Imputing Across Types of Item Nonresponse in Surveys

Natalie Jackson
nmjb09@gmail.com

ABSTRACT

Survey analysts often treat nonresponse on a given variable homogeneously, whether it comes from a refusal, a "don't know," or a skipped answer. In this work we show that these responses come from fundamentally different causes within a given individual and that the subsequent analytical treatment gives important differences in estimated models. A useful feature of the 2012 American National Election Study allows us to differentiate levels of reluctance to answer, and therefore to model those differences.

1 Introduction

In an ideal world, survey respondents would answer every question with a substantive response based on their existing knowledge and opinions. Since they do not, survey researchers have to decide how to deal with nonresponse behavior when respondents answer some questions and not others. Complicating the researchers' considerations, it is clear that there are different types of item nonresponse: The respondent could say they "don't know" the answer to a question, they could refuse to answer the question, or they could skip the question with no indication of why (this is most common on web surveys). Is it appropriate to treat all types of item nonresponse in the same way? Furthermore, when a respondent refuses to answer a question, no matter what the mode of the instrument, can we distinguish between a nonresponse that comes from ignorance of the issue or a nonresponse that conceals extreme, awkward, or embarrassing attitudes that the respondent possesses? We will demonstrate that these questions fundamentally affect survey analysis results and the questions cannot be simply ignored by researchers.

In this work we seek to distinguish between two theoretically distinct forms of item nonresponse: (1) nonresponse because the respondent does not possess an opinion, and (2) nonresponse because the respondent declined to state their existing opinion. Since the conventional approach to handling these responses is to treat them as missing data, there are inferential consequences to making assumptions about the source of the nonresponse. Since the two most common ways to deal with missing data in survey research are listwise deletion or imputation (see Appendix A), the resulting models are critically affected by the decisions researchers make about handling item nonresponse.

Our core theory is that item nonresponse because the respondent does not have an opinion creates missing at random (MAR) data, and item nonresponse because the respondent declined to state their existing opinion creates nonignorable (NI) missing data. To support this theory we first work to define the roots of the types of nonresponse in survey practice. We then explain how the two types of item nonresponse produce two fundamentally different types of missingness in the Rubin (1978) sense. Finally, we use ANES 2012 data to empirically show evidence of the distinction between the two types of item nonresponse, show evidence that these two types have different causes, and use imputation to highlight the differences in a modeling context and reduces biases in the associated coefficient estimates.

2 The Typology of Item Nonresponse

In the Total Survey Error framework, nonresponse error is one of the five major components contributing to errors in sample surveys, alongside sampling error, coverage error, measurement error, and postsurvey error (Weisberg 2009). Sampling error and coverage error are widely discussed in the effort to achieve representative sampling, measurement error receives quite a bit of attention in questionnaire construction, and postsurvey error is usually addressed by documenting and describing weighting and recoding techniques. But much of the survey analysis in academic journals and the media simply ignores nonresponse error unless the topic of the research is such forms of error.

Nonresponse error is comprised of both item-level nonresponse, in which respondents do not answer one or more questions on the survey, and unit-level nonresponse, where the sampled unit fails to respond to the survey at all. This work is concerned only with the former. Although unit-level nonresponse is extremely important to survey research, most researchers do not have to make decisions about how to handle unit nonresponse in data analysis; there is nothing one can do with completely nonexistent data. We are addressing item-level nonresponse, in which there exist partial data from a respondent, but for some reason the respondent did not answer all of the questions. Researchers have to decide what to do in this case - whether to use only the complete data and ignore the incomplete cases or impute the data somehow. Frequently, though, the question receives little attention and the problem is dealt with using listwise deletion, despite the fact that doing so is only statistically acceptable when the missingness is completely at random (MCAR) (Rubin 1978). Missingness in sample surveys is usually not completely at random; people have reasons for not answering questions.

We can derive the possible reasons for not answering questions from the types of item nonresponse typically recorded on surveys: (1) Saying "don't know" or the equivalent, if it is offered as an option, (2) "Refusing" to answer the question, if this is offered as an option, and 3) Skipping the question without selecting any option at all.

- Don't Know This comes from respondents who truly do not know the answer to a question. In practice, it is not clear whether respondents choose "don't know" because they really do not have an attitude, because none of the answer options quite fit their true attitude, or because they don not want to state their attitude. Social desirability and many other biases provide clear justification for why a respondent would decline to state an attitude and say "don't know"

instead.

- Refuse This originates from respondents who do not want to provide their answer to the question. Incidence of outright refusals is generally very low, which makes sense given the strong stance it requires, particularly in a face-to-face mode. Refusals to answer can often be used in the same way that "don't know" can be used - meaning that the assumption is that the respondent is taking a hard line in refusing to answer a question because they have an answer but do not want to give it, or because they do not have an answer to provide. Explicit refusals are typically rare responses.

- Skip This response occurs mostly with online survey modes, and is completely ambiguous. It is unclear what a skip means, whether it is really a "Don't Know," a refusal, or an accidental button click forward on a web survey. We have no information on that particular question by which to make this distinction in that mode. Note that this is not referring to *skip patterns* that are programmed into the survey in which a respondent is not asked a question because it was deemed not relevant based on branching or skipped over in a randomization assignment.

Despite these differences there are practical reasons for treating the three types of nonresponse as the same. We often cannot be certain how the respondent intended their response, given the above possible explanations. Mode effects complicate our understanding - the ability of the respondent to use skip versus record a "don't know" or refuse response varies based on the instrument and how it is delivered. Interviewer-administered (telephone or face-to-face) surveys usually have options for "don't know" and refuse, and sometimes even "skip" options, that interviewers can select. Sometimes these are not read to the respondent, so an interviewer is making the decision on how to categorize the lack of response to that item. Interviewers' categorizations can vary widely depending on what the respondent says, the interviewer's training, and the clarity of the interviewer-respondent interaction during the survey.

Self-administered (online or mail) surveys lack this interaction; the respondent selects from the available options and is at the mercy of the survey designer to put in adequate response options. These surveys vary a lot between providing explicit "don't know" and refuse options, and not providing a "don't know" option, but allowing respondents to skip questions. In the latter case, a skip could be interpreted as either a "don't know" or a refuse response, or it could be the case that the respondent accidentally advanced the survey without answering but intended to provide an answer.

