DIW Berlin

German Institute
for Economic Research

Discussion Papers

442

Martin Kroh

Taking ‘don’t knows’ as valid responses: A complete random imputation of missing data

Berlin, September 2004

Opinions expressed in this paper are those of the author and do not necessarily reflect views of the Institute.

DIW Berlin

German Institute
for Economic Research

Königin-Luise-Str. 5
14195 Berlin,
Germany

Phone +49-30-897 89-0
Fax +49-30-897 89-200

www.diw.de

ISSN 1619-4535

Taking ‘don’t knows’ as valid responses:
A complete random imputation of missing data

Martin Kroh

German Institute for Economic Research – DIW Berlin
Socio-Economic Panel Study (SOEP)

Contact:

Martin Kroh
DIW Berlin
Königin-Luise-Straße 5
D–14195 Berlin
Phone: +49 30 89789 678
Fax : +49 30 89789 109
Email: mkroh@diw.de

Abstract

Incomplete data is a common problem of survey research. Recent work on multiple imputation techniques has increased analysts' awareness of the biasing effects of missing data and has also provided a convenient solution. Imputation methods replace non-response with estimates of the unobserved scores. In many instances, however, non-response to a stimulus does not result from measurement problems that inhibit accurate surveying of empirical reality, but from the inapplicability of the survey question. In such cases, existing imputation techniques replace valid non-response with counterfactual estimates of a situation in which the stimulus is applicable to all respondents. This paper suggests an alternative imputation procedure for incomplete data for which no true score exists: multiple complete random imputation, which overcomes the biasing effects of missing data and allows analysts to model respondents' valid 'I don't know' answers.

Keywords: missing data, incomplete data, non-response, multiple imputation, survey methodology, mixture regression models, vote choice

JEL Classification: C81, D72, D80

1. Introduction

Problems of incomplete data have long been ignored by applied empirical research. In recent years, however, the development of advanced imputation techniques and their implementation in software packages has raised scholars' awareness of these problems.[1] One consequence of ignoring missing data can be a bias in parameter estimates, necessitating correction for such effects (e.g., Rubin, 1987; Little and Rubin, 1987). Multiple imputation procedures in particular make it possible to estimate unobserved data effectively, to include all data in the analysis, and thus to alleviate problems due to item non-response.

One of the basic assumptions about the applicability of such imputation techniques is that incomplete data is generated by measurement problems, which means that a true score for a survey question exists for all respondents but that these scores are not always observable. The literature on survey research provides various examples of measurement problems (for an overview see e.g., Tourangeau et al., 2000). These occur particularly if survey questions tap socially desirable behavior (e.g., voting) or illegal behavior (e.g., drug abuse), or if questions relate to subjects considered to be private by respondents (e.g., sexuality). Also, complex survey questions may increase non-response rates. For example, respondents simply may not know how much they contribute to their old age insurance or they may have difficulties recalling how many times they visited a doctor within a certain period, etc. Yet another example of measurement problems is the application of split-half techniques: analysts deliberately generate incomplete data by not observing variables for a certain subgroup or a random subsample of the original *N*. Existing imputation techniques are designed to compensate for measurement problems. These techniques estimate the unobserved (but existing) data on respondents by drawing on patterns in the observed data.

Measurement problems are not the only basis for incomplete data, however. Survey questions often are inapplicable to parts of the sample because respondents are simply not familiar with the stimulus of a question. As King et al. (2001: 50) point out:

> "In some cases [.] 'I don't know' given in response to questions about the national helium reserve or the job performance of the Secretary of Interior probably does not mean the respondent is hiding something, and it should not be treated as a legitimate answer to be modeled rather than a missing value to be imputed."

Asking respondents to give a statement in response to a stimulus presupposes that the stimulus in fact is relevant to all respondents surveyed. For some respondents, however, a true score may not exist. If respondents do not have adequate information with regards to the stimulus, an 'I don't know' answer is a valid response. One may argue that in such cases one does not encounter item non-response in some narrow definition of the term as respondents provide the true answer. Yet, the consequence of a valid 'I don't know' response is the same as an invalid refusal. Irrespective of whether measurement problems or the inapplicability of a survey question lead to an 'I don't know' answer, such a response will generate incomplete data and all the associated problems.

This paper focuses on this kind of incomplete data: valid 'I don't know' answers given in response to survey questions that are not applicable to some part of the sample. A procedure is proposed for modeling the information provided by respondents' valid 'I don't know answer' using multiple complete random imputation. The paper is structured in four sections. The first section very briefly reviews arguments as to why missing data are problematic for statistical inferences drawn from surveys and what statistical tools supposedly compensate for such

problems. The second section develops the idea of a multiple complete random imputation of incomplete data as an alternative to existing imputation procedures. The third section formally and empirically demonstrates the consequences of the technique for a simple linear model. The empirical example (vote choice in the 1994 Dutch parliamentary election) illustrates the consequences of different missing data treatments. The empirical example also focuses attention on the issue of mixed effects in samples consisting of observed and imputed data. The final section discusses the circumstances under which alternative treatments of incomplete data are most appropriate.

2. Missing data and how to handle it

Missing data is a widespread phenomenon in survey research. On the basis of empirical articles published in American political science journals, King et al. (2001) estimate that on average about a third of the original sample used in these papers is excluded from the analyses due to item non-response in any one of the underlying variables. The most common way to handle incomplete data is the listwise deletion of such cases. Deleting in the event of missing data confronts analysts with a trade-off: either they lose numerous cases, or they exclude covariates of interest when those share of item non-response appear too high. Both alternatives are unsatisfactory as they invalidate costly information and, more importantly, limit the statistical inferences that can be drawn from survey data. Yet the listwise deletion of missing data may not only obstruct practical restrictions on the number of analyzable units and variables; it may also generate selection bias in the models estimated. The biasing effect of incomplete data depends on the marginal distribution of non-response.

