command in Stata, and *PROC MI* in SAS.

model is desirable. If these variables are strongly related to the pattern of missing data in the analysis variables, they increase the extent to which missingness depends on observed data and thus reduce bias. If they are also highly correlated with the missing analysis variables themselves, they make imputed values more precise and thus increase efficiency. In practice, however, an "inclusive" strategy often causes the imputation model to become so large that the EM algorithm fails to converge (van Buuren 2012). In addition, very large imputation models have been found to actually reduce efficiency and increase bias, probably because they lower the ratio of observations to variables and thus generate instability in regression models (Hardt, Herke and Leonhart 2012).

I thus adopt the following rule of thumb, which allows us to capture the gains from including auxiliary variables while keeping the imputation model at a manageable size.[28] If there are less than 100 variables left after removing the four types of auxiliary variables described above, I include all of them in the model.[29] If 100 or more remain, I include only those auxiliary variables that meet the following two requirements: (1) they have a correlation of $r \geq 0.5$ with at least one analysis variable or at least one specially created dummy indicating whether observations for a given analysis variable are missing; and (2) less than 25 percent of their values are missing (since highly incomplete variables provide information at greater cost in terms of model size).[30]

The second step in building the imputation model is to add features that improve its fit to the data. The software manual for *Amelia II* recommends declaring categorical variables to the program to ensure that their imputed values are rounded off to the nearest discrete number (thus avoiding impossible values). In addition, it suggests applying

[28]Similar rules are proposed by Schafer (1997); White, Royston and Wood (2011).

[29]This threshold, which is based on my experience of when *Amelia II*'s algorithm fails to converge, is also recommended by Graham (2009). On average, 97 variables remained after removing the four types of auxiliary variables, with one-third of datasets exceeding the 100-variable threshold.

[30]The $r \geq 0.5$ cutoff is also suggested by Graham (2009); Hardt, Herke and Leonhart (2012). An average of 47 variables were included in the imputation model. In one study, the sample was so small that I was forced to depart from my rule of thumb and only include analysis variables (Keefer 2007).

logarithmic, square root, and logistic transformations to heavily skewed variables to normalize their distributions (Honaker, King and Blackwell 2011, 14-16). Recent research, however, has shown that these modifications tend to cause more problems than they solve. Rounding off imputed values for categorical variables has been found to produce biased parameter estimates because such values are typically not normally distributed around the cutoff point (for instance, 0.5 in the case of binary variables) (Horton, Lipsitz and Parzen 2003; Allison 2005; Cranmer and Gill 2013). Transforming skewed variables has also been found to increase bias because it alters their relationship with other variables in the imputation model; in effect, it is equivalent to assuming that they have zero correlation with such variables (von Hippel 2013). Thus, counterintuitively, analysts are better off leaving imputations at impossible values and non-normal variables skewed.[31]

I add only three features to the imputation model. First, for TSCS datasets I include a sequence of third-order time polynomials — a new capability in *Amelia II* — to better model smooth temporal variation within cross-section units. Second, I include lags of the dependent and key explanatory variables — or leads if they are already lagged — since data for one period tend to be highly correlated with data for the previous (or subsequent) period. Third, I add a ridge prior of one percent of the number of observations in the dataset, which addresses computational problems caused by high levels of missing data and multicollinearity as well as increasing the numerical stability of the imputation process (Honaker, King and Blackwell 2011, 20).

The final step is to decide how many imputations to conduct. Statisticians have traditionally recommended no more than five imputations, which is the default setting in

[31] As mentioned in fn. 10, an alternative (and potentially superior) approach to dealing with categorical variables is to draw imputations from conditional distributions or from observed values of similar units. As a sensitivity check, I used these two strategies to re-impute missing values in five randomly selected studies in which either the dependent variable or (at least) one of the key explanatory variables is categorical. Specifically, I employed the *mice* package in R to implement the former strategy and the *hot.deck* package to implement the latter. The results for the key estimation models, reported in Section II of the online appendix, are very similar to those derived using *Amelia II*.