Thus in both the interviewer-administered and the self-administered survey modes, the lines between the three types of item-level refusal are blurred.

3 Classifying and Treating Missing Data in Surveys

From the previous discussion it is obvious that we have to pay attention to the item nonresponse etiology. Surveys are different from many other social science datasets in that each case is an individual person who might have distinct reasons for not responding to each specific item, as discussed above, which means we have to consider the human element in labeling what type of missing data we have. Missing data on one question might not have the same root cause as missing data on another question. That means the type of missing data – missing completely at random (MCAR), missing at random (MAR), or non-ignorable (NI) – could differ based on the reason for the missingness.

In general, we know that survey item nonresponse is not MCAR, or really even MAR in many cases. MCAR would be the easiest to deal with, as listwise deletion would not bias the results, but almost no item nonresponse is MCAR. It is incredibly unlikely that the human interactions with a survey instrument would result in completely random missing data. Sometimes data are missing at random (MAR), meaning that the mechanism that caused the missingness is conditional data that are observed. MAR is possible in survey data when we know and have measured the respondent characteristics that correlate with item nonresponse on a particular question. Under MAR listwise deletion leads to biased model results and the only question is the extent of the bias. Fortunately a wide range of imputation tools are available for MAR data. Item nonresponse for which we do not know the respondent correlates and have not measured the causes of becomes a more substantial problem. These missing values are NI, meaning that they are conditioned on other data that we cannot see and regular multiple imputation does not solve this problem. Researchers still need to confront this problem and it is usually done by adding assumptions, subsetting the data, or additional data collection.

For purposes of missing data classification and treatment, then, we need to differentiate between the types of item nonresponse, but these three categories ("don't know," refuse, skip) do not help much given the blurred meanings. Theoretically, to classify the type of missing data, we need to separate the two basic causes behind any of the above item nonresponse categories: Either the

respondent has an opinion and did not report it for some reason, or the respondent has no opinion to report. The difference between those two causes is critical for how to handle item nonresponse as missing data, as they are fundamentally different types of missing data. There is an important gap in understanding between these two causes of item nonresponse.

- True Nonattitudes Social science literatures can generally adequately identify why respondents would not have an opinion or possess the factual knowledge on an issue to answer a question. Often these nonattitudes can be explained by respondent characteristics, such as lack of exposure to the topic and lower education and interest levels, as well as other such demographics. These variables are typically measured and can be used to explain a true lack of opinion. The missingness certainly is not random (not MCAR), but possibly qualifies for MAR.

- Hidden Attitudes The other cause of nonresponse – choosing not to report an attitude or opinion that the respondent does hold – is more problematic. As described above, there are many reasons a respondent might decline to state an opinion they have. Those might depend on variables measured, such as partisanship or ideology, a lack of certainty in opinion which might be explained by interest or education, or other demographics. But many of the possible reasons are unmeasured: Social desirability or other bias perceived by the respondent, a sensitive question for which the respondent does not want to reveal their response, the nature of the respondent's interaction with an interviewer, or privacy concerns - these are only a few possibilities. The missingness is not random, and cannot be accounted for with variables measured in the survey, making it non-ignorable.

There are a few options for dealing with item nonresponse regardless of the cause. One alternative is that these respondents are simply case-wise deleted from any analysis of the data. Since the data are not MCAR, this is inappropriate and leads to bias. Another strategy is to treat to treat item nonresponse as their own analysis categories. The appropriateness of this approach depends on what the researcher hopes to gain from the analysis, but is typically not useful unless the analysis focuses on item nonresponse. We are left with multiply imputing the missing responses as the most statistically appropriate way to deal with item nonresponse (Rubin 1978, 2004), which has been explored in survey research at length (e.g., deLeeuw *et al.* 2003, Andridge and Little 2010). But

the distinct causes for missing data described above brings up a few questions that have not, to our knowledge, been addressed in the literature:

- Are these two causes of missing data discernible in survey data: can we model the difference between true nonattitudes and hidden attitudes?
- Can we treat all missingness equally given the theoretical distinction between true nonattitudes and hidden attitudes?
- Is it appropriate to impute a response for analytical purposes when the respondent truly does not have an opinion?

4 Data

To test our theory about the connection between attitude existence and nonresponse we use the American National Election Studies (ANES) 2012 Time Series Study. This is the 29th installment in a longstanding series of election studies that go back to the US Presidential election in 1948. The 2012 data were selected due to ongoing updates to the 2016 data when this project was started (see "Errata" on the ANES 2016 website for details). The 2012 edition was the first in the ANES series to implement a dual-mode design by incorporating a traditional ANES face-to-face sample as well as a separate sample interviewed on the internet.

Our analysis focuses on the face-to-face portion of the survey in order to avoid the confounding effects of multiple modes in the same analysis. It has been shown that mode matters in this context (Homola *et al.* 2017), and ANES is no different: The rates of item nonresponse are quite different in the web portion than the face-to-face portion, likely for the reasons outlined above - a response is more likely to be coded as a "don't know" or refuse by an interviewer than on a self-administered survey that might not have such responses available. It is not exactly clear in the ANES questionnaire documentation how web respondents saw item nonresponse options, but the frequencies show clearly that there are fewer "don't know" and refuse responses on the web version.

The ANES for 2012 provides a convenient way to test the theory of two types of item nonresponse by using two questions to measure respondent ideology. The first is a seven-point ideology scale that goes from "extremely liberal" to "extremely conservative." Of the 2,054 face-to-face respondents in the survey, 590 opted to not answer this question by saying "don't know," refusing, or an additional

option for this question: "Haven't thought much about this." We classified "Haven't thought much about this" as item nonresponse because although it is a fairly common type of response option for ideology questions that respondents might truly not have thought about, it does not provide a substantive answer to the question that can be used in analysis. The second question in the respondent ideology measure is asked of those who did not provide a substantive response to the first question for any reason (the 590 respondents), plus those who answered "Moderate" on the scale. It asks "if you had to choose would you consider yourself a liberal or a conservative?" with "Moderate" as an unlisted, but accepted, option.

