Masteroppgave

NTNU
Norges teknisk-naturvitenskapelige universitet
Fakultet for samfunnsvitenskap og teknologiledelse
Institutt for Sosiologi og statsvitenskap

Stian Fagerli Arntsen

«The Application of Multiple Imputation in Correcting for Unit Nonresponse Bias»

Masteroppgave i sosiologi og statsvitenskap

Trondheim, høsten 2010

NTNU
Det skapende universitet

Acknowledgements

The data used in this thesis was provided by Statistics Norway. Any errors in this thesis is solely my responsibility

First of all, I want to thank my supervisor, Kristen Ringdal, for helping me find relevant literature and giving me useful comments.

I also want to thank Øyvin Kleven and Chen Xiaoming at SSB for helpful discussions on the theme and research question of my thesis.

I want to thank Joakim Døving Dalen for helping me preparing the data and for helpful discussions.

Of course, I want to thank my fellow students at Brakka for giving me a chance to unwind during lunch breaks and weekends. A special thanks to Fleskeorgelkompaniet is in order. This thesis would not have seen the light of day without you.

Trondheim 2010

Stian Fagerli Arntsen

Contents

Tables

Figures

Appendix

1. Introduction

Social surveys form the basis of a large body of research in the social sciences. For these surveys to be representative of the populations they are drawn from, they have to be a random sample of the total population. But in practice, it is very hard to get every person in the sample to participate in the survey. Failure to obtain a high response rate can lead to nonresponse bias in the data. In the recent years, reviews of the literature on social surveys have uncovered several disturbing trends that have startling consequences for the validity of the research results. An extensive review of the social research literature suggests that response rates of 50 percent in considered adequate for analysis and reporting (Babbie 2007:262). A response rate of 60 percent is considered good, while 70 percent is considered very good. King et al. (2001) found that in survey-based political science articles, only about half of the respondents answer every question. 94% of the articles in their review of the literature use listwise deletion[1], leading to an average loss of one third of the data. In an extensive study of survey nonresponse based on data from 16 countries and 10 different surveys, de Leeuw and de Heer (2002) concluded that the response rates have in general decreased over the years. Needless to say, with response rates as low as 50 percent being considered adequate for publishing research, listwise deletion being a common form of handling missing data, and response rates decreasing in general, there is a good chance that nonresponse might cause bias a disturbingly large amount of studies. The recent trend in survey response rates illustrates how important it is for researchers to be properly schooled in handling nonresponse. Sadly, it seems that far too many researchers simply ignore this problem, either because of ignorance or because they lack the time and skill to take the necessary measures.

There are two different forms of missing data; missing units and missing items. A missing item is a missing value on a variable. This is the case when a person in a survey does not answer one of the questions. A missing unit on the other hand is when an entire case is missing. In the social sciences, missing units are usually persons who for some reason do not participate in the survey. The most common form of correcting for *unit nonresponse* is weighting. This method of reducing bias has received much attention from researchers and is well understood. This method is very useful for finding point estimates such as means. But weighting cannot address the problem of *item nonresponse*. Most statistical techniques

[1] Most statistical techniques require complete data for all cases. Listwise deletion excludes any unit from the analysis that has one or more missing value(s) on the variables in the analysis.

assume a complete data matrix. But as we have seen, most surveys have substantial amounts of missing items. Statistical techniques that use listwise deletion ignore all units that do not have complete data on all variables. This reduces the statistical power of the analysis and increases the probability of making a type II error[2]. The problem of item nonresponse can be fixed by imputing plausible values for the missing items. Imputing missing values allows the researcher to utilize all the information in the dataset, thereby maximizing the statistical power. One such group of imputation methods is called multiple imputation (MI). In a multiple imputation, the process of filling in plausible values is repeated *m* times, creating *m* imputed datasets. The averages of these datasets are used to estimate the test statistic of interest. This gives a better estimation of the true variance in the data, thus eliminating the need for special methods of estimating variance. Multiple imputation allows the researcher to use standard methods of estimation, enabling him or her to capitalize on their previous knowledge and experience.

The most common application of multiple imputation is in the case of item nonresponse. In fact, according to Rässler (2003:3), "*unit-nonresponse evaluations for MI are quite rare if not a complete novelty*". This thesis will hopefully be able to expand on this neglected field of research. Multiple imputation can only be used to estimate missing values, and is therefore only applicable when at least some information about each respondent is known. The European Social Survey 2006 for Norway contains background information from the population register on almost all recipients of the survey. This has led me to the following research question:

Can multiple imputation be used to correct for unit-nonresponse bias in survey data that contains only a limited amount of information about the nonrespondents?

To answer this question, I first need to review some of the literature on nonresponse and the different approaches to correcting for it. Nonresponse is but one of the different sources of survey errors. To put this problem into context, I will briefly look at the other sources of total survey error. I will proceed by discussing what nonresponse bias is, how it can be detected, and present a discussion about the different methods for handling unit nonresponse bias in surveys. This will form the theory chapter of my thesis. The data source and method chapter

[2] A type II error is a failure to reject a false hyothesis

will deal with the practical implementation of a multiple imputation. The first part of my analysis chapter will deal with determining nonresponse bias, and in the second part, I will discuss the assumptions of multiple imputation, and present the analysis- and imputation models. A discussion on the practical implementation of multiple imputation using two different statistical software packages will round off the methods chapter. I will present the results of my analysis in the fourth chapter. Finally, in chapter five, I will discuss my findings.

2. Theory

Social surveys

Social surveys are a useful tool for social scientists. Surveys can provide invaluable information about numerous facets of the respondent's lives. Such studies are used in many different fields of research, from political attitudes and electoral behavior to happiness and public health. The simple logic behind the survey method is that when a sufficiently large sample of the population participates in a survey, the opinions and demographic makeup of the sample will be representative of the entire population. Most statistical tests assume that you have a complete dataset, consisting of independent units that have been sampled through a random process. Because of errors in measurement and sampling, these assumptions are rarely met in practice. The problem is especially important with regards to cross national surveys, such as the European Social Survey. Such surveys have a goal of creating comparable statistics across nations. To ensure the best possible quality, it is therefore important to minimize the *total survey error* (Groves et al. 2004). There are numerous opportunities for errors to occur. These can be divided into two main groups; measurement errors and errors in representation. The focus of this thesis will be on one of the representation errors; nonresponse bias. But first, we need to put the problem of sources of error in context. I will start by having a look at the first group; measurement errors.

Measurement errors

In the social sciences, we are often trying to measure an abstract *construct* such as political trust or personal wellbeing. But there is often a difference between the true value of the construct we are trying to measure and the value of the measurement. The goal is to minimize the difference between the two so that we have a high *construct validity* (Groves et al. 2004:50). Let's assume that *political knowledge* is the construct. We could try to measure this by a set of questions about politicians and political events. In this case we are trying to find the true value for each respondent; μ_i. But in reality, we can only obtain the value of the

measurement; Y_i. In addition we have an error term, ε_i , which is the deviation from the true value, so that;

$$Y_i = \mu_i + \varepsilon_i$$

A person might have a *true value* Y_i of 5 on a scale of 1 to 10 (though in reality, such a *"true"* value is hard to imagine). If our measure, μ_i is 4, our test has an error, ε_i of $Y_i - \mu_i = 1$. The error did not occur because the respondent didn't answer to the best of his or her abilities, but because our measurement does not give an accurate value of the construct. In other words, the measure we are using does not accurately represent the construct. *Validity* can be defined as the correlation of the measurement Y_i , and the true value μ_i , measured over all possible trials and persons (Groves et al. 2004:51).

A very similar notion is the gap between the ideal measurement and the response obtained, sometimes referred to as *measurement error.* This can be quite similar to the above, as the only difference is that in the latter case, the respondent might hold back on a sensitive question, or in the case of political knowledge, simply guess the correct answer even though he or she has little political knowledge. We call the reported value y_i and the true ideal measurement Y_i. If there is a systematic difference between the values $(y_i - Y_i)$ the result is *response bias.* This problem typically occurs on sensitive topics, such as drug abuse, which tends to be systematically underestimated in surveys.

Response bias can also stem from the same person not giving the same answer over several measurements, which Groves et al. (2004:53) refers to as *variability in response deviations.* More commonly, this is known as the *reliability* of the measurement. Some questions, such as *how happy are you,* may vary from day to day. Survey statisticians refer to this as *response variance,* to clearly separate it from *response bias. Response bias* is systematic, and can occur if the responses deviate because of the respondents change their mind in light of recent events. *Response variance* only leads to instability in the value of estimates over time, as respondents seldom give the exact same score on, say political trust, in for example an original survey and a follow-up survey. (Groves et al. 2004:53).

Processing of the data can also lead to errors. Extreme outliers might be removed, even if they in some cases are true values. Qualitative answers, such as when a respondent is allowed to answer with his or her own words, might subsequently be misinterpreted in the coding process (Groves et al. 2004:53-54). Measurement errors can to a large degree be avoided if the survey is well thought out and the questions are precisely worded. Measurement errors affect all

respondents equally and are seldom a source of serious bias. Obtaining precise measurements is important in surveys. But getting a correct measurement will only take you so far if the data is not representative of the population. Errors in representation can often lead to more serious bias.

Errors in representation

There are several types of representation errors that can occur. First, let's look at *coverage error*. Coverage error is the failure to accurately represent the target population. Even if simple random sampling (SRS) is used, coverage error might still occur if the sample has been drawn from an incomplete list of the population. This can happen when the *sampling frame* does not include everyone in the *target population*. In some countries it is difficult to find a complete register of people and their resident address. If names are drawn from a telephone register, households without telephones are naturally excluded. Such a situation is called *undercoverage.* The telephone-list approach might also lead to drawing *ineligible units,* such as businesses (also called *overcoverage*). *Coverage errors* occur before the sample is even drawn, thereby making it difficult to adjust for in the later stages (Groves et al. 2004:54-55).