Amelia II (Rubin 1987; Schafer 1997). This recommendation is based on Rubin's (1987) formula for the relative efficiency of a parameter estimate based on m imputations compared with a fully efficient one based on ∞ imputations: $(1 + \frac{\gamma}{m})^{-1}$, where γ is the "fraction of missing information," a complex quantity that roughly captures how much information about the parameter is lost due to missing data.[32] This formula indicates that the efficiency of an estimate when $m = 5$ is always close to that of one when $m = \infty$. Even when γ is as high as 50 percent, for instance, relative efficiency is still more than 90 percent. The implication is that the benefits of raising m above five will not outweigh the costs in terms of extra computation time.

However, recent research has shown that conducting just five imputations can have negative consequences for properties closely related to efficiency. As m decreases, studies have found, there is a sharp decline in statistical power and increase in confidence intervals and Monte Carlo standard errors (i.e., errors across repeated runs of the same imputation process) (Bodner 2008; Graham, Olchowski and Gilreath 2007; White, Royston and Wood 2011). These studies generally suggest that to avoid undesirable levels of statistical power and precision m should be approximately equal to the percentage of incomplete observations in the dataset (a conservative estimate of γ). Thus, if half of the observations are incomplete m should be around 50 — 10 times the number implied by Rubin's formula (assuming we desire relative efficiency of at least 90 percent).[33]

While sensible, this rule has a major weakness: the percentage of incomplete observations is sensitive to the number of variables in the imputation model. As this number increases, the percentage of incomplete observations rapidly falls to zero (since missing-

[32]For a parameter estimate $\hat{\beta}$, the fraction is missing information is defined as:

$$\hat{\gamma} = \frac{(1 + \frac{1}{m})B + \frac{2}{v_m + 3}}{\text{var}(\hat{\beta})} \tag{3}$$

where v_m is the number of degrees of freedom with m imputations.

[33]This rule was originally proposed by von Hippel (2009).

data patterns are not identical across variables), even if γ stays the same. I thus adopt an amended version of the rule: m is equal to the average missing-data rate of all variables in the imputation model.[34] Although this rate is a less conservative estimate of γ, it is typically a more accurate one because high correlations among variables — a feature of every imputation model in the reanalysis — tend to lower γ below the percentage of incomplete observations (Rubin 1987). In addition to being less sensitive to the number of variables in the imputation model, therefore, this rule is likely to result in a more appropriate number of imputations in terms of statistical power and precision.[35]

Having carried out the imputations, I re-estimate the main analysis using the m complete datasets. To verify that I am employing the correct analysis model, I first replicate the published results with the original dataset. Finally, I aggregate the m sets of new results using the Rubin combination rules (see Section 2.2).[36]

3.3 Results

The main findings of the reanalysis are summarized in Table 4 (full regression results are displayed in Section I of the online appendix). Most strikingly, the main empirical results of nearly half of the studies — 14 out of 30 — disappear (as defined earlier) when re-estimated with multiply imputed data. In nine of these studies, at least three-quarters of the regression coefficients on the key explanatory variable(s) cease to be statistically significant at the 10 percent level, experience a reversal in sign, or, in the case of negative findings, become significant where they were previously nonsignificant. In four of the

[34] A similar rule is suggested by van Buuren (2012). I impose a lower bound of $m = 5$. The average m in the reanalysis is 16; the highest is 67.

[35] To assess the sensitivity of the reanalysis results to variation in m, I re-imputed the missing values in the five studies mentioned in fn. 27 using the two alternative rules discussed above, i.e., setting m equal to (1) five and (2) the percentage of incomplete observations in the dataset. As shown in Section II of the online appendix, in both cases the results were almost identical to those based on the adopted rule.

[36] As I conduct the replications in Stata (version 13.1), I import the m complete datasets from R and perform the combinations using the built-in *mi* command.

studies, *every* key coefficient experiences one of these changes.