There are three basic types of incomplete data: data missing completely at random (*MCAR*), data missing at random (*MAR*) and data not missing at random (*NMAR*) (e.g., Rubin, 1987; Little and Rubin, 1987; Schafer, 1997; King et al., 2001; Allison, 2002). In the first case, *MCAR*, the occurrence of missing data is uncorrelated with values on any other variable in the dataset. The listwise deletion of incomplete data will produce an approximately random subsample of the original *N*. Statistical analyses on the basis of the original *N* and a random sub-sample will yield approximately the same, unbiased estimates. Nevertheless, the smaller sample size makes the estimation less efficient. To the extent that incomplete data are correlated with observed values on other variables in the dataset, i.e. if data are *MAR*, a listwise deletion of observations with missing data will lead to a biased subsample.[2] The third form of incomplete data, *NMAR*, describes a situation in which the missing data mechanism depends on the true values that are unobserved.[3] This paper omits the category of data missing completely at random (*MCAR)* as well as the category of data not missing at random (*NMAR*) and focuses on solutions to the problems of data missing at random (*MAR*) only.

The second type of incomplete data, data missing at random (*MAR*), is central for the further discussion in this paper. There is general agreement among scholars that imputation techniques are able to solve the problems related to *MAR* (e.g., Allison, 2002). Such methods estimate the true scores underlying item non-response based on patterns of relationships in the observed data. The literature suggests that of the host of parametric as well as non-parametric methods, multiple multivariate normal specifications and related methods approximate incomplete data most effectively (Schafer, 1997; King et al., 2001). A detailed discussion of these methods and their differences is beyond the scope of this paper (see e.g., Allison, 2002).[4] Nevertheless, it is crucial for the proceeding argumentation to note that these methods all belong to the same group of solutions. Applying these methods means replacing incomplete data by a reasonable guess. In terms of a linear model, some sort of $\hat{y}$ is substituted for missing scores on y.

3. Imputing valid 'I don't knows'?

What are possible solutions to the incomplete data problem if the likelihood of missing data is related to other variables in the analysis, i.e. when data is *MAR* and the cause of incomplete data is not measurement problems but the inapplicability of a stimulus? Of the host of solutions to this problem, this section briefly discusses the adequacy of existing (multiple) imputation techniques when modeling valid 'don't knows'.[5] Thereafter, an alternative approach is presented, a multiple complete random imputation of incomplete data.

Is it sensible to use advanced multiple imputation techniques, which are known to produce unbiased estimates of unobserved data, when the data is missing because survey questions are inapplicable? Are these methods robust against violations of the assumption that unobserved scores underlie the incomplete data? While it makes sense to estimate such true scores in the event of measurement problems, it makes less sense when survey questions are inapplicable. Generally, scores that do not exist should not be estimated if one is interested in a description of empirical reality, because such methods estimate counterfactual data. For instance, if all respondents hold a particular opinion on issue x, what impact would this opinion have on other variables analyzed? As this often is not the kind of information analysts are interested in,[6] existing imputation techniques do not provide an adequate solution to the problem of incomplete data generated by the inapplicability of survey items. The method is indeed not robust to violations of its assumptions.

If one nonetheless regards estimated results on the basis of imputed data as a reflection of the real world, it is a biased reflection. Less-informed respondents are 'transformed' into knowledgeable respondents and are therefore under-represented in the sample. The selection bias of listwise deletion is replaced by the selective misspecification bias of the imputation of incomplete data, the latter usually resulting in exaggerated effects. If, for instance, opinions

on issue *x* affect party choice among informed respondents, this relationship will also be replicated for respondents who are unfamiliar with issue *x*. But this result is, in my opinion, artificial. How can an opinion on *x* have a positive/negative effect on the choice of some party in the group of respondents who are unfamiliar with issue *x* or are unable to form an opinion? But if existing imputation techniques do not provide an adequate tool for handling this particular type of missing data, what else to do? What information is contained in a valid 'I don't know' response that can enable the analyst to relate this answer to the metric of a response scale and thus analyze all the cases in the statistical analysis?

A valid 'I don't know' answer to an inapplicable survey question is, in my opinion, best represented by a random answer on the response scale. If interviewers forced respondents who are unfamiliar with issue *x* to report an opinion anyway, one would usually expect a random answer. Thus, what is in fact unobserved by a valid 'I don't know' answer is an arbitrary answer on the survey question. It is therefore assumed that a substitution of random values for incomplete data is what comes closest to the information provided by 'I don't know' answers that originate from the inapplicability of survey questions.

But how are these random responses most likely distributed? First, it seems plausible to retain the metric of the answer scale. If respondents had been forced to give an answer, they would have done so by using the answer categories offered. Second, the likelihood with which they would have chosen certain answer categories is unknown. Different distributions of the random variable are possible and reasonable.[7] I propose the empirical univariate distribution in the complete cases. The randomly imputed values thereby are a random draw from the observed answers on a response scale. As an alternative to the univariate distribution, one could select the multivariate distribution in the observed cases for the random draw, that is, take into account the response distribution for 'related' cases in the substitution process of missing information.[8] Even if it is practical to use such multivariate distributions for the imputation of incomplete data by arbitrary scores, it is problematic from a conceptual point of view. If incomplete data results from the inapplicability of a stimulus, estimating different variances for the imputed data in different groups of the sample again means generating counterfactual information. The use of multivariate distributions presumes that respondents would have shown a certain response pattern if they had been familiar with the stimulus. Hence, it appears from a conceptual point of view more warranted to rely on less information in the complete cases (univariate distribution) when generating random values for incomplete data than to draw on observable patterns in the complete data (multivariate distribution). I therefore consider the complete random draw from the univariate distribution of the complete data as a reasonable approximation of the unknown distribution of 'forced' answers on a response scale in the group of respondents who provide a valid 'I don't know' answer.