When the data from these two variables are combined, 128 respondents still have not provided a substantive answer to the question, but 462 of the original 590 nonresponders did provide a substantive response when asked the second question. These can be interpreted as the two groups of item nonresponse described above: 128 are true nonattitudes – they do not have an ideological viewpoint to give. It could be that they are simply adamant in not revealing their preference, but the second probing question makes this less likely. The 462 who did answer the second question, but not the first, are the hidden attitudes. For whatever reason, they did not want to reveal their preference, but they do have a preference as revealed in the second question. This process is outlined in Figure 1.

The explanatory variables used in this stage of the analysis are a fairly typical set of alternatives in the literature. To measure sex differences we use `Gender`, which is dichotomized for male=1, and female=1. The respondents' interest in politics and elections is usually important with these models, `Interest`, with the responses Always, Most of the time, About half the time, Some of the time, and Never. Employment status is measured with the variable `Employed`, dichotomized for fully employed 1, not 0. Race is condensed in the variable `Nonwhite` such that white=0, and nonwhite=1. The respondent's age is given by `Age` giving rounded down years at the time of the survey ranging from 17 to 90. Party ID is measured as a dichotomous variable for self-identifying as a `Democrat` with any strength, and a dichotomous variable for self-identifying as an `Independent`, meaning that self-Identification as a Republican is the reference category in this treatment contrast. Finally, the variable `Education` goes from lowest to highest (less than high school, high school or equivalent, some college, college degree, and graduate degree). Any missing values for these variables were imputed using the `R` package `mice` (multiple imputation using chained equations) prior to the analyses below. The ideology variables were not imputed in that process.

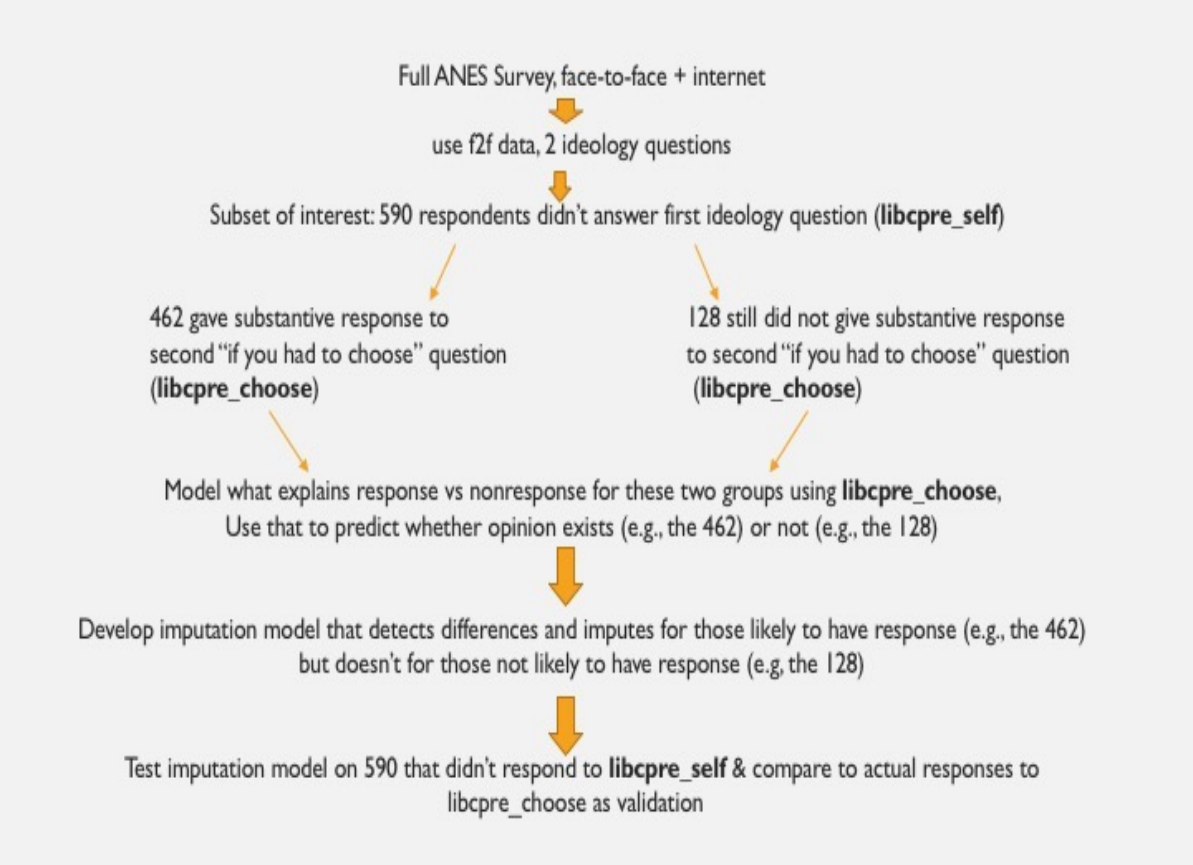

Figure 1: Analysis of Types of Missing Data

5 Distinguishing Between Two Types of Missingness Etiologies

To distinguish between the two types of cognitive missingness we construct two models with these data. The first model asks why 590 out of 2054 face-to-face respondents not provide a substantive response to the first ideology question. To make this test as clear as possible we used a dichotomized version of this variable (1464 : 0, 590 : 1), since out interest is not on the distribution of expressed ideology but rather than on the nonresponse. The second model asks whether we can learn about the different types of cognitive processes by looking at the difference between those that expressed an ideology when the "had to choose" language was added. In this case we use the combined ideology variable and dichotomize it so that we isolate the 128 hard-core refusers who refused on both opportunities (presumed true nonattitudes), and the 462 new cooperators (presumed hidden attitudes) are folded back into those giving an ideology response.