Even results that do not disappear are typically altered in important ways. In four studies, results become "weaker" in the reanalysis: at least one — but less than half — of the key coefficients drops out of significance or experiences a change in sign (in the case of positive findings) or gains significance (in the case of negative findings). In four studies, meanwhile, results become "stronger": at least one of the key coefficients becomes significant with the theoretically predicted sign (in the case of positive findings) or ceases to be significant (in the case of negative findings).[37]

Only in eight of the 30 studies are results not subject to any of these changes (and thus classified as experiencing "no change"). Even in these cases, however, multiple imputation modifies the size, sign, and significance level of coefficients on a host of control variables. Although I still consider the results of such studies to be robust to multiple imputation, these often substantial modifications deserve close attention from CIPE scholars.

Table 4 also provides information on the direction and magnitude of changes in results. In 16 of the 23 studies in which some form of change occurs, the direction is negative — that is, coefficients lose significance rather than gaining it (where they were previously nonsignificant).[38] This reflects the fact that the vast majority of studies in the reanalysis originally reported positive rather than negative findings.

The four columns on the far right display two measures of the size of changes in results. The two nearer columns show the mean percentage change in (absolute-value) t-ratios for all coefficients in the analysis (column 6) and only coefficients on the key explanatory variable(s) (column 7). Naturally, this change is largest in cases of disap-

[37]It is possible for results to become stronger and weaker simultaneously (for instance, if some key coefficients gain significance while others lose it). I classify such cases according to which effect predominates.

[38]In two cases, the direction of change is mixed because the study reported positive *and* negative findings (and both sets of findings were altered in the reanalysis).

Table 4 Summary of Reanalysis Results

Study	Dependent variable(s)	Key explanatory variable(s)	Reanalysis outcome	Sign of △	△ t-ratio (%)		O/R ratio	
					All	KEV(s)	All	KEV(s)
Allee and Scalera 2012	Trade flows (1950-2006)	Type of accession to GATT/WTO	*Stronger*	+	257.45	56.91	0.64	1.15
Dreher and Gassebner 2012	Government crisis (1970-2002)	IMF or World Bank program	*Disappear*	-	110.65	165.08	-0.47	1.02
Caraway, Rickard, and Anner 2012	IMF labor conditionality (1980-2000)	Domestic labor power, democracy	*Weaker*	-	222.93	52.62	1.25	-5.69
Brooks and Kurtz 2012	Capital openness (1983-2007)	Legacy of import substitution	*No change*		88.04	11.65	1.36	1.06
Pelc 2011a	Terms of WTO accession (1995-2008)	Industry product imports	*Disappear*	-	64.93	35.24	1.34	1.73
Pelc 2011b	Binding overhang (1996-2006)	Exchange rate regime, trade remedies	*Disappear*	-	208.08	90.98	1.53	1.44
Allee and Peinhardt 2011	FDI inflows (1984-2007)	Signing of BITs, ICSID filings	*Stronger*	+	86.43	211.87	4.32	0.54
Ramsay 2011	Democracy (1968-2002)	Instrument for oil income per capita	*No change*		120.86	26.15	2.46	1.24
Ward, Ezrow, and Dorussen 2011	Party policy position (1973-2002)	Median voter position, globalization	*Disappear*	-	51.00	51.43	-18.49	1.98
Oatley 2011	Tariff rate (1970-99)	Regime type, WTO membership	*No change*		149.09	45.57	-5.48	2.50
Broz and Plouffe 2010	Concern about inflation (1999)	Exchange rate regime	*No change*		26.06	38.45	1.13	0.84
Pandya 2010	Support for FDI (1995, 1998, 2001)	Individual skill level	*Weaker*	-	90.55	30.06	1.11	1.53
Cao and Prakash 2010	Pollution intensity (1980-2003)	Trade competition	*Disappear*	+/-	745.00	131.47	2.71	0.36
Winters 2010	World Bank lending (1986-2002)	Quality of governance in borrower	*Disappear*	+/-	160.08	458.59	1.94	1.26
Guisinger and Singer 2010	Inflation rate (1974-2004)	De jure and de facto exchange rate regime	*Disappear*	-	231.88	92.98	1.93	8.69
Hartzell, Hoddie, and Bauer 2010	Onset of civil war (1970-99)	Signing of IMF agreement	*Disappear*	-	158.69	63.76	0.30	1.64
Efrat 2010	Support for global arms rules (2006)	Negative externalities of arms trade	*Stronger*	+	21.80	17.64	0.95	0.96