The proposed complete random imputation of missing scores d_{mis} draws values from the univariate observed distribution d_{obs} with estimated mean $\hat{\mu}$ and estimated variance $\hat{\sigma}^2$,

$$d_{mis} \sim f(d_{obs} \mid \hat{\mu}, \hat{\sigma}^2) . \quad (1)$$

If one substitutes values for incomplete data that are the result of a single random draw, computed standard errors are underestimated (Rubin, 1987). Randomly imputed scores are treated as if they are observed ones. One thereby disregards that a random draw retrieves the distributional properties of a population with sampling error only (e.g., Greene, 2000: 97ff). The possibility of falsely relying on an outlier-draw introduces uncertainty in statistical models based on randomly imputed data. To take account of this uncertainty and to estimate more adequate standard errors, Rubin (1987) suggests the creation of multiple imputed datasets. This procedure is often applied in existing imputation techniques that rely on a random component (Schafer, 1997; King et al., 2001).

In each of these m generated multiple datasets, the observed data are identical, as is the algorithm that leads to random draws. However, each random draw generates parameters $\hat{\theta}_l$ (means, regression coefficients, etc.) with l=1,2,...,m (cf. Little and Rubin, 1987: 257). The mean point estimate $\bar{\theta}$ over all m simulated datasets is

$$\bar{\theta} = \frac{1}{m} \sum_{l=1}^{m} \hat{\theta}_l . \quad (2)$$

The total variance associated with $\bar{\theta}$ consists of the mean variance of $\hat{\theta}_l$ over all m imputations, i.e. the within-imputation variance,

$$\bar{v}^w = \frac{1}{m} \sum_{l=1}^{m} \hat{v}_l , \quad (3)$$

and the variance of $\hat{\theta}_l$ between imputations,

$$v^b = \frac{1}{m-1} \sum_{l=1}^{m} (\hat{\theta}_l - \bar{\theta})^2 . \quad (4)$$

To the extent that these results vary over m datasets (between imputation variance), a correction for small numbers of m is introduced so that the total variance is

$$v^t = \bar{v}^w + \frac{m+1}{m} v^b . \quad (5)$$

Although there is, of course, no limit to the number of datasets generated, five replications have proven to be sufficient in many applications (Allison, 2002).

4. Consequences of a multiple complete random imputation

What are the consequences of a complete random imputation of item non-response for statistical analyses that are based on a dataset of observed and imputed data? This section takes a formal and an empirical approach to illustrate this point. The empirical example is based on Dutch politics and demonstrates how three procedures of handling the incomplete data problem – listwise deletion, multiple importance weighted Expectation Maximization (*EM*is) imputation, multiple complete random imputation – affect the analysis of a substantive research question.

4.1 Formal illustration: multiple complete random imputation and regression models

Starting with the formal argument, suppose the simple example of explaining variable y (party preference) by variable x (opinion on issue x) by means of an ordinary least squares regression. One of the main interests will usually be the parameter estimate $\hat{b}$ in this regression,

$$\hat{b} = [x'x]^{-1}[x'y]. \tag{6}$$

Suppose furthermore that respondents often are not sufficiently familiar with the stimulus of x (opinion on issue x), which leads to many 'I don't know' answers. Let x_{obs} denote observed values and x_{mis} missing values on x. Splitting between complete and missing scores details the slope estimate,

$$\hat{b} = \left[\begin{bmatrix} x_{obs} \\ x_{mis} \end{bmatrix}' \begin{bmatrix} x_{obs} \\ x_{mis} \end{bmatrix} \right]^{-1} \begin{bmatrix} x_{obs} \\ x_{mis} \end{bmatrix}' \begin{bmatrix} y_{obs|x_{obs}} \\ y_{obs|x_{mis}} \end{bmatrix}. \tag{7}$$

Item non-response on x, i.e. x_{mis}, is imputed as suggested in Equation 1 by randomly drawing from the observed data on x, i.e. x_{obs}. Hence, a complete random imputation substitutes incomplete data x_{mis} by scores d_{mis} from these random draws, so that

$$\hat{b} = \left[\begin{bmatrix} x_{obs} \\ d_{mis} \end{bmatrix}' \begin{bmatrix} x_{obs} \\ d_{mis} \end{bmatrix} \right]^{-1} \begin{bmatrix} x_{obs} \\ d_{mis} \end{bmatrix}' \begin{bmatrix} y_{obs|x_{obs}} \\ y_{obs|x_{mis}} \end{bmatrix}. \quad (8)$$

Note that d_{mis} is a random variable. The covariance between d_{mis} and y is therefore zero. Note also that the distributional properties of d_{mis} follow the univariate distribution in x_{obs} (since d_{mis} draws randomly from x_{obs}). Hence, the variance of d_{mis} equals the variance of x_{obs}. Equation 8 can thus be reduced to

$$b = \left[x'x \right]^{-1} \left[x'_{obs} y_{obs|x_{obs}} \right] \quad (9)$$

Consecutive equations 6 to 9 point out two important implications of a complete random imputation of incomplete data. First, imputing random values as proposed results in parameter estimates that are approximately zero for respondents who are not sufficiently familiar with the survey question underlying variable x. The parameter estimate $\hat{b}$ of the effect of the issue position on x on party choice is approximately zero for respondents who do not hold an opinion on x. This seems a reasonable implication from what is known about this kind of incomplete data. Respondents' opinions on x cannot have affected party choice if respondents did not hold opinions on x. The second implication regards estimate $\hat{b}$ in the whole sample. Equation 9 illustrates that the estimation of $\hat{b}$ rests solely on the covariance in the complete cases but the variance in all cases. Hence, the size of $\hat{b}$ in the whole sample is reduced according to the proportion of incomplete data. Incomplete data is in fact replaced by random error or white noise. As a consequence, fit statistics are also reduced according to the proportion of missing scores.