The results for the first model explaining the 590 refusals and "Don't Know" responses with

Table 1: Modeling Missing Values at Stage One, Full Sample

	Estimate	Std. Error	t value	Pr(>|t|)	95% LCI	95% UCI
(Intercept)	-1.55	0.34	-4.55	0.00	0.12	0.37
Female	0.28	0.12	2.36	0.01	1.09	1.60
Interest	0.40	0.05	7.43	0.00	1.37	1.63
Employed	0.10	0.13	0.80	0.21	0.90	1.36
Nonwhite	0.50	0.13	3.87	0.00	1.33	2.04
Age	-0.00	0.00	-0.57	0.28	0.99	1.00
Democrat	0.36	0.14	2.59	0.00	1.14	1.80
Independent	1.21	0.18	6.55	0.00	2.48	4.55
Education	-0.53	0.06	-8.94	0.00	0.53	0.65

Null deviance: 2228.22 on 2053 degrees of freedom
Residual deviance: 1886.57 on 2045 degrees of freedom
AIC: 1743.101, Adjusted Degrees of Freedom from 10 Imputations: 556.9445

a standard logit model are given in Table 1. Notice that Model 1 fits extremely well both in terms of individual Wald statistics where 7 of the 9 coefficient estimates are statistically reliable at conventional levels, and the improvement in summed deviance over the null (mean) model which is easily in the tail of a chi-square comparison. We can reliably assert that those who do not provide an ideology answer are more likely to be: female, more interested in politics, nonwhite, Democrat versus Republican, and Independent versus Republican. Higher levels of education imply a great proclivity towards giving an ideology value.

Next we turn to the model in which there is no "Don't Know" options and we have 128 hard core refusals. For comparison sake we use exactly the same explanatory variables on the right-hand side of the model to understand this second outcome. Our claim is that many of those who said "Don't Know" or refused equivalent to first ideology question ($n = 590$) did answer the second ideology question when pushed (n=462), indicating that they did have an opinion, but were trying not to reveal it. Those who did not give an answer to the second question fall into the category of likely truly not having a position ($n = 128$). The results are given in Table 2.

As with Model 1, this is a very good statistical fit. Six of the nine Wald statistics indicate statistically reliable coefficients. Nonwhite joins employment status and age as not reliable. The summed deviance indicates a reliable increase in overall fit over the null (mean) model, a difference that passes the standard chi-square test. An interesting difference from the first model is not only are all of the reliable coefficient estimates in the second model are larger (and signed in the same

Table 2: Modeling Missing Values at Stage One, True Nonattitudes

	Estimate	Std. Error	t value	Pr(>|t|)	95% LCI	95% UCI
(Intercept)	-3.83	0.68	-5.62	0.00	0.01	0.07
Female	0.47	0.22	2.17	0.02	1.12	2.30
Interest	0.47	0.10	4.50	0.00	1.35	1.89
Employment	0.03	0.23	0.12	0.45	0.70	1.50
Nonwhite	0.15	0.23	0.66	0.25	0.80	1.71
Age	-0.00	0.01	-0.64	0.26	0.99	1.01
Democrat	1.23	0.34	3.59	0.00	1.95	6.04
Independent	2.49	0.37	6.76	0.00	6.57	22.06
Education	-0.56	0.11	-5.12	0.00	0.47	0.68

Null deviance: 849.4 on 1591 degrees of freedom
Residual deviance: 667.8 on 1583 degrees of freedom
AIC: 625.0275, Adjusted Degrees of Freedom from 10 Imputations: 86.06016

direction), indicating a stronger effect of the explanatory variables on explaining refusal. This implies that the 128 refusals modeled here are qualitatively different than the 590 refusals modeled before. Restated, the factors that lead to refusal are the same with the hardcore refusers, they are just stronger. This supports our claim that there is a different cognitive process between the groups that matters in terms of modeling and inference.

An important question is if the group of 462 casual refusers are fundamentally different than the 128 hardcore refusers. Table 3 and Table 4 compares these two groups on the explanatory variables used in the two models. These tables consist only of the 590 missing cases, with the 128 coded as 1, and the 462 coded 0. Table 3 shows that the 128 true nonattitudes are more likely to be nonwhite, less interested in politics and elections, Democrat, and Independent, compared to the 462 with hidden attitudes.

Notice from Table 4 that the difference of means test shows that only `nonwhite` and the party ID variables statistically reliable differences amongst the explanatory variables. This suggests that race and partisanship play an important role in determining the strength of refusal where there are less nonwhites in the hardcore refusals, fewer Democrats, and more independents. It is important to remember that these are just the explanatory variables used in the two models, there are many more in the full dataset, as well as those not modeled, that could affect the observed difference. However, based on these results, it seems clear that the true nonattitude group and the hidden attitude group share characteristics, but are also different in critical ways – they are less interested in politics and

elections, more likely to be independents (and less likely to be Democrats), and more likely to be white than the hidden attitude group.

Table 3: Modeling Missing Values at Stage One, True Nonattitudes vs. Hidden Attitudes

	Estimate	Std. Error	t value	Pr(>|t|)	95% LCI	95% UCI
(Intercept)	-2.36	0.73	-3.22	0.00	0.03	0.32
Female	0.15	0.23	0.65	0.26	0.80	1.70
Interest	0.18	0.10	1.75	0.04	1.01	1.42
Employed	0.10	0.24	0.41	0.34	0.75	1.63
Nonwhite	-0.36	0.25	-1.42	0.08	0.46	1.06
Age	-0.00	0.01	-0.43	0.34	0.99	1.01
Democrat	1.17	0.37	3.15	0.00	1.75	5.92
Independent	1.78	0.38	4.69	0.00	3.18	11.10
Education	-0.15	0.14	-1.06	0.15	0.69	1.08

Null deviance: 531.22 on 589 degrees of freedom
Residual deviance: 494.71 on 581 degrees of freedom
AIC: 465.1297, Adjusted Degrees of Freedom from 10 Imputations: 72.40175

Table 4: Explanatory Variables for Different Classes of Refusers

	462 Sample		128 Sample		t-test
	Mean	Std.Dev.	Mean	Std.Dev.	Pr(>|t|)
Female	0.63	0.48	0.66	0.48	0.52
Interest	3.14	1.11	3.24	1.23	0.39
Employed	0.56	0.50	0.56	0.50	0.97
Nonwhite	0.73	0.44	0.60	0.49	0.01
Age	43.31	16.38	42.21	18.80	0.55
Democrat	0.69	0.46	0.59	0.49	0.05
Independent	0.14	0.35	0.30	0.46	0.0001
Education	2.30	0.95	2.20	0.99	0.34

6 Effects of Types of Nonresponse on Imputed Data

With the differences between the groups of nonresponders established, we turn to the issue of what happens when we impute across the types of item nonresponse. In this section we use a form of imputation tailored for discrete random variables to impute the missingness in the two refusal groups. In both cases we are imputing the refusals that constituted the positive dichotomous outcomes (refusals) in Model 1 and Model 2 and them using them as an additional explanatory variable to

predict vote choice in the 2012 election.