(Table 4 continued)

Gawande, Krishna, and Olarreaga 2009	Government welfare concern (1988-2000)	Political institutions, socioeconomic factors	*Disappear*	-	123.96	123.96	42.59	42.59
Bueno de Mesquita and Smith 2009	Size of aid-for-policy deals (1960-2001)	Policy salience, resources, political institutions	*Weaker*	-	219.59	77.57	2.48	3.98
Morrison 2009	Regime change (1973-2001)	Government nontax revenue	*Disappear*	-	240.08	46.08	3.8	20.82
López-Córdova and Meissner 2008	Democracy (1870-2000)	Instrument for trade openness	*Disappear*	-	110.19	53.07	1.38	2.67
Kucik and Reinhardt 2008	WTO accession, adoption of AD law (1981-2003)	Possession of AD law, WTO membership	*No change*		83.09	48.62	0.69	1.09
Ansell 2008	Education spending (1960-2000)	Democracy, economic openness	*Weaker*	-	343.79	27.12	9.03	4.12
Boix 2008	Civil war (1850-1999)	Family farms, occupational diversification	*No change*		432.35	312.06	0.73	1.91
Rueda 2008	Income inequality, social policy (1973-1995)	Partisanship, corporatism, social policy	*No change*		140.69	62.83	1.22	-1.27
Accominotti and Flandreau 2008	Bilateral trade (1850-80)	Existence of MFN treaty	*Disappear*	+	3377.51	27613.60	7.37	50.14
Kurtz and Brooks 2008	Orthodox and embedded neoliberalism (1985-2003)	Partisanship, unionization, legacy of import substitution	*Disappear*	-	1874.22	1800.60	2.86	5.25
Baccaro and Rei 2007	Unemployment rate (2003)	Labor market rigidity	*Stronger*	-	565.89	115.86	2.76	-0.68
Ehrlich 2007	Tariff rate (1948-94)	Institutional access points	*Disappear*	-	265.92	86.98	0.61	0.92
Keefer 2007	Fiscal cost of financial crisis (various)	Electoral competitiveness, political checks and balances	*No change*		38.98	48.42	1.41	1.90