4.2 Empirical illustration: vote choice in 1994 Dutch parliamentary election

An empirical example of Dutch politics illustrates the consequences of a multiple complete random imputation as described above. Moreover, it contrasts these findings against a complete case analysis, i.e. results based on listwise deleted missing data, and multiple imputed data using an importance-weighted Expectation Maximization (*EM*is) algorithm (King et al., 2001). The analysis draws on data from the easily accessible (ICPSR or Steinmetz Archive) and well-documented (Anker and Oppenhuis, 1997) Dutch parliamentary election study of 1994 (DPES'94). In the example, a regression model aims to explain the reported probability of voting for the Green/Left Party, *Groen/Links*, in the Dutch parliamentary election of 1994.[9] The dependent variable is a ten-point scale on which respondents indicate how likely it is that they will ever vote for this party.[10]

Two explanatory variables are included in the ordinary least square regression. The first explanatory variable is respondents' self-placement on a nuclear-plants-scale that ranges from 1 (more nuclear plants should be built) to 7 (no more nuclear plants should be built). Table 1 reports that 96% of the respondents who report a value on the dependent variable (probability to vote) also had an opinion on the issue of nuclear power and provided an answer to the survey question.

The second explanatory variable of voting for the Green Party is a feeling thermometer (like-dislike) that measures sympathy for Mohamed Rabbae, one of the two Green/Left party leaders at that time. Mohamed Rabbae was widely unknown among the electorate: 41% of the sample reported not knowing him. The high rate of incomplete data on the question of the Green party leader results in all likelihood from the unfamiliarity of many respondents with this politician and not from measurement problems.[11] Asking the same question of like-dislike about Wim Kok (*PvdA*), Elco Brinkman (*CDA*) and Hans van Mierlo (*D66*) produces only marginal non-response. About 99% of the sample reported an opinion on these three party leaders, suggesting that respondents, who reported not knowing Mohamed Rabbae, provide the correct and legitimate answer. The result may not be surprising, keeping in mind that there are often more than ten parties represented in the Dutch parliament and party leaders of small parties do not receive as much attention by the mass media as leaders of large parties. Mohamed Rabbae was not even the most unknown party leader in the 1994 parliamentary election. A majority of 80% respectively 87% of the respondents is unfamiliar with the party leaders of the two small orthodox Protestant parties *SGP* and *RPF*, Bas van der Vlies and Leen van Dijke. In sum, about 95% of the individuals know at least five out of nine party candidates surveyed in the DPES'94, however, only 8% report an opinion on all nine candidates.

<Table 1>

Given a simple model that explains voting for the Dutch Green Party by a vital environmental party issue, nuclear power, and by the evaluation of the party leader, Mohamed Rabbae, a reasonable research question could be the following: what was more important for supporting the Green Party in 1994, voters' positions on the nuclear plant issue or their evaluation of the candidate Mohamed Rabbae?

Many scholars would probably not approach the question on the basis of the given data with the argument that 43% non-response comprises too much slippage to draw reasonable conclusions. Disregarding these reservations, a linear regression model of the complete cases (Model 1, Table 2) suggests that Mohamed Rabbae was more important for vote choice than the issue of nuclear plants. The estimated effects based on listwise deleted data are of course correct, as long as one is interested in the group of respondents who happen to know Mohamed Rabbae. However, since analysts usually are interested in the whole sample, these findings are at least controversial. It is implausible to conclude a strong effect of the

evaluation of Mohamed Rabbae in the whole sample if this variable causes 41% item non-response.

<Table 2>

If missing data is included in the analysis by means of an advanced imputation technique, one obtains results reported in the following column of Table 2 (Model 2). This solution is based on the importance-weighted Expectation Maximization (*EM*is) algorithm proposed by King et al. (2001). The multiple datasets were generated by using the *Amelia* imputation software developed by Honaker et al. (1999). Without going into detail about the procedure, it may be noted that there is some agreement among experts that this is an advanced solution to the missing data problem (Allison, 2002). Beside variables of the model, additional variables (age, sex, education, interest in politics) are included in the imputation process. Results of the *EM*is imputation illustrate that such imputation methods reduce standard errors of the parameter estimates by including all cases in the analysis. From a substantive point of view, however, one finds rather similar effect parameters than for the complete case analysis. The counterintuitive result is confirmed that Mohamed Rabbae was more important than nuclear plants for voting Green. But what one estimates here is the counterfactual prediction of what voting would have looked like if all respondents had known Rabbae.

The column of Table 2 on Model 3 reports the results based on a multiple complete random imputation of missing data. Again, standard errors are smaller compared to the complete case solution, but effect parameters and the goodness of fit statistic are also reduced. According to the number of missing cases (few for the first and many for the second explanatory variable) estimated parameters decrease in magnitude. These results correspond with what one intuitively would expect: the evaluation of Mohamed Rabbae is less important for voting Green/Left than the issue of nuclear plants. The low R^2 also more accurately reflects what one actually observes on the relationship between opinions on nuclear power and the evaluation of Mohamed Rabbae on the one hand, and the probability of voting for the Green Party on the other. One has, in fact, less information than the variance reduction in the first two models suggests. Given the set of explanatory variables included in the model, the solution of a multiple complete random imputation realistically reflects the fact that the model of vote choice in the example does not apply to almost half of the sample.

The example illustrates that different solutions to the missing data problem affect substantive findings drawn from statistical models. The results clearly are different, yet one cannot judge which result is correct based on expectations and prior knowledge. The question of which approach to choose has to be decided before the imputation of missing data, based on the applicability of the assumptions underlying the different methods. The concluding section of this paper will discuss this point.