To set up this analysis, we reverted back to the full ANES face-to-face survey dataset, prior to the recodes and imputations done in the preceding analysis. We trimmed the dataset down to only demographics and a small set of opinion questions that are typically used in a model to predict vote choice: Gender, age, education, employment status, race, Hispanic identity, interest in politics, interest in elections, whether registered to vote, whether voted in the primary, party identification, presidential job approval, tea party support, what religion respondent thinks Obama is, and the state of the economy. Variables are dichotomized as appropriate. The first ideology question (with 590 missing cases) and the combined ideology variable (with only 128 missing cases) are both included. This pared-down dataset is used for all of the following analyses, including regular imputation.

In order to impute the missing values for all variables, included the two ideology variables, we use the `R` package named `hot.deck`. Since the variables of interest are measured categorically, hot deck imputation - and more specifically, multiple hot deck imputation, is the most appropriate way to handle missing data (see Appendix B). The `hot.deck` package uses the multiple hot deck procedure developed by Cranmer and Gill (2013), which combines repeated imputation and estimation methods from parametric multiple imputation methods with the hot decking procedures that are appropriate for categorical data, and does not force assumptions of normality and can be shown to have superior properties to alternatives for categorical data (Bailar and Bailar 1997).

The `hot.deck` package calls `mice` to impute continuous variables, then follows these steps to impute categorical variables: The algorithm creates 10 copies of the dataset (by default), then searches down the columns for missing data. When missingness is found on a categorical variable, the code computes a vector of affinity scores that measure how closely other cases match the case with the missing value. These affinity scores range from 0 to 1 indicating the proportion of exact matches in categorical variables, and the procedure randomly draws the value to impute from the best matches based on the affinity scores. This process of random draws from the best matches is repeated across the 10 datasets. See the appendix for details. The output from the package combines the `mice`-imputed variables and the hot decked variables into 10 fully-imputed datasets. The following models use all of the datasets, thereby accounting for the standard errors of the imputations in the analyses.

We imputed the data in two different forms to allow for comparisons: The full dataset, including all of the cases for which ideology was missing and imputed; and the full dataset minus the 128 true

nonattitudes, but for which all else was imputed. These two slightly different datasets allow us to look at model fit when we do and do not impute cases for which respondents genuinely do not have an opinion - which stems from the question of whether it is appropriate to impute an attitude when we believe there truly is none.

Table 5 shows the results of a model predicting whether the respondent voted for Obama in 2012, using the first ideology variable with 590 missing values (which were imputed). The overall model fit improves and magnitude of the ideology variable increases in the next model, shown in Table 6, which uses the information gained in the second ideology question, and only relies on imputation for the 128 true nonattitudes. In these models, partisanship is a 1-7 variable in which 1 is strong Democrat and 7 is strong Republican. Job approval is also 1-7, where 1 is strongly approve and 7 is strongly disapprove. All other variables are as previously described.

Table 5: Modeling Vote Choice, All Missing Ideology Imputed

	Estimate	Std. Error	t value	Pr(>|t|)	95% LCI	95% UCI
(Intercept)	5.27	0.53	9.91	0.00	81.19	467.08
Ideology	-0.29	0.07	-3.95	0.00	0.66	0.84
Female	-0.14	0.21	-0.67	0.26	0.61	1.23
Interest	0.04	0.08	0.51	0.31	0.91	1.19
Presidential job approval	-0.89	0.06	-14.71	0.00	0.37	0.45
Employed	-0.55	0.23	-2.37	0.03	0.40	0.85
Nonwhite	1.19	0.21	5.77	0.00	2.33	4.59
Age	0.00	0.01	0.71	0.25	1.00	1.01
Partisanship	-0.53	0.06	-8.48	0.00	0.53	0.65
Education	0.13	0.08	1.60	0.06	1.00	1.29

Null deviance: 2837.75 on 2053 degrees of freedom
Residual deviance: 1140.29 on 2044 degrees of freedom
AIC: 1110.454, Adjusted Degrees of Freedom from 10 Imputations: 38.60938

The two models predicting votes for are obviously similar, yet there are some interesting differences. While both models are statistically distinct from the null (mean) model in terms of summed deviance, the model with no true nonattitudes fits somewhat better (1100 versus 1140, also a reliable difference). The coefficient estimates are similar but some differences are notable. The variable indicating that the respondent is employed is statistically reliable and negative with the model that imputes all but is not reliable in the no nonattitudes model.

Interestingly, however, an even better model fit comes from listwise deleting the 128 true nonat-

Table 6: Modeling Vote Choice, Ideology True Nonattitudes Imputed

	Estimate	Std. Error	t value	Pr(>|t|)	95% LCI	95% UCI
(Intercept)	5.92	0.58	10.28	0.00	144.58	962.81
Ideology	-0.42	0.08	-5.39	0.00	0.58	0.75
Female	-0.20	0.21	-0.95	0.19	0.58	1.16
Interest	0.02	0.08	0.30	0.39	0.90	1.17
Presidential job approval	-0.91	0.06	-15.88	0.00	0.37	0.44
Employed	-0.56	0.23	-2.49	0.02	0.39	0.83
Nonwhite	1.23	0.21	5.98	0.00	2.44	4.80
Age	0.00	0.01	0.84	0.20	1.00	1.01
Partisanship	-0.50	0.06	-8.10	0.00	0.55	0.67
Education	0.10	0.08	1.24	0.12	0.97	1.27

Null deviance: 2837.75 on 2053 degrees of freedom
Residual deviance: 1099.65 on 2044 degrees of freedom
AIC: 1070.551, Adjusted Degrees of Freedom from 10 Imputations: 21.28132

titudes from the dataset and imputing the 462 hidden attitudes (Table 7). Unsurprisingly, the best model, in Table 8, is created by listwise deleting the 128 true nonattitudes and using the data we get from the second ideology question to fill the 462 hidden attitudes. This leads to the question of quality control of respondents. While casewise deletion is generally ill-advised, it does make sense when the dropped respondents produce answers very low quality, and these cases are not fundamentally different than the rest of the sample. In other words, as sample can be polluted by uninformative cases.