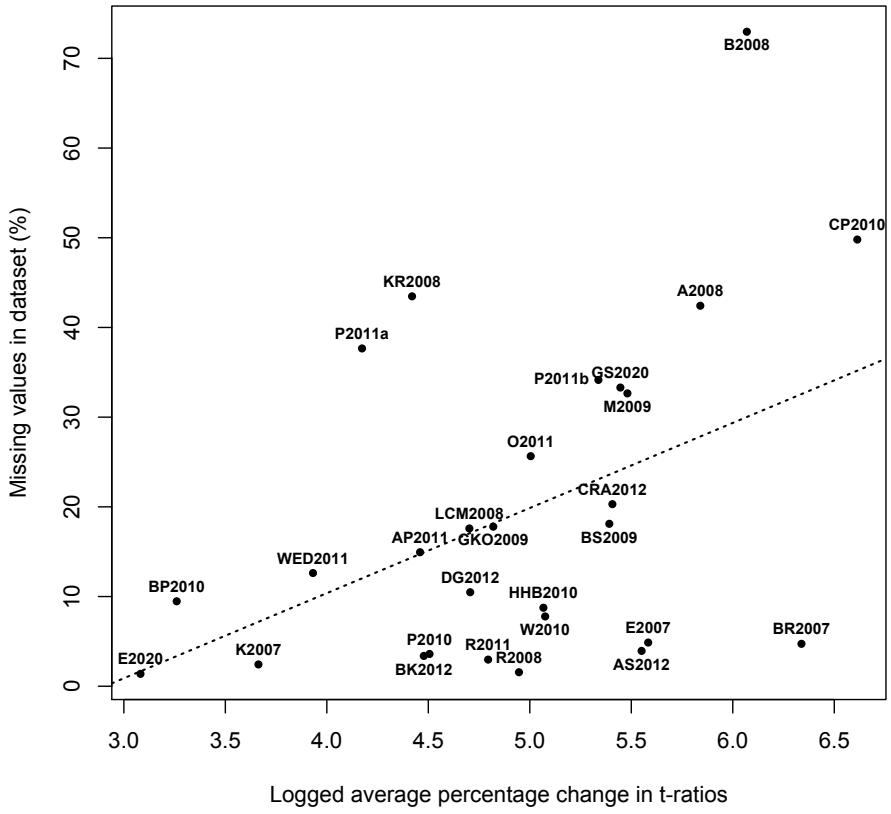
Notes: See text for technical definitions of reanalysis outcomes. The sign of change is positive if coefficients experiencing change lose significance at the 10 percent level and positive if they gain it. "All" refers to every explanatory variable in the main empirical analysis; "KEV(s)" stands for "key explanatory variable(s)." Percentage changes in t-ratios are based on absolute values. O/R ratio is the average ratio of original to reanalyzed coefficients.

pearance, averaging 552 percent for all coefficients and 2,201 percent for key coefficients. Note, however, that these averages are heavily influenced by two outliers in which both figures are more than two standard deviations above the mean (Accominotti and Flandreau 2008; Kurtz and Brooks 2008). Excluding these studies, the averages are 206 percent and 117 percent, respectively, similar to those in cases of strengthening (233 percent and 201 percent) and weakening (219 percent and 47 percent) but notably higher than those in cases of no change (135 percent and 74 percent).

The last two columns display the average ratio of original to reanalyzed coefficients (O/R ratio) for all variables (column 8) and only the key explanatory variable(s) (column 9). This ratio provides a measure of changes in the size of substantive effects (not accounting for uncertainty). A ratio higher than one indicates that coefficients have become smaller in the reanalysis, in which case the direction of change is likely to be negative; a ratio lower than one indicates that coefficients have become larger, in which case change is likely to be positive. Consistent with the fact that change is mostly negative, the ratio's absolute value exceeds one in more than 75 percent of studies for both all coefficients and key coefficients. Unsurprisingly, its mean deviation from one is higher in cases of disappearance (5.68 for all coefficients and 9.14 for key coefficients) than in cases of weakening (2.47 and 3.33), strengthening (1.37 and 0.58), and no change (1.21 and 0.77).

From a statistical point of view, the general tendency of coefficients to shrink and drop out of significance in the reanalysis is surprising. Bias can be positive or negative, and, other things equal, expanding the sample should lead to an *increase* in the size and significance level of coefficients. The outcome of the reanalysis indicates that there are systematic differences between the observations included in and excluded from the analyses — differences that are likely to render their inferences severely biased. It is thus consistent with the earlier discussion of the nonrandom process by which data tend to become missing in CIPE and the potentially sizable statistical changes that can occur

Figure 2 Relationship between Missingness and Change in the Reanalysis



Notes: Two studies in the reanalysis are excluded because their x-axis values are extreme outliers (Accominotti and Flandreau 2008; Kurtz and Brooks 2008).

when the sample is extended to all eligible observations in the dataset.

The results of the reanalysis are also consistent with Section 2.2's discussion of the conditions under which multiple imputation will make the greatest difference to inferences in CIPE. Excluding the two outliers mentioned above, there is a relatively high positive correlation (given the sample size) between the mean percentage change in t-ratios for all coefficients and (1) the percentage of missing values in the dataset ($r = 0.47$), a proxy for the level of missingness (see Figure 2); (2) the range of per capita incomes and Polity scores (mean $r = 0.24$), a proxy for political and economic heterogeneity; and

(3) the number of variables in the imputation model that have a correlation of $r \geq 0.25$ with a dummy variable indicating whether or not a given observation is included in the main analysis ($r = 0.17$), a proxy for the number of variables related to missingness.