On a related note, we have *sampling error*, which is the gap between the *sampling frame* and the sample. The sampling frame might be very large, and not every person in the sampling frame can be interviewed, as this would be immensely costly and represent a huge logistical challenge. Instead, subsamples of the sampling frame are selected. Ideally, every person should be just as likely to be selected at this stage. This error is deliberately introduced into the sample in survey statistics (Groves et al 2004: 57-58).

Nonresponse

Nonresponse is the source of total sampling error that will be the focus of this thesis. As previously mentioned, both item- and unit-nonresponse are prevalent problems in survey data. Nonresponse error occurs when a statistic, such as the mean of the respondents, is different to the mean of the gross sample. If respondents (units) or values (items) of individual questions are missing completely at random, this does not pose a major problem, since we essentially would be dealing with a slightly reduced sample size. But in some situations there might be systematic reasons for missing values or respondents. There might be a somewhat uniform group of the population is hard to reach by the chosen mode of data collection. For example, younger people might be difficult to contact at home. Or there might be some common characteristics shared by people who refuse to participate, for example low levels of

education. In such cases, we might end up with some groups being underrepresented, which can lead to a lack of precision and biased estimates (Groves et al. 2004:58-59). More thoroughly, Billet, Philippens, Fitzgerald and Stoop (2007:137), defines bias in a univariate distribution as a function of the nonresponse rate and the difference between the expected estimated mean from a survey and the true mean of a variable in the population:

$$\text{Bias}(\bar{y}_r) = \left(\frac{M}{N}\right)(\bar{Y}_r - \bar{Y}_m)$$

Bias($\bar{y}_r$) = nonresponse bias of the unadjusted respondent mean;
$\bar{y}_r$ = unadjusted mean of the respondents in a sample of the target population
$\bar{Y}_r$ = mean of the respondents in the target population
$\bar{Y}_m$ = mean of the nonrespondents in the target population
M = number of nonrespondents in the target population
N = total number in the target population

As seen, there are two factors that influence nonresponse bias($\bar{y}_r$); the difference between respondents and nonrespondents ($\bar{Y}_r - \bar{Y}_m$), and the rate of nonresponse $\left(\frac{M}{N}\right)$. To produce serious bias, either the difference between respondents and nonrespondents must be substantial, or the rate of nonrespondents to respondents must be high; or both. This means that even in cases of high rates of nonresponse, there does not necessarily have to be serious bias, as long as respondents and nonrespondents have similar means. But it also means that even moderate rates of nonresponse can introduce serious bias if respondents and nonrespondents are very different from each other.

In terms of handling nonresponse bias, we need to distinguish three different types of nonresponse. Rubin (1987) distinguishes between three different types of nonresponse; missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR). The most serious form of missing is NMAR (also called non-ignorable nonresponse). This occurs when the probability of being a nonrespondent depends on an unobserved variable. This could happen if nonresponse in the European Social Survey was not to depend on age, neighborhood or gender, but on a variable not included in the data, say hair color. Since hair color does not covary (at least to the best of my knowledge) with any observed variable in the survey, we cannot account for the nonresponse mechanism in our model. The difference between NMAR and MAR is actually dependent on the availability of information about the nonrespondents. If the survey sample is drawn from a population register, the researcher will have access to information such as gender, age and household composition for the entire sample, including the nonrespondents. When such information is

available, it can be used to correct for bias and change the situation from NMAR to MAR. Missing at Random is a somewhat misleading term, as it only assumes that nonresponse is dependent on observed, in contrast to unobserved variables. In other words: the data can be systematically missing and still be MAR, *if* we can model the nonresponse by using observed variables. Once we know which variables nonresponse are dependent on, we can explain the nonresponse mechanism and thus ignore it once the model has taken it into consideration. A more fitting term, *ignorable nonresponse*, is also used at times (Lohr 1999:264-265). If however, the missing data is not dependent on any other variable, observed or not, the data is said to be Missing Completely at Random. If the data was missing because of some random bug in the machine used for scanning the paper sheets from the interview, we would not find any systematic reason for missing data. This means that our sample would in essence be a random subsample of the original sample. Any such random subsample should be approximately unbiased under this circumstance. Of all the sources of total survey error, nonresponse is perhaps the most influential source of bias. Luckily, there are methods for correcting for nonresponse bias after the data has been collected.

Ideally, the best way to get a representative sample is to pay careful attention to the planning stages, and to obtain as high a response rate as possible. This is however both time-consuming and expensive. But when all else fails, there are several methods of reducing bias. First, we need a way to determine if the data is missing completely at random (and thus relatively harmless), or if the missing data causes bias. Next, I will look at several ways of determining if the data is biased.

Determining nonresponse bias

So how do we determine if nonresponse causes bias in our data? Groves (2006) propose five techniques for determining if the data is biased by nonresponse:

Response rate comparison across subgroups

One way to assess nonresponse bias is to look at the response rates in different groups of the sample. If we find that response rates are lower in young men with low levels of education, we might assume that this group is biased towards not responding. If the response rates are similar across all groups however, we might assume that there is no evidence for nonresponse bias. One underlying assumption here is that the subgrouping variables are the only variables that are affected by nonresponse, and that any missing data on other variables are missing completely at random. This is not a very solid assumption in most cases, since biases in any other variable (other than the subgroups) will remain undetected. As such, this is at best a

superficial way to determine if any subgroup is under- or overrepresented, and not in any way a guarantee to detect nonresponse bias (Groves 2006: 654). In other words, this technique can only detect the *presence* of bias, and not confirm that bias is *absent.* Still, this is a very simple procedure that can easily let the researcher get an impression of the nonresponse situation.

Comparing data with previous or more accurate sources

If we have access to reliable data on population statistics, say population register data, we can compare the distribution of these well documented variables, to the distributions in our sample. Groves (2006:657) refers to these kinds of statistics as the *'Golden Standard'*. If our sample has an age distribution that is not similar to that of the age distribution in the population, or *'Golden standard'*, we might reasonably assume that the variable is subject to nonresponse bias. But this method also relies on the often flawed assumption that any missing data in the other variables are missing completely at random. For example, there might be a bias toward a characteristic that is both prevalent in, and independent of, any age grouping, and as such it will not leave any bumps or holes in our age distribution. This again will only serve to inform us of any bias in variables where there is a *'Golden standard'* for comparison. But the absent of such bias is not evidence that the data is unbiased on any other variable (Groves 2006:655).

Variation within the survey

In some cases, the data is collected in several phases. Most studies also collect data over time, with some respondents being more cooperative than others. Some answer the phone or accept the interview immediately, while others are more reluctant and take some convincing before joining the survey. In these cases, some researchers assume that the late, or reluctant, respondents are more similar to the final refusals than to the cooperative respondents. Groves (2006:658) refers to this idea as a *"continuum of resistance"*. The information about the reluctant respondents can thereby be used to estimate approximate values for the refusals. The European Social Survey, among other studies, uses the resource intensive approach of converting refusals. Some research also supports the continuum of resistance hypothesis. Lynn et al. (2002:142) has found that easy-to-convert refusals are quite similar to the easy-to-get, but that they are rather different to the remaining refusals. Lin and Schaffer (1995:252) also found that respondents and nonrespondents appear to be somewhat different from each other. But not all research supports this hypothesis.

There is an alternative to the *"continuum of resistance"* hypothesis called the *"classes of nonparticipants model"*. This model assumes that there are several different classes of

nonparticipants. Some of these classes, for example the reluctant respondents, and those who are difficult to contact, are thought to be similar to final nonrespondents. Respondents who refuse because of lack of time, illness or the like, are not likely to be similar to the above, and constitute a second group. Likewise, respondents who refuse because of characteristics of the survey, such as objecting to the subject topic, are also assumed to be different from the other groups (Billet, Philippens, Stoop and Fitzgerald 2007:148). Stoop (2005:152) found that there seemed to be distinct groups that do not participate for varying reasons. It seems that there might be support for each hypothesis in different surveys. But some researchers find little evidence for either hypothesis. Teitler et al. (2006:136) concluded that cases requiring a high level of effort provided poor proxies for the final nonrespondents and that they failed to reduce nonresponse bias. Curtin, Presser and Singer (2000) also find no evidence to suggest that the difference between reluctant and cooperative respondents is large enough to substantially change the estimates (However, their study was based on consumer attitudes, not on the topics typically studied in social sciences. Therefore, I am not confident that their results can be extrapolated to the field of social surveys). Billet, Philippens, Stoop and Fitzgerald (2007) found that the different types of respondents in the European Social Survey round 3 (reluctant, cooperative, easy to convert, hard to convert) differed in attitudes and background variables across countries, suggesting that there are few common traits for different types of respondents[3]. In summary, several researchers find that converting refusals by time consuming and expensive strategies does not pay off, since their inclusion probably doesn't have a great effect on reducing bias in the data either way. But the different results seem to indicate that the situation is unique in each survey. Regardless of which hypothesis fits best in a given survey, both approaches can be used to test if there are differences between levels of cooperation.