6.1 What About Missingness On the Outcome Variable?

Previously we evaluated ideology on the left-hand-side of regression models to demonstrate differences in nonresponse types. The issue of handling missing data for the outcome variable remains unresolved in the literature. However, our purpose is to demonstrate that different handling of an important litmus test question in American political behavior has consequences related to the handling of missingness.

Recall that the imputation process doesn't care about the eventual model, so including the $\mathbf{y}$ vector along with the $\mathbf{X}$ matrix has no effect on the imputation stage of the analysis. When there is missingness in $\mathbf{y}$ but $\mathbf{X}$ is complete then the incomplete cases contribute no information to the regression of $\mathbf{y}$ on $\mathbf{X}$ (Little 1992, p.1227, Little 1998), but $\mathbf{X}$ values can help imputing $\mathbf{y}$ values.

Table 7: Modeling Vote Choice, No True Nonattitudes, Others Imputed

	Estimate	Std. Error	t value	Pr(>|t|)	95% LCI	95% UCI
(Intercept)	5.75	0.62	9.29	0.00	113.06	865.29
Ideology	-0.34	0.07	-4.86	0.00	0.63	0.80
Female	-0.22	0.19	-1.21	0.12	0.59	1.08
Interest	0.01	0.08	0.10	0.46	0.88	1.16
Presidential job approval	-0.92	0.07	-13.77	0.00	0.36	0.45
Employed	-0.27	0.20	-1.36	0.10	0.54	1.06
Nonwhite	1.14	0.21	5.32	0.00	2.19	4.43
Age	0.01	0.01	0.87	0.21	0.99	1.02
Partisanship	-0.58	0.06	-9.56	0.00	0.50	0.62
Education	0.14	0.08	1.69	0.05	1.00	1.31

Null deviance: 2679.6 on 1925 degrees of freedom
Residual deviance: 1000.09 on 1916 degrees of freedom
AIC: 984.8838, Adjusted Degrees of Freedom from 10 Imputations: 15.09675

Controversy arises from performing the latter procedure since the goal of the regression model is to explain variance in $\mathbf{y}$ that is attributable to levels of $\mathbf{X}$ variables, thus to some political scientists this feels like "using the data twice." This is an unsupported concern. First, the model that produces the conditional posterior for imputation draws is different than the model that will be specified for research purposes. Therefore relationships in the data are used in different ways: "An imputation model does not represent causal relationships among the data." (Young and Johnson 2010). Second, the set of explanatory variables available for the imputation process on $\mathbf{y}$ is almost always different than the set of explanatory variables used in the final model specification: it is rare to have k covariates in a dataset and use exactly k variables on the right-hand-side of a model. For instance, the ANES 2012 Direct Democracy Study has 1037 variables available to researchers and of course it would be ridiculous to have anywhere near this number specified as explanatory variables. In addition, when there is missingness in both $\mathbf{y}$ and $\mathbf{X}$, the non-missing $\mathbf{y}$ values can contribute to the prediction of missing $\mathbf{X}$ values, as well as the reverse. Third, Graham (2009) notes that leaving $\mathbf{y}$ completely out of the imputation procedure imposes a zero correlation between $\mathbf{y}$ and all of the other variables which biases coefficients in the resulting model.

So these facts about outcome variable imputation are somewhat complicated about our general research strategy here. The ideology variables were used as outcomes in Model 1 and Model 2 to show the implications of the different types of refusal. Now the the refusals and "Don't Know"

Table 8: Modeling Vote Choice, No True Nonattitudes

	Estimate	Std. Error	t value	Pr(>|t|)	95% LCI	95% UCI
(Intercept)	6.43	0.67	9.58	0.00	205.98	1877.44
Ideology	-0.48	0.09	-5.55	0.00	0.53	0.71
Female	-0.27	0.19	-1.47	0.08	0.56	1.03
Interest	-0.00	0.08	-0.01	0.50	0.87	1.15
Presidential job approval	-0.94	0.07	-12.74	0.00	0.35	0.44
Employed	-0.27	0.20	-1.33	0.10	0.55	1.07
Nonwhite	1.25	0.22	5.60	0.00	2.42	5.03
Age	0.01	0.01	1.10	0.16	1.00	1.02
Partisanship	-0.55	0.07	-8.42	0.00	0.52	0.64
Education	0.09	0.08	1.15	0.13	0.96	1.26

Null deviance: 2679.6 on 1925 degrees of freedom
Residual deviance: 968.57 on 1916 degrees of freedom
AIC: 952.8098 , Adjusted Degrees of Freedom from 10 Imputations: 11.67363

values are contained within explanatory variables in two corresponding models and the effect is shown. Differences in the results demonstrate our hypothesis that underlying cognitive processes that differ lead to different "flavors" of missingness in the form of denying researchers an answer.

7 Limitations and Future Work

There are some obvious limitations to doing this type of analysis. The first is that it would be potentially infeasible to conduct such an analysis on every variable in a large model. The results could, and likely would, point to a different set of true nonattitudes for each variable, which would mean that the researcher still has a missing data problem with some items missing for some respondents and not others if the guidance of not imputing true nonattitudes is to be followed. The item nonresponse and missing data problem might be smaller than it was prior to the analysis, but it would not be resolved. Therefore, we recommend that this guidance only be followed for the main variable of interest in a model. If there is more than one variable of interest, it might be best to impute all cases, but be aware of the impact that including imputed true nonattitudes has on the data.

A second limitation is that ANES set up a clean way to distinguish types of item nonresponse in the ideology questions, but that does not exist for most questions. Modeling the true nonattitudes versus the hidden attitudes will be much more dependent on the theoretical background that bore

out in our analyses: True nonattitudes are more likely to correlate with a lack of experience with the topic, lower education, and indicators specific to having an opinion on the topic.

Additionally, these models need to be run on the ANES internet samples for comparison, as well as other datasets. Those additions are planned, but not executed at this time.

8 References

Allison, Paul D. Allison. 2001. *Missing Data.* Thousand Oaks: Sage.

Andridge, Rebecca R. and Roderick J. A. Little. 2010. "A Review of Hot Deck Imputation for Survey Non-response." *International Statistics Review* 78 (1): 40-64..

Bailar, John C.III and Barbara A.Bailar. 1997. "Comparison of the Biases of the 'Hot Deck' Imputation Procedure with an 'Equal Weights' Imputation Procedure." *Symposium on Incomplete Data: Panel on Incomplete Data of the Committee on National Statistics, National Research Council)* 422-447.