In addition, the results provide support for Section 2.2's related predictions regarding which issue areas will experience the largest changes in results. The all-coefficient percentage change in t-ratios (again ignoring the two outliers) is above the mean (191 percent) in the issue areas of economic performance (399 percent), political regimes (201 percent), trade (223 percent), foreign aid (220 percent), governance (267 percent), and public goods (343 percent).[39] It is well below the mean, meanwhile, in studies of income inequality (141 percent), redistribution (90 percent), and policy diffusion (88 percent).

Finally, it is worth emphasizing that many of the changes in the reanalysis are significant from a *substantive* as well as a statistical perspective, altering our understanding of empirical relationships in ways that have important practical as well as theoretical implications. To offer a few notable examples from different issue areas: participation in World Bank loan programs reduces rather than increases the likelihood of major government crises (by 4.84 percent for each structural loan received in the previous year), challenging analyses suggesting that such programs tend to cause political turmoil by forcing governments to implement unpopular market-oriented reforms (Dreher and Gassebner 2012); the effect of oil and other nontax revenues on political stability is not positive but exactly zero, undermining recent theories positing that such revenues stabilize democratic as well as authoritarian regimes by enabling leaders to pursue fiscal policies that appease social groups who pose a threat to regime survival (Morrison 2009); and official exchange rate targets make virtually no difference to the effectiveness of fixed exchange rate regimes in controlling inflation (the predicted inflation rate for fixed-rate regimes

[39]Results in all six issue areas are also more likely than average to disappear or experience some form of change.

with targets is just 0.01 percent lower than that for all such regimes), a finding with important implications for the practice of central banking as well as theories of credible commitment in economic policymaking (Guisinger and Singer 2010).[40]

4 Conclusion

This article has shown that multiple imputation can make a substantial and striking difference to existing empirical knowledge in political science by reanalyzing the results of a large number of recently published studies in the area of CIPE. The results of the reanalysis naturally raise serious questions about the validity of many of the statistical findings accepted by the CIPE community in recent years. At the very least, sizable changes in parameter estimates should encourage CIPE scholars to reflect carefully on the scope conditions — both spatial and temporal — of their theoretical propositions. As suggested in Section 3.3, in many cases such changes may warrant a more fundamental re-examination of the assumptions and causal mechanisms underlying these propositions. This exercise may open up interesting and fruitful avenues for further research.

The article's findings also have significant implications for quantitative research in other areas of political science. Since the pattern of missing values in political science datasets is probably *never* completely random, as discussed in Section 2.1, inferences produced by listwise deletion are likely to always be (to some extent) biased as well as inefficient. While adopting multiple imputation can be expected to reduce bias and alter parameter estimates in most studies, such changes will be largest under the three conditions set out in Section 2.2: (1) the proportion of missing data is high; (2) the dataset contains a large number of variables that are strongly related to missingness; and (3) hypotheses are tested on a heterogeneous sample in terms of missingness correlates.

[40] These substantive effect estimates are based on the results of, respectively, Model 1 in Table A4, Model 2 in Table A22, and Table A17.

Where, other than CIPE, are these conditions likely to be satisfied? While it is hard to generalize about domestic studies, if wealthier and more democratic nations also tend to have more complete *non*economic data — which seems plausible — we might expect the first condition to be only rarely satisfied in American politics and areas of comparative politics that focus primarily on advanced democracies, such as electoral institutions and party systems.[41] Similarly, we might expect cross-national studies in the latter areas to be less likely to meet this condition than those focusing on poorer and less democratic nations, which are common in areas such as state-building and clientelism. Neither type of study, however, is likely to meet the second and third conditions: since they focus on a relatively homogeneous set of nations, missingness in their datasets is liable to depend primarily on unmeasured idiosyncratic factors (see Section 2.2). The two conditions are thus more likely to be fulfilled in areas such as such as political violence and regime change in comparative politics and security studies in international relations, where analysts typically test their propositions on a diverse sample of countries and variables of theoretical interest are known to be strongly related to missingness (Gleditsch 2002; Fearon and Laitin 2003; Ross 2004). Given that such variables themselves tend to have a high proportion of missing values — ensuring that all three conditions are satisfied — multiple imputation is likely to make as large a statistical and substantive difference as in many of the CIPE studies reanalyzed in this article.