Enriching the sample with information from other sources

In some cases, data from other sources than the main survey can be used to augment the dataset and give valuable insight into *who* the nonrespondents are. There are several examples of this. Individual background data (for example population register data) can be used to find out how old the nonrespondent is, his or her gender, how many people live in the same household etc. This is only possible if the identification of the respondent has not yet been made anonyms. The ethics of this approach might be viewed as questionable in some cases though (Groves 2006:657). In the ESS there is data on interview length, number of calls

[3] For more on this discussion, see Curtin et al. (2000) and Stoop (2005)

before first contact, cooperation rates, response rates and reason for refusal. Such information is called paradata. This information can be used to determine who the more reluctant respondents are, and whether this is due to permanent (thinks surveys are a waste of time, do not trust surveys) or circumstantial reasons (do not have the time, is traveling abroad). Interviewers can record the approximate age and sex of the respondent, as well as neighborhood characteristics (such as condition of neighborhood, littering and graffiti) and residence characteristics (apartment or house, condition, size) (Groves 2006:656). Direct information on nonrespondents can be obtained through having all initial respondents answer a short list of questions. This could be done by the *door step approach* or a *follow-up survey.* In the ESS 2006 data for Norway, a follow-up survey was undertaken. This survey contained about a dozen questions from the original survey as well as a couple of questions on how the respondent feels about surveys in general. Examples of the door step approach are the Basic Questions Approach (Kersten and Bethlehem 1987) and PEDAKSI (Lynn 2003). The PEDAKSI (Pre-Emptive Doorstep Administrator of Key Survey Items) approach is a set of basic questions that were found to be most sensitive to nonresponse bias. These questions are given to nonrespondents as well. This approach ensures that at least some information on the non respondents can be obtained. The Basic Questions Approach is a short set of questions given to both respondents and nonrespondents in the hope that some auxiliary information can be extracted that can help adjust any nonresponse bias.

Comparing the original data to a dataset that has been adjusted for nonresponse

The final approach is to examine the difference between the original data and an adjusted dataset. There are several ways of adjusting a dataset, which will be discussed in more detail in the next chapter. In short, the researcher performs one or more adjustments on the data and compares these to the original, unadjusted dataset. When the variable means remain stable across different adjustments, the data is probably not biased. On the other hand, if the means are substantially different between the adjusted and the unadjusted data, bias is likely to be present in the data (Groves 2006:658).

Methods of correcting for bias

If the researcher finds evidence for nonresponse bias, there are several ways of handling the problem. Voogt (2004:133) distinguishes between four classes of methods: weighting, imputation, extrapolation and modeling. Of these methods, weighting and imputation have received the most attention in the research literature. Extrapolation and modeling on the other hand, are difficult to find any decent discussion on, and will not receive any attention in this

thesis. I will give a description of weighting and imputation and discuss briefly their different areas of application.

Weighting techniques

Weighting techniques can be divided into two groups; design weights (sometimes also referred to as inverse probability weights) and post-stratification weights (Bethlehem 2002, Gelman and Carlin 2002:290). Design weights are used to correct for differences in probability of selection. Many surveys use complex sampling methods to ensure a representative sample. However, this might lead to some groups of respondents having a higher probability of selection. The inverse of this probability can be used to correct for these differences. These weights are typically calculated prior to collecting data for the survey. Post stratification on the other hand is performed after the survey has been completed, and is typically used to correct for errors in representation, such as coverage error and nonresponse. This is the most common method for treating unit nonresponse bias.

Post stratification relies on auxiliary information about the target population. The logic of post stratification is very straight forward. First, you need an auxiliary variable, *X*, that divides the population into *L* strata. In the simplest terms, *X* could be gender, divided into two strata. If we know, for example, the number of men and women in a population, we can make sure that the ratio between the two genders is the same in both the population and the sample. We can assign a given weight to each group, so that our sample stratum is approximately equal to our known population stratum. Again, in the simplest term, if we know from other sources that the distribution of males and females is 50/50 in the population, and our sample has a distribution of 60% women and 40% men, we can assign a weight to each respondent to make our sample stratum match that of the population (Bethlehem 2002:277). The weight can be calculated by dividing the observed percentage with the expected percentage of each stratum. In this case we can assign a weight of (60\50=1.2) to the men and (40\50=0.8) to the women. The result is that the two groups are now of approximately equal weight. Even better, if we know the distribution of age as well, and also how age is distributed among the genders, we can perform an even more precise post stratification weighting. But as more variables are used for weighting, the risk of having strata with too few cases also increases. Sometimes this lack of information about the population is a problem when using several auxiliary variables. For example, we might know the distribution of both gender and age in the population, but lack the necessary cross classification information, such as the number males in the oldest age group (as opposed to the other age groups). There are other ways of weighting the dataset,

which Bethlehem (2002:278) refers to as Linear and Multiplicative weights. These weighting techniques can overcome the problem of empty strata by using both marginal frequencies distributions simultaneously. However, these methods will not receive further discussion here. For more on these weighting techniques, see Bethlehem 2002 and Gelman and Carlin 2002.

All forms of weights do however suffer from some common disadvantages. Since weights only utilize complete cases, a lot of information can be lost due to listwise deletion. This in turn will lead to a loss of efficiency, or statistical power. In addition, weighting is an ad hoc approach. This makes it difficult to replicating results accurately. Rässler (2003) argues that weighting can be considered a sort of single imputation, which therefore tends to gives biased estimates of the sampling variance. Nevertheless, weighting is a valuable approach to reducing unit nonresponse bias that is used frequently and has well understood statistical consequences. However, weighting cannot be used to correct for item nonresponse. Instead, imputation methods are traditionally used in these cases.

Imputation methods

While weights are used to correct for *unit* nonresponse, imputation methods can be used primarily to correct for *item* nonresponse. The basic idea behind imputation is to insert a plausible value for each missing items. That way, none of the information in the dataset is lost due to listwise deletion. Let's say that some respondents in the European Social Survey dataset has left a couple of questions (items) unanswered. In a univariate statistic, the missing values might not be prevalent enough to pose a problem if they are missing completely at random. But suppose we want to look at a multivariate statistic that relies on some sort of covariance structure. In this case, a respondent might have a missing value on only one variable, and another respondent might be missing a value on the other. Since both items need to be observed to calculate the covariance, both respondents will be excluded from the analysis, and the total valid N will be reduced. In a univariate statistic, two variables might both have 5 % missing, but if different respondents have missing values on different variables, the total number of missing when looking at a bivariate statistic might be doubled, to 10%. If you are running a multiple regression analysis with a dozen independent variables, the loss of data can become quite substantial. In simpler terms, covariance structures rely on every respondent having answered both questions (Schafer 1997:2). In such a case we could simply impute the mean value of the sample on missing items. If the missing item is highly correlated with another variable, for example if it is part of a series of sub questions on political trust, and thus part of an underlying scale, we can impute the *respondents* mean value

on the relevant items instead. But this general approach to single imputation usually ends up imputing the same value, say sample mean, to every respondent with a missing item. The major advantage of this technique is that it is a simple procedure that can be done with ease in any statistical package. There are however, more sophisticated methods of single imputation that can give each unit a different imputed value based on other observed variables (as opposed to giving the same mean value to all cases). These are called semi-parametric imputation methods. Among them, we have the Hot Deck and Nearest Neighbor techniques.

Semi-parametric imputation methods

The term 'Hot Deck' stems from the onset of survey research, when survey data was stored using punched cards. Hot deck imputation uses data from respondent that are similar to the respondents with missing values. Hot deck imputation finds values from "donors" that have complete values on all relevant variables, and transfers these values to the "beggars", who lack values for the target variables. The "donors" are selected for "beggars" with similar characteristics. Hot deck imputation only requires a moderate amount of work since you do not need to fit a model for the imputed variable. Since this method relies on using already observed values, the values cannot span outside the natural range of values observed in the original dataset (sometimes referred to as *univariate plausibility*). Also, a hot deck procedure will consistently produce the same estimates for all users (Marker, Judkins and Winglee 2002:329-330). But there are some limitations. Hot decks can only be used on univariate statistics as there are no methods of estimating variance that are appropriate for bivariate or multivariate statistics (Marker, Judkins and Winglee 2002:331). Generally, continuous predictors must be categorized for a hot deck to perform well, which can lead to a loss of information. It can sometimes be difficult to describe the procedures used to deal with cells with small samples transparently, making it difficult to replicate the data (Marker, Judkins and Winglee 2002:331). Attenuation of association is also a problem; imputed values might be perfectly reasonable when seen isolation, but might yield nonsensical results in contingency tables (Marker, Judkins and Winglee 2002:330 and Kalton and Kasprzyk 1986). Also, estimations of variance tend to be unnaturally low, leading to narrow confidence intervals increasing the risk of rejecting a false null hypothesis (type I error). Often, a single partition is used to impute a large number of variables. This is far from optimal, as different target variables usually have different sets of predictors that would give the best estimates for each variable.

Nearest Neighbor (NN) techniques are closely related to hot decks. NN is similar to hot deck imputation, but while hot decks must categorize continuous predictors, NN more fully utilizes continuous covariates. As with hot decks, NN leaves little or no randomness in the imputation process. NN techniques, like Bayesian methods, can use multiple continuous predictors, and have more transparent descriptions. Still, NN techniques suffer from a lot of the same drawbacks as other single imputation methods; most notably the underestimation of variance (Marker, Judkins and Winglee 2002). This is an important point that deserves some attention.

When single imputation has been performed, it is common to treat the data as true observations, thus allowing standard estimators of variance to be used. But these kinds of ad hoc solutions also have weaknesses. One of the serious weaknesses, according to Lee, Rancourt and Särndal (2002), is that this may lead to serious deflation of the variance. Variance can be estimated by

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

When $x = \bar{x}$, s^2 is equal to zero. If x is larger or smaller than $\bar{x}$, s^2 increases. Therefore, as long as the imputed value is within the observed range of values, the variance will decrease. This means that in practice, any single imputation is bound to decrease the variance. This will again lead to confidence intervals that are too narrow, increasing the likelihood of making a Type I error[4]. This problem can however be solved using multiple imputation (Rubin 1987).