Cox, Brenda.G. 1980. "The Weighted Sequential Hot Deck Imputation Procedure." *Proceedings of the Section on Survey Research Methods, American Statistical Association.*, 721–726.

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, 425-449.

de Leeuw, Edith D., Joop Hox, and Mark Huisman. "Prevention and Treatment of Item Nonresponse." *Journal of Official Statistics* 19, 153-176.

Graham, John W. 2009. *"Missing Data Analysis: Making it Work in the Real World." Review of Psychology* 60, 549-576.

Jonathan Homola, Natalie Jackson, and Jeff Gill. 2016, "A Measure of Survey Mode Differences." *Electoral Studies* 44, 255=27.

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, 49-69.

Little, Roderick J. A. 1992. "Regression with Missing X's: A Review." *Journal of the American Statistical Association*, 87, 1227-1237.

Little, Roderick J.A. 1988. "Approximately Calibrated Small Sample Inference about Means from Bivariate Normal Data with Missing Values." *Computational Statistics & Data Analysis*, 161-178.

Little, Roderick J. A. and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data.* New York: Wiley, 2nd ed.

Rockwell, Richard C. 1975. "An Investigation of Imputation and Differential Quality of Data in the 1970 Census." *Journal of the American Statistical Association* 70 39-42.

Rubin, Donald B. 1978. "Multiple imputations in sample surveys- a phenomenological Bayesian approach." *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 20-34.

Rubin, Donald B. 2004. *Multiple Imputation for Nonresponse in Surveys.* Second Edition. New York: Wiley.

Weisberg, Herbert F. 2009. *The Total Survey Error Approach.* University of Chicago Press: Chicago.

Appendix A: Imputation in Survey Research

There is a considerable focus on nonresponse bias at the item level in the science of conducting representative sample surveys, so it follows that many survey researchers have paid particular attention to methods of adjusting for item nonresponse. Until recently, the most common method of addressing missing data in empirical models has been to delete every case with even a single missing value: a technique called *casewise deletion* (also called listwise deletion or euphemistically complete-case analysis). With casewise deletion, every row of the data matrix, corresponding to individual cases, are removed from the data if there is one or more missing values, which reduces the effective sample size for any subsequent model. If some survey respondents did not answer all of the items on a questionnaire needed for a particular analysis. This leaves a dataset of only respondents who answered every single necessary question for the model at hand. The *convenience* of this approach is obvious, making it tempting for empirical research.

Listwise deletion usually induces bias with political data, and survey respondents typically do not completely randomly choose to not answer a question. The extent of the bias induced by casewise deletion depends on how much of the data are missing, how far the pattern of missing values is from being completely random throughout the data matrix, and how different the missing data are from the observed data (Cranmer and Gill 2013, Little and Rubin 2002, 42). In the rare circumstance where missing values occur *completely* at random (MCAR), listwise deletion can be used without biasing the data; the only harm to subsequent analyses is the reduction in sample size and power. When data are not MCAR, however, the bias from casewise deletion can be negligible, extremely damaging, or anywhere in between (Allison 2001, Little and Rubin 1992, Rubin 1978). Replications of published political science work have shown that the bias induced by casewise deletion can be severe enough to produce faulty inferences (King, Honaker, Joseph, and Scheve 2010). It is difficult to imagine a situation in which casewise deletion is an ideal treatment of missing values since it is unbiased in very few realistic survey situations and always reduces the effective sample size. Therefore a consensus has developed in the statistical, econometric, and methodological literatures that missing values should be imputed using the best available method for the data and model at hand.

Traditional hot decking was developed in the 1970s by market researchers and census takers, and the term is a literal reference to drawing cards from a deck of computer punch-cards (e.g., Bailar

and Bailar 1997 Cox 1980, Rockwell 1975). Hot decking is a two-step process: first the sub-sample of cases most like the case with missing data is selected, then the value to impute into the missing datum is selected from within that subselection. In early hot deck imputation, if a respondent had a missing value, a set of "similar" respondents were pulled by hand from the data deck, and one of those "similar" respondents was randomly selected to provide a fill-in value for the missing datum. The biggest challenge is determining how similar the "similar" cases should be to the case with missing data in order to be part of the subset, but practitioners of the time seemed to do relatively well at performing this task. Hot deck imputation methods subsequently grew from simple random sampling to choose out of the subset of "similar" cases to more complicated algorithms in attempts to find respondents as similar as possible to those with missing responses.

Hot deck imputation, as it has been implemented thus far in sample surveys, has one primary disadvantage: the method is almost always deterministic in nature, which means there are no estimates of uncertainty included in the imputation values. When imputed values are produced using regression or other forms of modeling, the values have inadequate standard errors to indicate the certainty of the estimate as do any results stemming from statistical models since a single value is taken from another case and put into the place where a piece of data is missing (Little and Rubin 2004). The lack of an uncertainty measure in this method is problematic in that we do not know for sure that the accurate value was imputed, but by simply replacing a missing value with the value from a closely matching case and moving forward with analyses, we are pretending that it is the certain value for that case. Measures of uncertainty, such as a standard error, would allow for some adjustment to estimates based on the fact that we cannot be completely certain that the imputed value is the true value.

The best available method of imputation can be difficult to identify for survey data since many, if not most, survey responses are measured as categorical values. Standard forms of *multiple imputation* (Rubin 1978) are often deemed inappropriate for imputing categorical data, since the imputations take on unrestricted continuous values, producing unbiasedness under MAR but possibly unrealistic actual outcomes. However, multiple hot deck methods impute values defined within the dataset itself, making it an ideal method for imputing missing values in categorical-response survey questions, while still accounting for added certainty from the missingness like multiple imputation. The technique works best for discrete data because it relies on similarities between cases that have missing data and cases that are complete, and it is more likely that several respondents will have the same value

for discrete variables (e.g. gender, partisanship, or vote choice) than it is for two respondents to have the exact same value of a continuous variable before rounding. An additional advantage is that the algorithm for selecting the value to impute does not depend on a regression or other type of model that has to be "fit" to the data, and therefore the method is less subject to bias in subsequent model selection (Andridge and Little 2010).