4.3 Mixed effects in samples consisting of observed and imputed data

In a multiple complete random imputed dataset, one obtains two (possibly different) coefficients for one relationship: a zero-effect, b_{mis}, in the group of respondents who validly say that they do not know the answers, and an estimate b_{obs} in the group of respondents who use the answer categories provided. The latter corresponds to the estimate one derives from a complete case analysis. Model 4 in Table 2 illustrates this point (see Section 4.1 for a formal argument). If one uses an interaction term to distinguish the candidate effect for respondents with complete information from the effect for respondents who do not know Mohamed Rabbae, one obtains an effect for the reference category (respondents who know Rabbae) of $b = 0.55$ (Model 4), which is very similar to the estimate based on complete cases only, $b = 0.54$ (Model 1). The estimated deviation in the party candidate effect on vote choice for respondents who do not know the politician is $b = -0.55$. Hence, the effect among those who say 'I don't know' is approximately zero.[12]

In a joint analysis of both imputed and complete data (Model 3), the overall estimate $\hat{b}$ is the weighted average of respective estimates $\hat{b}_{mis}$ and $\hat{b}_{obs}$ in both parts of the sample. This describes what is also referred to in the statistical literature as a situation of a heterogeneous sample or mixed distribution of effects, i.e. groups with different coefficients mixed in one sample (e.g., Laird 1978; Arminger et al., 1995; Böhning and Seidel, 2003). In estimating the overall effect, the procedure of a multiple complete random imputation may bear the risk of producing coefficient b in the whole sample which describes none of the two groups accurately if they are homogeneous.[13]

As long as effect heterogeneity is uncorrelated with variables of the model, ordinary least square regressions will provide an unbiased though not efficient estimation of parameters (Greene, 2000: 501). Nevertheless, to control for mixed effects and to investigate such differences, which are of interest from a conceptual point of view, additional model specifications are applied to the empirical vote choice data: continuous and discrete random effect models.[14]

The hypothesis of mixed effects as a consequence of the multiple complete random imputation implies that estimates are more variable in a dataset consisting of complete and imputed data than in a dataset of only complete data. To test this hypothesis one requires variation in estimated coefficients. This variation can be obtained by drawing on hierarchical (multilevel data) or repeated observations (e.g., across time or items). In the DPES'94, the probability to vote question and the evaluation of party leaders are surveyed not only for Green/Left but for nine parties in parliament (see Table 1), which leads to 11,043 observations of the relationship between candidate evaluation and party preference (= 1,227 respondents x

9 parties). The repeated observations for respondents *i* across parties *j* allow to estimate individual variation in the candidate effect $\widehat{b_i}$ on vote choice. If the hypothesis of unobserved heterogeneity holds true, $\widehat{b_i}$ should be clustered in imputed datasets between imputed and observed data or should at least in imputed datasets be more variable between respondents than in complete datasets.

To test for the possibility of higher variability of $\widehat{b_i}$ in imputed datasets, models in Table 3 fit continuous random effect regressions of the probability to vote for a party on the evaluation of respective party candidates. This is done for the data deleted listwise (Model 5), for the *EM*is imputed data (Model 6) and for the complete random imputed data (Model 7).[15] Note that about 29% of the observations contain missing information on candidate evaluations.

<Table 3>

The estimates of Models 5 to 7 indicate that variation in effect parameters is evident for all datasets. The estimated standard deviation of the candidate effect across respondents reported in the second part of the table exerts statistical significance for these models. However, in contrast to the hypothesis of inflated heterogeneity due to imputations, the *EM*is and the complete random imputed data do not show more variation in the candidate effect than the data deleted listwise. Thus, there is no indication of inflated effect heterogeneity attributable to the joint analysis of complete and imputed data in one sample.

Although continuous random effect models do not indicate increased heterogeneity in a sample that consists of complete and imputed data, one may argue that this kind of unobserved heterogeneity is more appropriately modeled by discrete random effects. To the extent that effect parameters differ between observed and imputed data, a discrete random effects model[16] should detect these two groups and ascribe different estimates to them. To test for such latent classes, Table 4 reports discrete random effect models regressing the reported probability to vote for a party on respective candidate evaluation in the complete data (Model 8), the *EM*is imputed data (Model 9) and the complete random imputed data (Model 10).

<Table 4>

The estimates of Models 8 to 10 show correspondences. In all three cases, mixture models identify two latent classes of respondents that significantly differ in the weight they give to party candidates when considering whether to vote for a party.[17] In class 1, which contains in all three models less then 20% of the sample, party candidates have a weaker effect on vote choice than in class 2.[18] The negative deviation of class 1 from the mean effect does not differ substantively between the datasets consisting of complete and imputed data and the data

deleted listwise. The similarity of random effects between the complete case analysis and analyses on imputed samples suggests that the latent classes defined by these mixture models cannot unanimously be attributed to the difference of observed and imputed data.[19] Apparently there are other reasons for these two groups of respondents relying more heavily either on their evaluation of party candidates when assessing their party choice. In brief, continuous random effect models do not show signs of inflated effect heterogeneity and discrete random effect models do not show signs of mixed effects due to the imputation of incomplete data.

The discussion of mixed effects between imputed and observed data in this section has been the discussion of a technical obstacle. Conversely, one may argue that effect heterogeneity is not a consequence of the imputation technique but – provided that the assumptions underlying the method hold – an accurate reflection of empirical reality. Hence, mixed effects are a characteristic of the phenomenon analyzed and is therefore of substantive interest. It is not, however, a defect of the imputation procedure. This leads then to the crucial question under which circumstances the assumptions of the method do or do not hold true.