Multiple Imputation

In the late 1980's, Rubin (1987) proposed a solution to the nonresponse problem that was *not* an ad hoc approach, such as the aforementioned techniques, but rather a principled and generalized method that could be used in all cases (Zhang 2003:584). This approach is called multiple imputation. As we have seen, the problem with single imputation is that the variance becomes unnaturally low. Rubin (1987) proposed an elegant solution to this problem; adding uncertainty to the estimated values for missing data. Since we do not know for certain what the missing value actually should be, in a multiple imputation, each missing value is replaced by $m > 1$ simulated values. These reflect the uncertainty we face. After several imputations, we are left with m number of data sets. These are in turn analyzed by standard complete-data methods, whose results are combined to produce confidence intervals and p-values that account for the uncertainty in the data (Rubin 1996:476). The simulated values are drawn

[4] A Type I error occurs when the researcher mistakenly rejects a true hypothesis

from a form a probability function. There are especially two different approaches than need mentioning, the Expectation Maximation (EM) Algorithm, and Markov Chain Monte Carlo (MCMC) methods.

Markov Chain Monte Carlo

Markov Chain Monte Carlo is a combination of two methods. The Monte Carlo method of statistical computation has its roots from the first electronic computers ever built. In the mid 1940's, John von Neumann and Stanislaw Ulam was working on the Manhattan Project, the United States secret nuclear weapon development program. One of their tasks was to develop a technique for simulating random draws. This secret project required secret code names. A college, Nick Metropoils (1987:127), suggested the name Monte Carlo, referring to the Monte Carlo Casino in Monaco. The suggestion was related to the fact that Ulam's uncle would borrow money from relatives because he "*just had to go to Monte Carlo*" to gamble. The name was very fitting, as both the technique and the casino relied on repeated random draws of numbers (Robert and Casella 2004:2). The basic problem was finding the probability of a particular event in a complex system. Ulam first developed the method to deal with a seemingly simple question: what is the probability that a particular solitaire would come out successfully? However, the combinatorial mathematics required to solve this simple question proved very hard. Instead, Ulam started to record the results of different plays and subsequently calculate the probabilities of each outcome (Andrieu, de Freitas, Doucet and Jordan 2001:1-2). This is a very slow procedure when done by hand, but with the invention of the computer, this became a feasible method for solving such difficult combinatorial mathematics. A Monte Carlo method uses simulations in this way to arrive at the probability of a certain outcome across numerous trials. The Monte Carlo method was then used to arrive at a desired distribution using a Markov Chain. A Markov Chain is a random draw of numbers that changes randomly in discrete steps. The probability distribution for the system at the next iteration depends only on the current state of the system, but not of the state of the system at previous steps. Here's an example: A normal dice has an equal probability of rolling each number, 1/6, regardless of what line of numbers have been rolled previously. If we have rolled three 6's in a row, the probability of rolling another 6 is still 1/6, no more, no less. This in contrast to a game of cards, where the probability of drawing a certain card would depend on which cards have previously been played, as there is a limited number of cards in the deck. According to Schafer (1997:3), these distributions stabilize to a common distribution; the *stationary distribution.* This stationary distribution is an already known distribution of

interest. Using this method of simulation, we can obtain values that are both plausible and contain variance, while upholding the covariance structure. MCMC is a form of Bayesian inference, which expresses information about an unknown posterior probability distribution. The main difference between Bayesian and frequentist statistics is that the Bayesian view assigns a probability of rejecting a hypothesis, in contrast to the frequentist view, which uses statistical hypothesis testing and confidence intervals. MCMC requires a fair bit of computing power. Because of this, other less computationally demanding techniques were requested. One such technique is the EM-algorithm.

The EM Algorithm

The Expectation Maximation (EM) algorithm is a form of maximum likelihood method of estimation. EM is an iterative method of estimating unknown parameters (p). Understanding and explaining the mathematics behind this procedure is quite hard (and beyond my mathematical skill). But the basic idea is based on the idea of alternating between an estimation step and a maximation step. This is an iterative procedure that continues until there is a convergence. Although the algorithm was formulated in the late 1950's, convergence was not proven until Dempster, Laird and Rubin formalized EM and published the mathematical proof in 1977 (Dellaert 2002:1). The EM algorithm proved to be much faster than the relatively slow MCMC method, but the method still produced results comparable to the MCMC method. Since then, EM spawned what Schafer (1997:3) describes as no less than a revolution in the analysis of incomplete data. The EM approach is applicable in a wide class of statistical problems, and should in theory have obliterated the need for ad hoc methods for treating nonresponse bias. But as we have seen in the first chapter, treating nonresponse in a proper manner is still not common among social science researchers. In all, there are several viable approaches to treating item nonresponse bias. But which method is best suited for my research question?

Semi-parametric or Multiple imputation?

Despite the advantages multiple imputation has over semi parametric imputation methods, there is some controversy in the literature. According to Marker, Judkins and Winglee (2002) there are situations where semi-parametric methods are advantageous. The Bayesian techniques, especially MCMC, are useful for all *parametric* distributions[5] (Marker, Judkins and Winglee 2002:331). The weakness is that Bayesian models require that an explicit

[5] A parametric distribution is one that can be specified by a set of parameters. These include the normal distribution, exponential distributions and logarithmic distributions.

parametric model is formulated for each target variable. This is a problem for variables that do not follow any known distribution, such as personal income. Skewness can be fitted using transformations, but there are no techniques for modeling discontinuities in the distribution. The semi-parametric methods; hot decks and Nearest Neighbor (NN) method, handles this without any problem. Any natural gap and limits in variable range will remain, as NN and hot decks never impute values into gaps or outside the original range of the variable. In these cases, semi-parametric methods outperform Bayesian methods. However, values outside of the observed limits and between integers, is not necessarily a problem, depending on what method of analysis you intend to use. For analysis that utilizes means or covariance structures, such as t-tests or correlations, such values are not a problem. But if you are using contingency tables, fractions will cause problems if the original data only contains integers. Consider if the original data contains the values 1, 2, 3 and 4, while the imputed data contains fractions, such as 1.112, 2.981 etc. While the original data can be expressed by a 4x4 contingency table, the imputed values might require a 400x400 contingency table, with one row and column for each fraction. It is however worth noting that this problem can be solved by categorizing the fractions or simply by rounding off the value. Nevertheless, Marker, Judkins and Winglee (2002:341) argue that hot decks and NN techniques are generally easier to implement, easier to explain to laymen, avoid dependence on prior distributions, and are flexible enough to handle nonstandard distributions. Further, Marker, Judkins and Winglee (2002:341) recommend that users avoid using multiple imputation since the improvement in quality is uncertain. Bayesian methods also require the specification of prior distributions on parameters, which Marker, Judkins and Wingle (2002:332) argue is baffling to most consumers of government statistics and should thus be avoided in official statistics (although they do admit it makes the process transparent to professionals). Bayesian methods are also more difficult to program, and thus more costly and time consuming. Rubin (1996, 2003) has answered to many of these criticisms. Rubin (1996) does to some degree recognize the increased workload that multiple imputation necessitates. But to the question: Is multiple imputation too much work for the user? Rubin (1996:480) answers with a rather good question: "*To much work relative to doing what?*". Rubin (1996:480) argues that even without the appropriate macros, which he argues are easy to write, multiple imputation is still the best method that can *validly* address nonresponse. Filling in the mean, only using complete data, and single imputation, are *not* statistically valid methods, not even for point estimates such as means, variances and correlations, and should not be used in public databases, Rubin (1996) argues. In a later discussion paper on the issue of multiple imputation, Rubin (2003:619) also

argue that multiple imputation is a generally applicable solution because it allows the researcher to capitalize on the already accepted complete data analysis, thereby avoiding the complications of using ad hoc statistical methods of analyzing data. In some cases, multiple imputation can lead to a loss in efficiency, or even loss of validity in rare cases, but Rubin (2003) argues that this is a small price to pay, considering the advantages of MI. Another major advantage of multiple imputation is that the imputation process can be separated from the researcher. Once the missing values have been imputed, the data can be shared and analyzed separately. This means that datasets can be imputed before being shared with researchers, thus saving the researcher the time and intellectual investment needed to fully understand the complexities of multiple imputation (Zhang 2003:590). Perhaps the greatest advantage is one that directly impacts the researcher: Standard methods of estimation can be used, making it easier for the average researcher to do the research. Although it is common in single imputations to treat imputed values as true observations, this approach is clearly flawed because of the inevitable variance deflation. Because of this, standard methods of variance estimation for both NN and hot decks do not provide good estimates. Instead, more complex methods of variance estimation must be used. This is especially problematic when multivariate statistics are required, as the semi-parametric methods are primarily suited for univariate statistics. The goal of this thesis is to find a method that is well suited for a practical application in the research society. Since multivariate analysis methods are such an important tool for social science research, the obvious choice of method is multiple imputation. The approach is generally applicable, which alleviates the need for ad hoc procedures. Multiple imputation is a more statistically valid method than semi-parametric methods. Multiple imputation is also more statistically efficient since it uses the entire observed dataset. This gives us more statistical power compared to other approaches, as we need fewer complete cases to achieve the same statistical power. Multiple imputation provides by far the most flexible method in terms of subsequent analysis of the data, and will therefore serve as my choice of imputation method.