Some have argued that this issue with hot deck imputation is not applicable to hot decking with sample survey data because in those analyses it is not the individual values that matter, but rather the ability to use the values to estimate population parameters based on the samples. The goal of imputation in this case is not to get the best estimate of the imputed value, but to fill in values that make sense in order to use the full dataset to make inferences (Little and Rubin 2002). However, obtaining the best possible imputed value and using the whole dataset for estimating a population are not mutually exclusive goals. If the variable is important enough to warrant going through the process of imputing missing values, it is very likely that the expectation is the variable will be used in analysis. Just as drawing a sample from a population results in error which affects the ability to infer to the population, choosing a value for an individual missing response from a much smaller sample of similar cases introduces error into the estimates that use the variable with imputed values. The latter error may be very small and inconsequential, but without an estimate of the uncertainty in the imputation it is impossible to know how that error affects the estimates. The effect of assuming no error, as the standard hot deck imputation method does, is to underestimate error in overall estimates which could lead to deceptively lower standard errors and inflated estimates of significance.

Appendix B: The Multiple Hot Deck Imputation Procedure

The problems associated with traditional hot deck imputation can be resolved by using the multiple hot deck procedure developed by Cranmer and Gill (2013). The procedure combines old-style hot decking with the repeated imputation and estimation methods generally used in parametric multiple imputation. Thus the problem of nonsensical imputations for categorical survey data which could result from other imputation methods is resolved by using the hot deck procedure, and the problem with deterministic hot-decking and assumed factual imputations is also resolved by the repeated imputation and analysis method built into multiple hot decking. Additionally, the method is truly

nonparametric, meaning that it does not require the assumptions of normality (or related) that parametric methods require–a notable advantage for survey data variables which are often far from normally distributed.

The steps of the multiple hot deck imputation algorithm as laid out in Cranmer and Gill (2013) are quite simple. The first step of the algorithm creates m copies of the dataset, each with both missing and observed values. There should be two or more copies of the dataset; the user defines how many in the provided `R` package named `hot.deck`, otherwise it uses a default of 5. The second step of the algorithm searches down each column of the dataset sequentially looking for missing values. When a missing value is found, a vector of affinity scores ($\boldsymbol{\alpha}$) is computed, which measures how close the other cases are to the one with missing data. Each individual affinity score, denoted $\alpha_{i,j}$, provides the degree of similarity the recipient case, i, has to each potential donor, j, in terms of the observed discrete covariate values in the $n \times k$ matrix $\mathbf{X}$. A perfect donor has exact matches in the $k-1$ observed values, and a completely unacceptable donor has no matches in the $k-1$ observe values. Obviously most potential donors lie in between these to extremes. Now define $z_{i,j}$ as the number of variables in which the potential donor j and the recipient i have different values. Therefore $k - z_{i,j}$ is the number of variables on which donor j and the recipient i are perfectly matched on these discrete variables. The affinity score is now defined as:

$$\alpha_{i,j} = \frac{k - z_{i,j}}{k-1}, \tag{1}$$

which ranges from 0 to 1. So essentially draws for categorical missing data are drawn from a normalized histogram of observed values where centrality is determined by levels of this affinity score.

Any case without missing data for the column of interested is considered in this process: only cases without missing values on this variable can be used as donors to impute the missing data. The affinity score is used to identify the closest matching cases for that missing value, from which imputations are randomly drawn. The procedure draws from the M donor values and imputes these values into the appropriate cells in the M datasets. This procedure is repeated until all missing values in the datasets have been imputed. Finally, the statistical model of interest is fit for each of the M datasets independently. The separate analyses are combined to produce a single estimate using the combination rules standard in parametric multiple imputation. These steps are outlined

below.

1. Create several copies of the dataset.
2. Search down columns of the data sequentially looking for missing observations.
   - When a missing value is found, compute a vector of affinity scores for that missing value.
   - Create the cell of best donors or calculate affinity scores and draw randomly to produce a vector of imputations.
   - Impute one of these values into the appropriate cell of each duplicate dataset.
3. Repeat Step 2 until no missing observations remain.
4. Estimate the statistic of interest for each dataset.
5. Combine the estimates of the statistic into a single estimate.

There are two different approaches to multiple hot deck imputation written into the `hot.deck` package. Each works according to the algorithm detailed above, but selection of the exact value to impute is slightly different for the methods. The first is the best cell method. The technique works by taking each observation with missingness, one at a time, and finding the "best cell," the set of other observations in the dataset that, ideally, match the observation with missingness on all observed values. In other words, the best cell contains identical observations that vary only on the variable for which the observation in need of imputation is missing. This is powerful because all observations in the best cell are expected to have the same data generating process as a the observation with missingness. As a result, we can draw randomly from the observed values of the variable with missingness in the best cell and use those draws to impute the missing value. The result is very high quality imputations that are naturally on the scale of the variable with missing values and that variance in the draws appropriately reflects imputation uncertainty.

The other option for multiple hot deck imputation is the "probabilistic draw" method. This method works by calculating *affinity scores* for each observation that measure how closely the observation matches the case for which missing data needs to be imputed. The algorithm selects potential donors for the missing cell from the entire dataset, but uses the affinity scores as weights on the selection so that observations more closely matching the case with the missingness are more likely to be selected.

In both methods, the entire dataset is imputed five times by default. The command `combine.mids` included in the package creates one dataset of the five by averaging them together, resulting in a single dataset in which the imputed values are the median of the five imputations. All of the statistics

reported below are from this single merged dataset, therefore even though the numbers are discussed as single estimates the imputed proportions were calculated using all of the multiply imputed datasets.

While hot deck methods are not dependent on a regression model for imputing values, the researcher still has to choose which variables the algorithm will use to calculate the affinity scores and identify the closest matching cases. In the case of political variables, such as electoral choices, there are quite a few variables which give reliable clues as to which candidate choice should be imputed. Demographics, including age, education, gender, and race are known correlates of vote choice, as are political variables such as partisanship, ideology, impressions of the candidates, and previous vote choices. These variables will be used to impute the values for the missing candidate choice cases. Missing data on these demographic and political variables used to predict vote choice will be imputed in the same procedure – the multiple hot deck procedure is capable of imputing all missing data at once – but only the cases which were complete prior to imputation are used as potential donors for the missing values. Since the missing demographic and political variables will be imputed at the same time, estimates based on these variables will be assessed as well.