These are, of course, only preliminary and very general speculations. As illustrated by the example of CIPE, there is considerable variation in missing-data patterns within as well as across subfields. Neither these patterns nor the effects of substituting multiple imputation for listwise deletion can be properly ascertained without the kind of systematic empirical examination conducted in this article. Expanding this investigation

[41]Survey data are unlikely to be an exception: King et al.'s (2001) review of survey-based articles published in five leading political science journals in the period 1993-97 found that they lost an average of around one-third of their data due to listwise deletion, implying a relatively high missing-data rate.

to other areas of the discipline is an important task for future methodological research.

References

Accominotti, Olivier and Marc Flandreau. 2008. "Bilateral Treaties and the Most-Favored-Nation-Clause: The Myth of Trade Liberalization in the Nineteenth Century." *World Politics* 60 (2):147–188.

Allison, Paul D. 2002. *Missing Data*. Thousand Oaks, CA: Sage Publications.

Allison, Paul D. 2005. "Imputation of categorical variables with PROC MI." Paper 113-30, 30th Meeting of SAS Users Group International (SUGI 30). `http://www2.sas.com/proceedings/sugi30/113-30.pdf`.

Bodner, Todd E. 2008. "What improves with increased missing data imputations?" *Structural Equation Modeling* 15 (4):651–675.

Collins, Linda M., Joseph L. Schafer and Chi-Ming Kam. 2001. "A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures." *Psychological Methods* 6 (4):330–51.

Cranmer, Skyler J. and Jeff Gill. 2013. "We Have to Be Discrete About This: A Non-Parametric Imputation Technique for Missing Categorical Data." *British Journal of Political Science* 43 (2):425–449.

Dreher, Axel and Martin Gassebner. 2012. "Do IMF and World Bank Programs Induce Government Crises? An Empirical Analysis." *International Organization* 66 (2):329–358.

Edwards, Martin S., Kelsey A. Coolidge and Daria A. Preston. 2011. "Who Reveals? Transparency and the IMF's Article IV Consultations." Seton Hall University Working Paper Series. `http://wp.peio.me/wp-content/uploads/2014/04/Conf5_Edwards-30.09.11.pdf`.

Fearon, James D. and David D. Laitin. 2003. "Ethnicity, Insurgency, and Civil War." *American Political Science Review* 97 (1):75–90.

Gleditsch, Kristian Skrede. 2002. "Expanded Trade and GDP Data." *Journal of Conflict Resolution* 46 (5):712–724.

Graham, John W. 2009. "Missing Data Analysis: Making It Work in the Real World." *Annual Review of Psychology* 60:549–76.

Graham, John W., Allison E. Olchowski and Tamika D. Gilreath. 2007. "How many imputations are really needed? Some practical clarifications of multiple imputation theory." *Prevention Science* 8 (3):206–213.

Graham, John W., Scott M. Hofer and David P. MacKinnon. 1996. "Maximizing the Usefulness of Data Obtained with Planned Missing Value Patterns: An Application of Maximum Likelihood Procedures." *Multivariate Behavioral Research* 31 (2):197–218.

Guisinger, Alexandra and David A. Singer. 2010. "Exchange Rate Proclamations and Inflation-Fighting Credibility." *International Organization* 64 (Spring):313–337.

Hardt, Jochen, Max Herke and Rainer Leonhart. 2012. "Auxiliary variables in multiple imputation in regression with missing X: A warning against including too many in small sample research." *BMC Medical Research Methodology* 12 (1):184–197.