Applying imputation to correct for unit nonresponse bias

In social science research, researchers are often interested in the relationship between variables and not just univariate point estimates, such as means. As such, we often use bi- or multivariate analytical models such as covariance matrixes and regression analysis. When dealing with incomplete data, both weighting and imputation methods have their uses. Traditionally, weighting is used to correct for unit nonresponse while imputation is used to

correct for item nonresponse. Weighting works very well for point estimation, but incomplete data leads to a loss of information and statistical power in analysis that rely more heavily on complete cases. Imputation on the other hand utilizes all the information in the dataset and is therefore more efficient. Some statistical software packages, such as SAS v9.2 allow for both methods to be used simultaneously (Berglund 2010). This allows the researcher to both impute missing items and to correct for unit nonresponse bias through weighting procedures. But to my knowledge this is the only software that has this functionality. As such, it would be helpful if a single, commonly available method would be able to correct for both item- and unit nonresponse at the same time. The solution I propose in this thesis is to use multiple imputation to correct for *unit* nonresponse as well as for its common application of *item* nonresponse. As mentioned in the introduction, evaluations of the application of multiple imputation for unit nonresponse are very hard to come by (Rässler 2003). In fact, Rässler (2003) is the only article I could find on this subject. Rässler (2003:15) does however conclude that using multiple imputation came closer to the unknown true values than weighting methods. This conclusion is based on the fact that the multiple imputation gave results that were in the expected direction. Since the true values were unknown, this conclusion might be overconfident. Still, the results are encouraging for further inquiry into the use of multiple imputation for unit-nonresponse. For this approach to be viable, we have to be able to treat the unit nonresponse as a special case of item nonresponse. This is only possible in datasets that have at least a limited amount of information about all survey recipients. The European Social Survey 2006 data for Norway provides just that. To see if the data is applicable for bias correction, I will first have to check the dataset for unit nonresponse *bias*. In this process, I will use within survey variation as well as paradata to obtain information about the nonresponse mechanism. This should allow me to find out if the continuum of resistance or the classes of nonparticipants model is better suited to explain the missingness mechanism in the dataset. This information will also be used to create an imputation model that should be able to correct for both item- and unit nonresponse. MCMC and EM-algorithm are two different approaches to multiple imputation. Since there is little publicized research on the advantages and disadvantages of these methods, it would be relevant to see if the two methods produce different results. To answer this question, I will use both the MCMC and the EM-algorithm to impute the data. After reviewing the literature, I have come across two secondary research questions in addition to my main question. To sum up, the three research questions are:

1. *Can multiple imputation be used to correct for unit-nonresponse bias in survey data that contains only a limited amount of information about the nonrespondents*?
2. *Is there a "continuum of resistance" or does the "classes of nonparticipants" model better explain the missingness mechanism?*
3. *Do MCMC and EM produce different results in terms of imputed values and subsequent regression analysis results?*

In the next chapter, I will first discuss the data source. I will proceed by checking for nonresponse bias and testing the two hypothesis related to the missingness mechanism. Next, I will look at the assumptions of multiple imputation and discuss the two software packages I will be using.

3. Data source and methods

I am using the European Social Survey data from Norway 2006 (round 3) in this thesis[6]. As of now, four rounds have been performed on a biannual basis, in 2002, 2004, 2006 and 2008. As such, the ESS is a fairly recent study compared to other such studies (for example the World Value Survey and European Values Survey). Nevertheless, it has made a solid mark in the field of Social Sciences. The European Social Survey has from the outset aimed to be a high quality survey. The goal of the ESS is not only to give produce high quality surveys, but to also improve research methodology in the social sciences in general (Billet, Philippens, Fitzgerald and Stoop 2007:136-138). In 2005, the efforts made by the ESS were recognized; the ESS was the first social scientific research program to be awarded the prestigious Descartes Prize (The European Social Survey Webpage 2010).

The survey consists of roughly 200 questions, and two additional rotating modules of questions that are selected among suggestions from different research groups. The European Social Survey for 2006 had a goal of achieving a response rate of 70 percent. Despite elaborate testing and preparation, only about half of the countries in the first round achieved this goal, with five countries having response rates lower than 50 percent. For all but one country, refusal to participate was the most common source of nonresponse. The goal of having low (less than 3 percent) levels of noncontact on the other hand, was more successful,

[6] The data is prepared by Statistics Norway (Statistisk sentralbyrå). Any errors in this thesis are solely my responsibility.

with only 6 out of 22 countries superseding this limit. In Norway, the response rate was 65.5 percent (Billet, Matsuo, Beullens and Vehovar 2009:1-3).

I am using is a dataset prepared by Statistics Norway (Statistisk Sentralbyrå) specifically for nonresponse research. This is dataset that contains contact data and information from the population register about all recipients of the survey, in addition to the survey questions. The data also contains data from a limited follow-up survey. The target population is people of ages 15 years and older. Simple random sampling has been used to draw a sample from the population register. All persons received a pre-notification letter in advance. A lottery ticket was the incentive for the advance letter. Initial refusals received a motivational letter, emphasizing the purpose of the interview. If the person could not be persuaded, the respondent was contacted by a new interviewer. The total number of cases in the gross sample is 2750. Altogether, 1750 persons participated in the survey. 77 respondents are system missing. There is no data whatsoever on these missing participants (not even data from the population register). It is unclear why these are system missing, as opposed to being coded as non-contact. These are therefore excluded from the data, leaving me with a total number of 2673 survey recipients.

Table 1. Types of final response

Type	Frequency	Percent
Respondent	1750	65.5
Refusal	735	27.5
Unable	61	2.3
Language problem	28	1.0
Non contact	62	2.3
Other	37	1.4
Total valid	2673	100.0
System Missing	77	
Total Gross Sample	2750	

As is the case in almost all of the countries in the survey, the largest group of nonrespondents is the refusals. After attempts at refusal conversion, 735 persons ended up refusing to participate. 61 respondents were unable to participate, either because of psychological reasons or because it was not physically possible to obtain an interview. 28 did not participate because of language problems. 62 persons could not be contacted. 37 did not participate for other reasons than the main categories.

Before testing for bias, I have taken closer look at the different types of nonrespondents. I have looked the mean values for some of the relevant variables from the population register in table 2. Some of these groups have a low valid N, and the variables are not normally distributed for all subgroups of nonrespondents. Therefore, parametric tests are not applicable to test for significant differences between these groups. The *age* variable is continuous. *Number of persons in household* is divided into 1, 2, 3, and 4 or more persons. *Level of education* is divided into low (1), middle (2) and high or university (3) level of education. The *centrality of municipality* variable ranges from *least central*, (0) to *central municipality* (3). Because of the ordinal measurement level of *education*, *number of persons in household*, and *centrality of municipality*, looking at frequencies and percentages would be more statistically correct. However, I would argue that since these variables are approximately continuous, a mean comparison makes the interpretation easier.

Table 2. Means for different types of final response

Variable	Respondent	Refusal	Unable	Language problem	Non contact	Other
Age	45.369	47.422	67.574	39.429	38.79	44.189
Number of persons in household	2.45	2.37	1.98	2.54	2.05	2.49
Centrality of municipality	2.18	2.25	2.13	2.5	2.39	2.27
Level of education	2.09	1.85	1.67	1.14	1.79	2.03
Valid N	1750	735	61	28	62	37

The nonrespondents differ from the respondents in several respects. The refusals are slightly older than the respondents. They also have a lower mean level of education. That the non contacts are younger and less educated than the respondents. Non contacts also live in households with on average fewer people than most other groups. This fits well with the observations made by Stoop (2004:35) about non contacts; they are likely to be young (and thus have not completed a university degree) and often spend time out side of the home. Both non-contacts and the unable group live in households with a mean value of about 2 persons. This supports the notion that those who are unable to participate are mostly elderly people (perhaps couples where all children have moved out) with health problems causing them to be unable to participate. The non contacts are likely to be young people or couples who in general have not yet had children. The *Other*-group seems to be the group that is most similar to the respondents, and they are likely to be a fairly random subsample with non-systematic reasons for refusal. Those in the Language problem-group are likely to be of a minority background. They have the lowest levels of education and have a somewhat younger mean

age than the respondents. They also live in households with the most people, and live in more central municipalities than all the other groups.

The following variables contain complete data for all contacted recipients of the survey: gender, age of respondent, number of people in household from register, county, centrality of municipality and highest level of education. In addition we have paradata on the number of visits the interviewer has made, the degree of cooperation and type of final response. Having access to this information makes it easier to determine bias and eventually correct for it. As we have already seen in table 2, there are some differences between the groups of final response type, suggesting that there might be bias caused by nonresponse. Next, we will take a closer look at the potential for bias caused by nonresponse by testing for differences between the respondents and nonrespondents.

Determining bias in the ESS 2006

The first step in determining if the data is biased is to look at how the respondents differ from the nonrespondents. If the data is biased, there should be significant differences between the respondents and the nonrespondents. I have compared the respondent's and nonrespondent's composition on the variables; gender, age, education, centrality of municipality, and number of persons in household, using contingency tables. I chose contingency tables since most of the variables are categorical. The exception is age. The reason why age is also tested using contingency tables is because it will give more information about the differences than a simple t-test. Even if the mean values are similar, there might be differences in the distributions of the age variable that will not be uncovered using a mean comparison. Age has been coded into groups spanning 10 years, with the oldest category being 66 years and older. This gives a roughly equal number of persons in each group. A chi square test will determine the level of significance. As we have seen above, there are some differences in the means of the different types of respondents. However, the important question is whether or not the nonrespondents as a whole are different from the respondent, not if there are differences between subgroups of nonrespondents.

Table 3 shows that the distribution of age groups is significantly different between nonrespondents and respondents. As seen in the previous section, there is a larger proportion of older people among the nonrespondents. It seems that middle aged people are more likely to participate, while younger and older people are more likely to be nonrespondents. This fits well with the differences seen in mean age among the different categories of nonrespondents.