5. Conclusion: choosing a method

Every potential solution to the missing data problem rests on assumptions about the data. The use of different methods therefore hinges upon the correctness of assumptions, something that has to be judged for each variable in the dataset separately. Deleting all missing data without considering the consequences can produce biased results. But the idea of cleaning datasets by universally applying multiple imputation techniques is also problematic.

If one encounters data missing at random (*MAR*) and is confident that measurement problems are at the root of item non-response, i.e. that a true score for each missing score exists, obtainable multiple imputation techniques are the proper choice. They provide estimates of the unobserved scores. Such methods apply in all cases in which respondents are asked to report facts like demographics or past behavior. Irrespective of whether respondents answer questions on their age or education, they nonetheless have a certain age and education. Also, even if respondents state that they do not know whether they attended religious services, either they did or did not go to church. For a number of variables that do not concern facts, there exists established evidence that measurement problems often occur in form of refusals. Questions on socially desirable behavior, such as vote intentions, can be named in this respect (cf. Tourangeau et al., 2000). It seems plausible to assume that all respondents have an idea what is meant by this question and whether or not they consider casting a ballot. Missing values on this question most likely reflect measurement problems.

The multiple complete random imputation is the preferable choice if one can be confident that 'don't know' answers are valid, i.e. in the absence of a true score to the survey question. When respondents are asked to report their opinion on an external stimulus one has to take into consideration that the stimulus of the question may be unknown to some respondents or that they are not affected by it. Preferably, one would select respondents who are unfamiliar with a certain item before prompting them to report an opinion on the stimulus. However, many surveys do not contain such filter questions.[20] In general, if the likelihood of measurement problems is low but many respondents are probably really not familiar with the stimulus, a multiple complete random imputation is a sensible solution.
The method avoids problems related to data missing at random (*MAR*) when incomplete data results from inapplicable survey questions. By including all cases in the analysis, the estimation of statistical models becomes more efficient. Moreover, variables can be incorporated in the analysis according to their substantive interest to the analyst rather than to their proportion of item non-response. Most importantly, this procedure prevents sample selection of the listwise deletion as well as selective misspecification bias in the form of counterfactual estimates due to falsely applying alternative imputation techniques. Finally, in the empirical example provided in this paper, there is no clear indication that the method of a multiple complete random imputation generates problems of effect heterogeneity between observed and imputed data. The existence of mixed effects is, however, a possibility that should preferably be tested and if necessary be modeled by means of, for instance, random effect models or simply interaction models when applying imputation procedures.

In many situations one cannot know with certainty what caused certain scores to be missing. Research on survey response may indicate how likely measurement problems are for different variables. If one does not have such information to reject or support either the application of existing imputation techniques or the application of a multiple complete random imputation, the latter may be regarded as the more conservative method. As illustrated in the empirical example, statistical models based on multiple complete random imputations tend to decrease effect parameters. Methods of data augmentation, conversely, tend to generate effect parameters more similar to those found in the complete data. Hence, the error one makes when applying falsely existing imputation techniques is that of overestimating a relationship that does not exist. The other case, where a multiple complete random imputation is applied although there is a relationship underlying incomplete data means underestimating relationships in the whole sample. The choice of complete random imputation is therefore the more conservative approach.

[1] *SAS* users, for instance, can run multiple imputations with the *MI PROX* tool or with *IVEWARE*, which is a *SAS*-based application by Raghunathan et al. (2000). *MICE* by van Buuren and Oudshoorn (2000) contains *S-PLUS* software. Honaker et al. (1999) wrote *Amelia*, an imputation software which can be used as *GAUSS* application or as Windows program.

[2] For instance, if politically uninformed respondents show item non-response on questions regarding their political opinions more frequently than politically informed respondents, uninformed respondents will have a higher likelihood of being excluded from analyses of the multivariate patterns of such opinions, leading to over-representation of knowledgeable respondents. If the patterns in opinions differ between informed and uninformed respondents, one will introduce bias in the estimation of these patterns due to the omission of observations with incomplete data.

[3] Not missing at random (*NMAR*) can be encountered, for example, if wealthy respondents are more reluctant to report their income. Heckman models for selection bias are a tool used to handle this specific missing data problem (Heckman, 1979).

[4] The difference derives basically from the application of different estimation techniques in arriving at a prediction of unobserved data.

[5] Researchers often add a category of missing information to the variable affected by non-response to jointly analyze all cases. Collapsing the variable affected by high rates of item non-response into a discrete one with (few) values that represent the topic surveyed and one value for 'I don't know' answers means to lose information of a more detailed (possibly metric) response scale. Alternatively, a modified zero-order regression is often applied to model missing data. This refers to a model in which missing scores on an explanatory variable are replaced with a constant (e.g., zero) and an additional binary variable is included in the analysis that distinguishes between observed and unobserved scores on this variable. As Greene (2000) points out, this procedure is algebraically identical to the simple mean substitution. But this mean substitution is also not a solution to the problem of incomplete data. As, for example, Little (1992: 1231) notes, the simple mean substitution biases the sample (co)variance and "cannot be generally recommended".

[6] Bartels' (1996) analysis of uninformed and fully informed vote choices is an example of studies in which such counterfactual data is explicitly sought.

[7] One would prefer to draw on established evidence from experimental research on response functions for different 'fake' questions. In other words, how do respondents in an experiment form an opinion on policies, politicians, etc. that do not exist if no 'I don't know' answer category is provided? The problem with such an approach is, however, that respondents' answers may depend heavily on cues included in the specific stimulus and experimental knowledge on response functions may thus be difficult to generalize. For instance, responses may depend on the position of the question in the questionnaire, the domain (e.g., international politics, public health, the economy), the concept (e.g., evaluation, feeling, importance, judgment), certain formulations (e.g., using phrases like "problem", "crises", "benefits") and other associated characteristics of the survey item (e.g., time reference or labeling of answer categories).