Hollyer, James R., B. Peter Rosendorff and James Raymond Vreeland. 2011. "Democracy and Transparency." *Journal of Politics* 73 (4):1191–1205.

Honaker, James, Gary King and Matthew Blackwell. 2011. "Amelia II: A Program for Missing Data." *Journal of Statistical Software* 45 (7):1–47.

Horton, Nicholas J., Stuart R. Lipsitz and Michael Parzen. 2003. "A Potential for Bias When Rounding in Multiple Imputation." *The American Statistician* 57 (4):229–232.

Houle, Christian. 2009. "Inequality and Democracy: Why Inequality Harms Consolidation but Does Not Affect Democratization." *World Politics* 61 (4):589–622.

Keefer, Philip. 2007. "Elections, Special Interests, and Financial Crisis." *International Organization* 61 (3):607–41.

King, Gary, James Honaker, Anne Joseph and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95 (1):49–69.

Kropko, Jonathan, Ben Goodrich, Andrew Gelman and Jennifer Hill. 2014. "Multiple Imputation for Continuous and Categorical Data: Comparing Joint Multivariate Normal and Conditional Approaches." *Political Analysis* 22 (4):497–219.

Kurtz, Marcus J. and Sarah M. Brooks. 2008. "Embedding Neoliberal Reform in Latin America." *World Politics* 60 (2):231–280.

Lall, Ranjit. 2016. "Replication Data for: How Multiple Imputation Makes a Difference." `http://dx.doi.org/10.7910/DVN/CRLKIF`, Harvard Dataverse.

Little, Roderick J.A. 1988. "A Test of Missing Completely at Random for Multivariate Data with Missing Values." *Journal of the American Statistical Association* 83 (404):1198–1202.

Little, Roderick J.A. 1993. "Pattern-Mixture Models for Multivariate Incomplete Data." *Journal of the American Statistical Association* 88 (421):125–134.

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

Little, Roderick J.A. and Donald Rubin. 2002. *Statistical Analysis with Missing Data (Second Edition)*. Hoboken, NJ: Wiley.

Meng, Xiao-Li. 1994. "Multiple-imputation inferences with uncongenial sources of input." *Statistical Science* 9, no. 4:538–558.

Morrison, Kevin M. 2009. "Oil, Nontax Revenue, and the Redistributional Foundations of Regime Stability." *International Organization* 63 (1):107–38.

Obinger, Herbert and Carina Schmitt. 2011. "Guns and Butter? Regime Competition and the Welfare State during the Cold War." *World Politics* 63 (2):246–270.

Ross, Michael. 2006. "Is Democracy Good for the Poor?" *American Journal of Political Science* 50 (4):860–874.

Ross, Michael L. 2004. "What Do We Know About Natural Resources and Civil War?" *Journal of Peace Research* 41 (3):337–356.

Rubin, Donald B. 1976. "Inference and Missing Data (with Discussion)." *Biometrika* 63:581–592.

Rubin, Donald B. 1977. "Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys." *Journal of the American Statistical Association* 72 (359):538–43.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley & Sons.

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 (394):366–74.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. London, UK: Chapman and Hall.

Scheve, Kenneth and David Stasavage. 2009. "Institutions, Partisanship, and Inequality in the Long Run." *World Politics* 61 (2):215–253.

van Buuren, Stef. 2012. *Flexible Imputation of Missing Data*. Boca Raton: Taylor and Francis.

von Hippel, Paul T. 2009. "How To Impute Squares, Interactions, and Other Transformed Variables." *Sociological Methodology* 39 (1):265–91.

von Hippel, Paul T. 2013. "Should a Normal Imputation Model Be Modified to Impute Skewed Variables?" *Sociological Methods and Research* 42 (1):105–138.

White, Ian R., Patrick Royston and Angela M. Wood. 2011. "Multiple imputation using chained equations: Issues and guidance for practice." *Statistics in Medicine* 30 (4):377–399.