Table 3. Contingency tables for respondents and nonrespondents

Variable	Respondent	Nonrespondent	Sig	Chi square	Df
Age group			0.001	24.363	5
15 through 25	16.1 %	17.2 %			
26 through 35	16.6 %	14.0 %			
36 through 45	19.4 %	17.8 %			
46 through 55	18.4 %	15.9 %			
56 through 65	14.9 %	13.5 %			
66 and older	14.6 %	21.6 %			
Gender			0.001	10.039	1
Male	51.0 %	44.5 %			
Woman	49.0 %	55.5 %			
Municipality			0.001	16.208	3
Least central	14.9 %	10.4 %			
Less remote	7.0 %	7.3 %			
Fairly central	23.0 %	28.4 %			
Central	55.1 %	54.0 %			
Level of Education			0.001	98.463	2
Low	17.90 %	31.00 %			
Middle	55.00 %	55.80 %			
High (Univ)	27.10 %	13.20 %			
Number of persons in household			0.042	8.18	3
1	28.9 %	30.4 %			
2	26.1 %	29.7 %			
3	16.4 %	15.9 %			
4 or more	28.6 %	23.9 %			
Sum	100 %	100 %			
(n)	1750	923			

In total, there are slightly more woman than men in the sample. For final respondents, the distribution is the opposite, with slightly more males than females. For non respondents however, there is a proportionally larger amount of women. The difference between the groups is statistically significant. This seems to indicate that women are less inclined to participate in the survey than men.

There are some small but significant differences in centrality of municipality between respondents and non respondents. The distribution is very similar in central and less remote municipalities, while there is a small difference between the fairly central and the least central municipalities. 28.4 percent of the nonrespondents reside in a fairly central municipality as opposed to 23 percent among respondents. Remarkably, the proportion of respondents from least central municipalities is larger (14.9 percent) than that of the non respondents (10.4 percent). The only difference between municipalities seems to be that people from the least

central municipalities are more prone to participate while those in fairly central municipalities tend to be more prone to not participate.

The proportion of people with low education is higher among non respondents (31%) than among respondents (17.9%). The proportion of highly educated is smaller for nonrespondents (13.2%) than for the final respondents (27.1%), while the proportion having a mid-level education is very similar. The differences are statistically significant. This suggests that lower educated respondents are more likely to refuse, and higher educated are more likely to participate in surveys. Differences in level of education have been found in other studies as well (Vehovar 2007, Billet, Philippens, Fitzgerald and Stoop 2007)

Nonrespondents are slightly more likely to live in a household with 2 or fewer persons, and are less likely to live in a household with 4 or more persons. But the differences are quite small and barely significant at a 0.05 level.

This table shows us that there are significant differences between the respondents and nonrespondents in composition of gender, centrality of municipality, number of persons in household, age and in level of education. Other studies have found similar results as well (Billet, Philippens, Fitzgerald and Stoop 2007, Vehovar 2007). To sum up, middle aged and higher educated people are overrepresented in the survey. Women, people with lower levels of education, and the young and the elderly are underrepresented. If there was no bias in the data, the groups should not be significantly different from each other. The results so far suggest that the data is biased by nonresponse. To delve deeper into the mechanisms behind the bias I will look at the variation within different groups of survey participants and nonparticipants.

Variation within the survey

The ESS 2006 data for Norway contains valuable paradata for the respondents. This data allows us to separate between levels of cooperation within the survey. This information can be used to test the "continuum of resistance" and "classes of nonparticipants" hypothesis. To test which hypothesis fits the data, I have made a distinction between reluctant and cooperative *respondents*. In addition, we can use this information to see if there are differences between those who *refused* to participate early in the survey and those who refused after being re-contacted. When comparing the two groups of nonrespondents, the only basis for comparison will be the data gathered from the population register.

In total, 339 initial refusals were re-contacted. These are respondents who were reluctant to participate in the survey, but who were deemed to be candidates for refusal conversion. Sadly, there is no direct information about who these respondents are. According to the data file documentation, all refusals with the exception of those marked "*Will definitely not cooperate in the future*" on the cooperation rate variable, were re-contacted. This category was instructed to be used only in the most extreme cases. However, this does not seem to be the case in the data, as there are respondents who were put in this category, but were re-contacted anyway. Because of this, I have tried to find the re-contacted refusals by following the steps in the protocol described in the data file documentation. According to the protocol, respondents who refused to participate were reassigned to a new interviewer, which made a second attempt at convincing the persons to participate. Thus, I have coded *reluctant respondents* as respondents who have refused an interview and subsequently been reassigned to a new interviewer. Using this approach, I find a total number of 370 reluctant and subsequently reassigned respondents. This is somewhat higher than the reported number of 339 recontacted respondents reported in the data file documentation. I have tried to use other variables, such as estimation of cooperation rate to get further to the true number, but answers in this variable are not consistent with the other the variables. As such, I will have to make do with the estimate of 370 reluctant respondents as opposed to the real number of 339. According to this approach, there are 79 *reluctant respondents* among the *final respondents*. These respondents refused initially, but were subsequently reassigned to another interviewer and were persuaded to participate. Among the *final refusals* there are 291 *reluctant nonrespondents* that were candidates for refusal conversion, but finally refused to participate. This allows me to divide the respondents and nonrespondents into four groups. First, we have the *cooperative respondents*, who participated without the need for incentives or conversion. Secondly, we have the reluctant respondents, who refused initially but were subsequently reassigned to another interviewer and converted. The third group is the reluctant refusals (or reluctant nonrespondents) who were reassigned but could not be converted. The final group is initial refusals, who were not subject to refusal conversion. These 444 persons refused to participate on the first visit, and were not reproached for refusal conversion, presumably because they showed no signs of cooperation. This leaves us with four groups, as showed in Figure 1.

Figure 1. Groups of respondents

Respondents	1750		1671	Cooperative respondents
		Reluctant	79	Reluctant respondents
Refusals	735	370	291	Reluctant refusals
			444	Initial refusals

If the continuum of resistance hypothesis is correct, there should be differences between these groups, and the differences should go in the same direction. For example, there should be a trend towards lower education the more negative the response. My first hypothesis is;

Hypothesis 1: The more reluctant the respondent is, the more the mean values of education, centrality of municipality, age and number of persons in household deviate from the cooperative respondents' mean.

Comparing these groups will hopefully shed some light on the two different models of nonparticipants; the continuum of resistance-, and the classes of nonparticipants-model. If there is a continuum of resistance, I would assume that the harder the respondent is to get an interview with, the more different they are from the cooperative respondents. In other words, the reluctant respondents should be different to the cooperative respondents. The reluctant respondents should also be different from the reluctant refusals. And finally, the initial refusals should be different from the reluctant refusals. In contrast, if there is a marked distinction between cooperative respondents and the other groups as a whole, this might indicate that the only real distinction is between cooperative and non-cooperative respondents, thereby supporting the "classes of nonparticipant" model.

Table 4 shows the mean values for the different groups. If the continuum of resistance hypothesis is true, there should be differences among the levels of cooperation. The table below seems to indicate that the cooperative respondents are different from the rest, but that the different groups of non-cooperative respondents do not differ in the few variables available. The cooperative respondents are on average younger, from slightly more central municipalities, and have lower levels of education than the reluctant and refusing groups. Number of persons in the household does not seem to follow any such pattern. There is still a possibility that these types of respondents are different from each other on other values, such as attitude towards immigration or political trust. Still, it is not farfetched to believe that the

reluctant respondents and the nonrespondents are more similar to each other than to the cooperative respondents.

Table 4. Means for different levels of cooperation

	Level of cooperation			
Variable	Cooperative Respondent	Reluctant respondent	Reluctant nonrespondent	Initial refusal
Age of respondent	45.22	48.46	48.51	47.33
Number of persons in household	2.44	2.51	2.31	2.42
Centrality of municipality	2.18	2.28	2.21	2.28
Level of education	2.10	1.86	1.87	1.85
Valid N	1671	79	291	414

So far, it seems that the cooperative respondents are different from the other levels of cooperation. In the previous section, we found that the nonrespondents are significantly different from the respondents. What remains is to find out if the reluctant respondents are *significantly* different from the cooperative respondents, and whether or not they are part of the same demographic as the nonrespondents. If they are, they can potentially give valuable insights into the characteristics of the nonrespondents.

I have tested the differences between reluctant respondents and cooperative respondents below using the same variables as above. Reluctant respondents are more likely to have lower levels of education and have a larger proportion of women, similar to the nonrespondents in the above chapter. There are no significant differences in centrality of municipality and number of people in household. Previous research has showed similar results; Billet, Philippens, Fitzgerald and Stoop (2007:150) found that converted refusals had higher proportion of lower educated people. However, in Germany, a higher proportion of converted refusals were living in big cities and suburbs, but this distinction is not found here. (Only the significant and close to significant results are shown here)

Table 5. Contingency tables for reluctant and cooperative respondents

Variable	Reluctant respondent	Cooperative respondent	Sig	Chi square	Df
Age group			0.023	13.094	5
15 through 25	17.7 %	16.0 %			
26 through 35	16.5 %	16.6 %			
36 through 45	16.5 %	19.6 %			
46 through 55	6.3 %	19.0 %			
56 through 65	19.0 %	14.7 %			
66 and older	24.1 %	14.1 %			
Gender			0.057	3.626	1
Male	40.5 %	51.5 %			
Woman	59.5 %	48.5 %			
Level of Education			0.005	10.725	2
Low	30.4 %	17.3 %			
Middle	53.2 %	55.1 %			
High (Univ)	16.6 %	27.6 %			
Sum	100 %	100 %			
(n)	79	1671			

I have also tested the difference between reluctant respondents and reluctant nonrespondents, as well as between reluctant respondents and initial refusals. There were no significant differences between reluctant respondents and the other groups. Therefore, I have to reject hypothesis 1: *The more reluctant the respondent is, the more the mean values of education, centrality of municipality, age and number of persons in household deviate from the cooperative respondents' mean.*

This lends some support to the "*classes of nonparticipants*" hypothesis. There is no sign of a continuum of resistance in these variables. On the other hand, there is some evidence to suggest that cooperative respondents are significantly different from the reluctant respondents and refusals. Also, the reluctant and refusing respondents are quite similar to each other, suggesting that they might be part of the same group.