[8] Suppose, for instance, that on average women tend to have a different opinion on issue *x* than men do, thus the distribution of opinions differs between the two groups. One could use these two distributions for the generation of two random variables and replace item non-response for women and men separately.

[9] For contextual information on the Dutch political system and the 1994 parliamentary election see e.g. Anderweg and Irvin (2002).

[10] For discussion of the validity and applicability of the measure see for example van der Eijk (2002), Kroh and van der Eijk (2003) and van der Brug et al. (2003).

[11] The DNES'94 distinguishes three forms of non-response on this variable: (a) don't know, (b) don't know this politician, and (c) refusal. Of the missing data on the feeling thermometer of Mohammed Rabbae, more than 90% falls into the category 'do not know this politician', about 9% of the respondents with missing information said 'I don't know' without giving a specific reason, and 3 respondent refused to answer the question (Anker and Oppenhuis, 1997: 77). These figures are indicative of the assumption that non-response on the feeling thermometer on Mohammed Rabbae is motivated predominantly by the lack of information about this politician and not by measurement problems.

[12] Table 2 reports for Models 4 a significantly positive effect of not having an opinion on Mohammed Rabbae on the probability of voting for the Green party (b = 2.29). The effect denotes a comparison between respondents who evaluate Rabbae with the score 1 (dislike Rabbae very much) with respondents who report not to know this politician. The former have, as one would expect, a significantly lower probability of voting for the Green Party than the latter.

[13] Although it may be the case that a multiple complete random imputation generates variation in estimates

between observed and imputed data and thus larger confidence intervals for $\hat{b}$, one may expect that these two groups are not discrete ones and therefore not homogenous in their effect. Some respondents may lack relevant information and may thus be so uncertain of their opinion on an issue that they decide not to respond. Others may feel fairly uncertain, though, just sufficiently informed to give an answer, and again others may be fully informed about the topic surveyed and may therefore be very certain of their answer (cf. Alvarez and Franklin, 1994). A multiple complete random imputation defines the effect of the variable as zero for which respondents validly say that they do not know the answer. However, in the group in which respondents are fairly uncertain of their response, this answer in all likelihood has an unsystematic or weak impact on other variables, while in the group of respondents with answers given with certainty, this variable presumably has more relevance for third variables analyzed. In other words, in many instances the distribution of $\hat{b}$ in the overall sample may form a new, continuous distribution and not two discrete ones. The overall mean point estimate $\hat{b}$ may then properly describe the average respondent who is only fairly certain of her answer.

[14] As argued in the previous footnote, one may often expect a continuous distribution of effects rather than a discrete one. The overall sample of complete and imputed data may comprise not only of two distinct and therefore homogenous groups of respondents fully informed and respondents fully uniformed about a topic surveyed, but also of respondents that can be located on a scale of certainty, which positively moderates the effect of the variables analyzed. This is why the existence of effect heterogeneity is also tested by means of a continuous random effects model.

[15] The imputation of missing scores is performed for each party separately. As in the previous example of voting Green, the *EM*is imputation is based on the variables of the model (the probability to vote for a party and the candidate evaluation associated) and additional covariates (age, gender, education and interest in politics).

[16] Discrete or non-parametric random effect models are often also referred to as latent class or (finite) mixture regression models.

[17] Allowing for more than two latent classes does not improve the model fit significantly. In general, altering model specifications in the example often leads to unstable results or even non-convergence of the ML algorithm, which is indicative of identification problems of latent classes in the data.

[18] For instance, in the multiple complete random imputed data (Model 10), the candidate effect on the reported probability of voting for a party in class 1 is the estimated deviation from the main effect, thus 0.55 – 0.15 = 0.40, whereas the candidate effect in class 2 is 0.55 + 0.04 = 0.59. Note that Table 4 does not report standard errors for the location of the second class, because these parameters, z_2, deterministically derive from the estimated location of class 1, z_1, and the estimated group size, π_1, as $z_2 = z_1 \pi_1 / (1 - \pi_1)$.

[19] Continuous and discrete random effect regressions test for unobserved heterogeneity. However heterogeneity due to imputation can be easily identified. Fitting a covariate effect of a variable indicating which values are observed and which ones are imputed on the estimated random candidate effect (not reported in form of a table), leads to similar results as reported in Tables 3 and 4. The covariate effect of not knowing certain candidates is stronger in the case of complete random imputed data (see also Model 4 in Table 2) than in the case of *EM*is imputed data, yet it does not notably contribute to the explanation of continuous or discrete random effects.

[20] This leads to a situation in which some respondents, who are unfamiliar with the content of a survey question, voluntarily admit that they do not have an opinion. Others, who are unfamiliar with a stimulus will provide an substantive answer because they feel that this is expected in the interview situation.

6. References

Allison, P.D. (2002). *Missing Data.* Sage Series: Quantitative applications in the social sciences. Thousand Oaks: Sage.

Alvarez, R. & Franklin, C. (1994). Uncertainty and Political Perceptions. *Journal of Politics* 56: 671-688.

Anderweg, R.B. & Irwin, G.A. (2002). *Governance and Politics of the Netherlands.* New York: Palgrave.

Anker, H. & Oppenhuis, E. (1997). *Dutch Parliamentary Election Study 1994*. Ann Arbor: ICPSR (Study Nr. 6740).

Arminger, G., Clogg, C.C. & Sobel, M.E. (eds) (1995). *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. New York: Plenum.

Bartels, L. (1996). Uninformed Votes: Information Effects in Presidential Elections. *American Journal of Political Science* 40: 194-230.

Böhning, D. & Seidel, W. (eds.) (2003). Recent developments in mixture models. *Computational Statistics & Data Analysis* 41 (Special Issue): 349-678.

Brug, W., van der, Eijk, C., van der & Franklin, M. (2003). *Designs for the empirical analysis of electoral preferences, utilities and choice*. Paper prepared for the joint sessions of workshops of the ECPR in Edinburgh, March 2003.