Assessing bias using information from follow-up survey

As a final approach to assess bias, I will look at a comparison of the variables in the follow-up survey. There are 242 nonrespondents who were persuaded into participating in a short nonresponse survey, along with about 245 of the original respondents. This survey only has 12 questions, some of which have too low response rates to be of any use (especially *main activity last seven days*, where only 473 gave an answer in the original survey). My second hypothesis is;

Hypothesis 2: Reluctant respondents are more similar to the refusals than to the cooperative respondents on the follow-up survey variables.

I have compared the mean values of the cooperative respondents to the reluctant respondents and the initial refusals. In addition, I have compared the answers in the original survey to those of the follow-up survey for the cooperative and reluctant respondents. This is to see if there is response variance between the values given in the original survey and the follow-up survey.

There are however several problems with using the follow-up survey. The number of valid cases is low, especially when comparing the few reluctant respondents' values in the original and follow-up survey (only 14 of the reluctant respondents participated in the follow-up). For a comparison to be valid, any variability in response deviations should be low. If the answers from the original data are different from the responses obtained in the follow-up, the responses may be unreliable. This makes it difficult to draw any conclusion about differences between respondents to the main survey and respondents of the follow-up survey. Regrettably, the responses are not very reliable for participants of both surveys, with correlations between answers ranging from 0.510 to 0.684[7]. Table 12 shows the mean values for cooperative and reluctant respondents in the original and the follow-up survey.

Table 6. Means in original and follow-up survey

	Cooperative respondent		Reluctant respondent	
	Original Survey	Follow-up	Original Survey	Follow up
TV watching	3.72	3.62	4.54	4.71
Take part in social activities	2.96	2.86	2.96	2.57
Feeling of safety	1.60	1.50	1.63	1.57
Trust in people	6.85	6.81	6.36	5.43
Political interest	2.48	2.58	2.22	3.00
Satisfaction with democracy	6.62	6.04	6.44	5.47
Trust in politicians	4.47	4.75	3.97	3.67
Attitude towards immigration	5.11	4.53	4.42	3.33
Involved in work for charity	4.22	3.91	4.69	4.93
Valid N (listwise)	1578	241	79	14

Since the two surveys were given at different points in time, some attitudes might be prone to change in light of recent events, such as *trust in politicians*, and *attitudes towards immigration*. If we compare the mean value of political trust in the original survey for

[7] See appendix table 1 for full list of correlations

example, with the nonparticipant's means at the time of the follow up survey, we might risk drawing the wrong conclusion. Any change between the two groups might be due to a change in the general attitude of the population, perhaps in light of a political scandal, and not because of differences in the two groups. But when comparing the mean values of respondents who participated in both the follow up and the original survey, their mean values are however mostly consistent. One notable exception is satisfaction with democracy, which decreases from 6.64 to 6.04 from the original answer to the follow-up. This difference is found even when using listwise deletion and comparing only the 241 that gave valid answers in both surveys, in which the means are 6.63 and 6.02 respectively. Attitudes towards immigration also decreased from 5.04 to 4.53. Using listwise deletion here also yields very similar means, of 5.05 and 4.53. The changes might indicate a change due to current events. If for example in the after the original survey, a crime committed by immigrants got a lot of media coverage, and politicians got the blame for not being able to enforce stronger regulations, such a change in public opinion could be a consequence. However, trust in politicians is increased from 4.47 to 4.75, which is not consistent with this particular hypothesized event (Once again, using listwise deletion here also yields very similar means, 4.42 and 4.75). There are quite large deviations between reluctant respondents in the original and follow-up survey. But comparing the reluctant respondents in the two surveys is very dubious, as only 14 participated in the follow-up survey, which is too low to be considered representative. Since there are some serious deviations between the means and relatively low correlations between the two points, I am, not too confident in the results below.

Table 7 shows a comparison of the means in the original survey participants, the reluctant respondents and the initial refusals who participated in the follow-up. As I hypothesized, the reluctant respondents show some similar tendencies to the refusals on several variables. For example, they both deviate from the cooperative respondents in all but two variables (*social participation* and *political interest*). However, the cases of political interest and trust in politicians, the refusals are closer to the cooperative respondents than to the reluctant respondents. Still, this is only the case in these two variables; in all other cases, the refusals are more similar to the reluctant respondents.

Table 7. Means for different levels of cooperation on follow-up survey variables

	Level of cooperation		
Variable	Cooperative	Reluctant	Refusal
TV watching	3.72	4.54	4.26
Take part in social activities	2.96	2.96	2.69
Feeling of safety	1.60	1.63	1.67
Trust in people	6.85	6.36	6.55
Political interest	2.48	2.22	2.79
Satisfaction with democracy	6.62	6.44	5.91
Trust in politicians	4.47	3.97	4.32
Attitude towards immigration	5.11	4.42	4.35
Involved in work for charity	4.22	4.69	4.54
Valid N (listwise)	1578	79	215

The two variables that are most clearly similar are TV-watching and attitudes towards immigration. For TV-watching, the mean of the cooperative respondents is 3.72, whilst the reluctant and refusals have 4.54 and 4.26 respectively. Cooperative respondents also show a more positive attitude towards immigration than the reluctant respondents and refusals, with a mean of 5.11 compared to 4.42 and 4.35. In both these variables, the reluctant respondents are slightly more dissimilar to the cooperative respondents than the refusals, once again supporting the notion that there is no "*continuum of resistance*", and that the classes of nonparticipants model is more fitting in this dataset. The refusals have a lower value for social participation than both the respondent groups (who have almost identical means). There is a negligible difference in feeling of safety when walking alone. Cooperative respondents have a slightly higher trust in other people, but this difference is too small to be considered significant. There are differences in political interest. A score of 1 is *very interested*, while 4 is *not at all interested*. Surprisingly, reluctant respondents are more interested in politics than the other groups, with the refusals being the least interested. Another surprising find is that refusals have the lowest satisfaction with the way democracy works, compared to the cooperative and reluctant respondents. But as table 6 shows, there seemed to be a general decrease in this variable from the time of the original survey to the follow-up survey. The reluctant respondents have lower trust in politicians than the cooperative respondents and the refusals who participated in the follow-up study.

While the population register data seems to be clearly affected by nonresponse, bias is more difficult to trace in the variables in the follow-up survey. The only two variables in the follow-up that can be said to be biased with any degree of certainty are *TV-watching* and *attitudes*

towards immigration. Of course, we cannot expect the refusals to be different to the cooperative respondents in every way. Some variables will be more affected by bias than others. When looking at which variables show little sign of bias, there are some variables that are, unsurprisingly, similar to each other. Taking part in social activities is slightly lower for the refusals than the other groups. When it comes to *fear of walking alone at night*, there are reasons to believe that there should be differences; if nonparticipants are skeptical towards immigration and watch more TV, they could very well also be more paranoid about being assaulted. Cooperative respondents have a slightly higher trust in other people, but the difference is not as marked as with immigration or TV watching. For trust in politicians, I would expect a large difference between cooperative respondents and nonrespondents, but there hardly is one. Finally, being involved in charitable work is more common in the cooperative respondents than the other groups (This variable goes in the opposite direction of most of the other variables. Hence, a low value is often, a large value is seldom or never). Although the reluctant respondents are similar to the refusals in some variables, the differences are in many cases very small or not in the expected direction.

The time difference and the discrepancies between answers from the main survey and the follow-up study make comparisons difficult. Are the differences between cooperative respondents and initial refusals due to differences among groups of participants, or are they heavily influenced by current events? Cooperative respondents are more negative towards immigration in the follow up study, and the same is the case for the reluctant respondents. Is this due to current events or are people who are more negative towards immigration are more prone to participate in a follow up? Or are the answers in general too unreliable to draw any conclusion? The two main findings that prove to be robust across all comparison are that reluctant respondents and initial refusals seem to watch more TV and are more negative towards immigration. Also, on all but two variables, the refusals have more similar means to the reluctant respondents than to the cooperative respondents. These findings do support the hypothesis that refusals are, in general, more similar to the reluctant than to the cooperative respondents. With such low reliability I am reluctant to draw any conclusion besides to two strongest findings. Nonrespondents are more similar to reluctant respondents on the amount of TV the watch and their attitude towards immigration. Hypothesis 2: *Reluctant respondents are more similar to the refusals than to the cooperative respondents on the follow-up survey variables*, is only partially confirmed.

Conclusion on bias

The data is clearly biased in the population register variables as well as in at least a couple of survey variables and is thus a good candidate for testing the application of multiple imputation in correcting for unit nonresponse bias. The cooperative respondents have higher levels of education, are slightly less prone to live in central areas, and are younger than the other groups of respondents. Taken together, the results of the bias analysis indicate that there are different classes of nonparticipants, and not a continuum of resistance in this survey. It seems that the groups of nonrespondents are more similar to the reluctant respondents than to the cooperative respondents on several variables, but there is little evidence to suggest that nonrespondents have more extreme values than reluctant respondents. If we assume that the reluctant respondents are representative of the nonrespondents, any variable that is associated with reluctance is likely to also be associated to the missing-mechanism. If I can find other variables in which the cooperative and reluctant respondents differ, I can use these variables to create a better imputation model. Hopefully, including such variables will result in a better imputation model. Next, I will briefly discuss the assumptions of the method of imputation and the limitations and advantages of the statistical software I will be using.