Buuren, S., van & Oudshoorn C.G. (2000). *Multivariate imputation by chained equations: MICE V1.0 User's Manual*. Leiden: TNO Preventie en Gezondheid.

Eijk, C. van der (2002). Design issues in electoral research: taking care of (core) business. *Electoral Studies* 21: 189-206.

Greene, W. (2000). *Econometric Analysis.* 4th Edition. London: Prentice Hall.

Heckman, J. (1979). Sample Selection Bias as a Specification Error. *Econometrica* 47: 153-161.

Honaker, J., Joseph, A., King, G., Scheve, K. & Singh, N. (1999). *Amelia: A Program for Missing Data.* Cambridge: Harvard University.

King, G., Honacker, J., Joseph, A. & Scheve, K. (2001). Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *American Political Science Review* 95: 49-69.

Kroh, M. & Eijk, C., van der. (2003). *Utilities, Preferences and Choice.* Paper presented at the joint sessions of workshops of the ECPR in Edinburgh.

Laird, N. (1978). Nonparamteric maximum likelihood estimation of a mixture distribution. *Journal of the American Statistical Association* 73: 805-811.

Little, R. (1992). Regression With Missing X's: A Review. *Journal of the American Statistical Association* 87: 1227-1237.

Little, R.J. & Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley.

Raghunathan, T.E., Solenberger, P. & Hoewyk, J., van. (2000). *IVEware: Imputation and Variance Estimation Software: Installation Instructions and User Guide.* Survey Research Center, Institute of Social Research, University of Michigan.

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

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Tourangeau, R., Rips, L.J. & Rasinski, K. (2000). *The Psychology of Survey Response.* Cambridge: Cambridge University Press.

Table 1 Descriptive statistics.

	Range	Mean	Variance	Missing
Probability to vote *Groen/Links* (GL, Green-Left)	1 – 10	4.13	7.93	-
Nuclear plants issue	1 – 7	5.14	3.60	0.04
Party candidates				
Wim Kok (*PvdA*, Labor)	1 – 10	7.37	4.37	0.01
Elco Brinkman (*CDA*, Christian-Democrats)	1 – 10	5.53	5.59	0.01
Hans van Mierlo (*D66*, Left-Liberals)	1 – 10	7.11	3.81	0.01
Hans Janmaat (*CD*, Extreme Right)	1 – 10	1.34	1.03	0.02
Frits Bolkenstein (*VVD*, Right-Liberals)	1 – 10	5.65	5.08	0.05
Mohamed Rabbae (*GL*, Environmentalists)	1 – 10	5.51	5.79	0.41
Gerrit Schutte (*GVP*, Orthodox Protestants)	1 – 10	5.26	5.96	0.46
Bas van der Vlies (*SGP*, Orthodox Protestants)	1 – 10	4.76	5.25	0.80
Leen van Dijke (*RPF*, Orthodox Protestants)	1 – 10	4.57	5.98	0.87
Number of unknown party candidates	0 – 9	2.64	1.84	-

Data Source. DPES'94. *N*=1,227

Table 2 Linear regression models of the reported probability of vote for the Dutch Green Party.

	Listwise Deletion		Multiple *EM*is imputation		Multiple complete random imputation			
	Model 1		*Model 2*		*Model 3*		*Model 4*	
Intercept	- 0.48	(0.32)	- 0.23	(0.28)	0.42	(0.29)	- 0.33	(0.28)
Nuclear plants issue	0.39***	(0.05)	0.33***	(0.04)	0.38***	(0.04)	0.34***	(0.04)
Evaluation of Rabbae	0.54***	(0.04)	0.49***	(0.04)	0.32***	(0.04)	0.55***	(0.04)
Rabbae *x* don't know Rabbae	-		-		-		- 0.55***	(0.09)
Don't know Rabbae	-		-		-		2.29***	(0.49)
N	704		1,227		1,227		1,227	
Adjusted R^2	0.29		0.25		0.15		0.22	

Note. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$; standard errors in parentheses. *Data Source*. DPES'94.

Table 3 Continuous random effect models of the reported probability to vote for a party.

	Listwise deletion *Model 5*	Multiple *EM*is imputation *Model 6*	Multiple complete random imputation *Model 7*
Fixed effects			
Intercept	0.76*** (0.06)	0.55*** (0.06)	1.00*** (0.06)
Evaluation of party candidates	0.69*** (0.01)	0.65*** (0.01)	0.55*** (0.01)
Random candidate effect, second level			
Standard deviation of effect	0.10*** (0.01)	0.08*** (0.01)	0.08*** (0.01)
$N_{Respondents}$, second level	1,224	1,227	1,227
$N_{Observations}$, first level	7,796	11,043	11,043

Note. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$; standard errors in parentheses. *Data Source*. DPES'94.

Table 4 Discrete random effect models of the reported probability to vote for a party.

		Listwise deletion *Model 8*	Multiple *EM*is imputation *Model 9*	Multiple complete random imputation *Model 10*
Fixed effects				
Intercept		0.77*** (0.06)	0.56*** (0.06)	1.01*** (0.06)
Evaluation of party candidate		0.69*** (0.01)	0.65*** (0.01)	0.55*** (0.01)
Random candidate effect, second level				
Deviation of effect in	class 1	- 0.19*** (0.03)	- 0.17*** (0.04)	- 0.15*** (0.05)
	class 2	0.04	0.03	0.04
Size of	class 1	0.19	0.15	0.19
	class 2	0.81	0.85	0.81
$N_{Respondents}$, second level		1,224	1,227	1,227
$N_{Observations}$, first level		7,796	11,043	11,043

Note. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$; standard errors in parentheses. *Data Source*. DPES'94.