Multiple imputation – methods and assumptions

Multiple imputation is commonly used to impute missing values on single items. This goal of this thesis is to use multiple imputation to deal with bias caused by *unit* nonresponse. Performing an imputation without any information at all is of course impossible. Multiple imputation uses the posterior distribution of the observed data to predict a plausible value for missing values. So in order to perform a multiple imputation, the dataset needs to somehow be augmented. The European Social Survey contains both some individual background data from the population register, as well as paradata. These variables are complete for the entire gross sample. This makes it possible to treat the situation as a special case of item nonresponse. The big question is whether or not this limited information is enough to correct for the nonresponse bias, or if the procedure will have the same effect as a bootstrapping procedure, namely just increasing the sample size and thus also the statistical power. This thesis deals with the practical implementation of such an approach. One approach could be to use data from the follow-up survey to correct for bias. But the follow-up survey is problematic in some respects, most notably the low reliability of the answers. In most cases, data from follow-up surveys are not always available to the researcher. A more common resource however is register data and paradata. Since I want to explore a practical and generally applicable approach, I will only use the more common auxiliary information from the population register

and paradata to correct for bias. To find out if multiple imputation can be used in this special case, I will use two statistical packages, each with a different algorithm, namely SPSS with the Markov Chain Monte Carlo and Amelia with the Expectation Maximation algorithm. This will allow me to see how stable an imputation model is across different algorithms. First, I will look at the assumptions of doing a multiple imputation, followed by a discussion on the two software packages and their algorithms.

Missing at random assumption

For an imputation to be valid, the data must be missing at random. As previously noted, this means that the missing data is dependent on observed variables and values, as opposed to unobserved ones. In other words, the data can be biased by nonresponse as long as you have information about respondents that can be used to correct for the bias. As such, imputation in the case of unit nonresponse is only possible if the few variables available can turn the situation from a NMAR to a MAR situation. With both paradata and individual background data available, this assumption should be met in the European Social Survey data.

The imputation model must be proper

The model used to generate the imputed values must be "proper" according to Rubin's rule (1987). Rubin's rule (1987) states, that a multiple imputation is proper when all the sources of variability that affects the imputed values have been incorporated into the imputation procedure. Multiple imputation requires the specification of two models; the imputation model and the analysis model. It is generally believed that this assumption is met when the model used for the multiple imputation and the analysis are compatible, although there are some objections to the validity of this belief (Nielsen 2003). This assumption should be met as long as the imputation model is at least as complex as the analysis model. Rubin (1996) advices using as many predictors as possible when performing multiple imputation. But according to Schafer (1999:6), analyses that utilize means, variances and covariances, such as regression, should perform well even if the imputation model is rather simple[8]. However, analysis that are more sensitive to tail behavior should use more complex imputation models (Schafer 1999:6). In social science research, such robust analysis methods are often used. If performing a multiple imputation to correct for unit nonresponse does not affect the results of a regression analysis, there is perhaps little reason for social scientists to invest a lot of effort into creating a very complex imputation model. In light of this, I have a third hypothesis;

[8] For more on model specification, see Rubin 1987, Rubin 1996 and Schafer 1997.

Hypothesis 3. The substantial interpretation of a regression analysis will not differ between a simple and a complex imputation model.

To see how sensitive my analysis model is, I will specify both a simple and a complex imputation model, and see if the two give different results. The simple imputation model will only use the variables in the analysis model. To create the complex imputation model, I will perform a data exploration to find out what variables might be associated with nonresponse.

Number of imputations

One of the main advantages of multiple imputation is to be able to account for the true variance, and thus to avoid using complicated methods of variance estimation. The way to achieve this is by creating several imputed datasets and to use the average of the estimators. The recommended number of datasets, *m*, varies. According to Rubin (1987:15), there is very little benefit to using more than 2-10 imputations unless the rate of nonresponse is very high, (which they arguably are in this case). The Stata Reference Manual (2004) for multiple imputation recommends using at least 20 imputations, but as many as 50 have been shown to be required in some cases (Kenward and Carpenter 2007, Horton and Lipsitz 2001). Since these recommendations vary, I will be using a moderate number of imputations[9] (10) in the simple imputation model, and a very high number (50) in the complex imputation model. This will allow me to see how big a difference the number of imputations has on the estimates[10].

Assumption of normally distributed and continuous variables

For an imputation to produce good estimates, the data has to come from a continuous multivariate distribution and contain missing values that can occur on any of the variables. Also, the data must be from a multivariate normal distribution when either the regression method or MCMC method is used (Schafer 1997:9-10). As such, categorical data are not suited for imputation, but there are ways of getting around this problem in most statistical packages. This is handled differently by SPSS and AMELIA, and will therefore be discussed further in the software section.

Arbitrary vs monotone missingness

The pattern of missing data has consequences for the choice of imputation method. Monotone patterns allow for a bigger range of methods, while an arbitrary pattern necessitates the use of more complex imputation methods. A data matrix has a monotone missing data pattern when,

[9] One could argue that the simple model should use a lower number of imputations. However, with the data processing power available today and the ease of use of most statistical software, it is unlikely that any researcher would run an imputation with a lower number than 10 imputations.

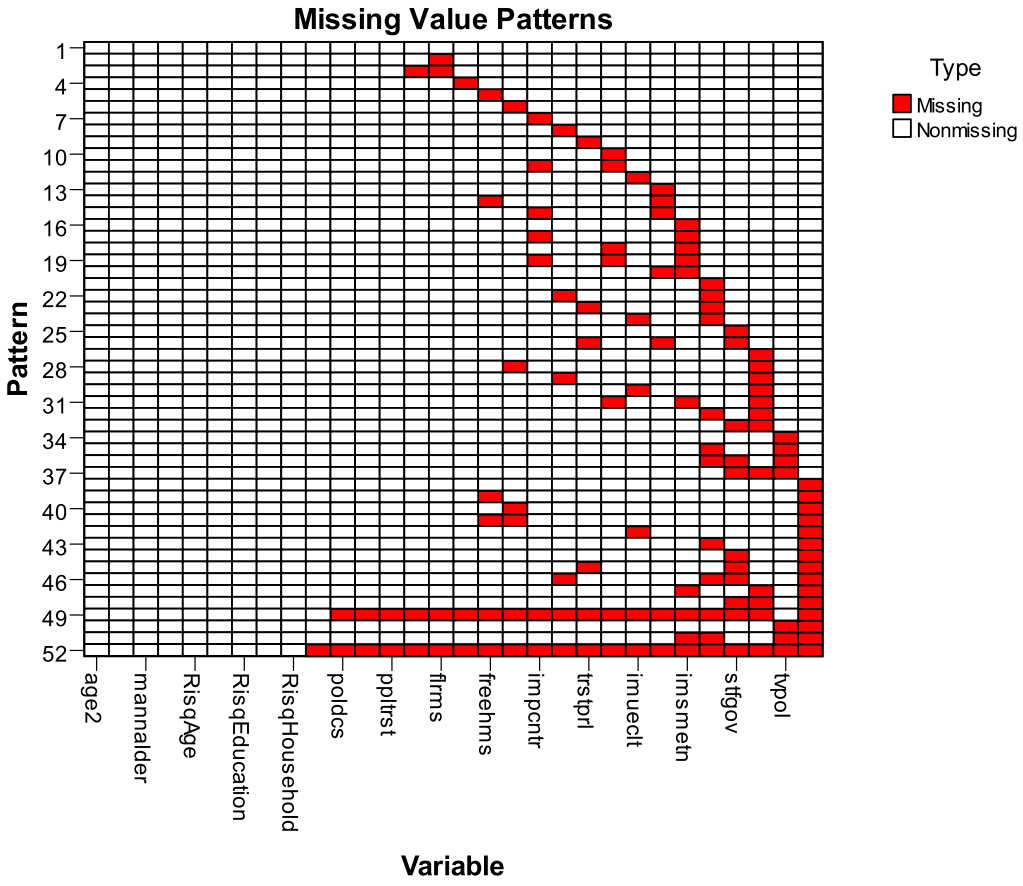
[10] For more details about imputation modeling, see Rubin (1996), Schafer (1997, 139–144), Kenward and Carpenter (2007).

for example, a respondent refuses to participate any longer, and all subsequent values from that point onward are missing. This is common in medical clinical trials with repeated measures. Some patients will drop out at a certain time, and all subsequent measurements will of course be missing from this point. In mathematical terms; whenever an element y_{ij} is missing, the elements y_{ik} are also missing for all k > j. This will give a staircase-shaped pattern (Zhang 2003:586). If the missing data has a monotone pattern, several methods can be used. A monotone missing data pattern can be imputed from independent univariate distributions. If the data is nonmonotone, more complex methods that use a multivariate distribution, such as EM or MCMC must be used. Because of the nature of the data normally found in social surveys, such a nonmonotone pattern of missing data should be very rare. In the social sciences, there are often several different reasons why the data is missing (do not want to answer, do not know, not applicable), as opposed to the typical example of a nonmonotone pattern, in which patient drop outs are most common.

Figure 2 shows the missing data pattern for the ESS 2006 data[11]. The variables are ordered from left to right in increasing order of missing values. Each line represents a group of cases with similar patterns of missing values. For example, pattern 52 seems to be nonrespondents, as they have missing values on all variables save for the variables from the population register (variables that start with *"Risq"*). The other patterns are groups of respondents that have missing data on similar sets of values. If the has monotone, stair-case shaped pattern, each variable can be imputed in order, steadily increasing the information in the imputation model. In this case however, figure 2 clearly shows that the data is not monotone, and non-monotone methods will have to be used.

[11] This figure is created in SPSS although Amelia also has the option of creating a missing data pattern.

Figure 2. Missing data pattern



The assumptions of multiple imputation are heavily dependent on the analysis- and imputation models. In order for the assumptions to be met, the researcher needs to have a proper imputation model. For them imputation models to be proper, they need to match up with the analysis models. The first step is therefore to specify the analysis models.

The analysis models

The main goal of this thesis is to see how well imputation works for correcting for unit nonresponse bias. The ideal model for this experiment is therefore a somewhat complex multiple regression model[12] that contains both non-linear associations and interactions between independent variables. But it is also relevant to see how well a simple regression model holds up in comparison, since a simple model will probably be less sensitive to misspecification. I have specified two different regression models, one simple and one

[12] One could argue that an analysis that is more sensitive to tail-behavior is better suited. However, such methods are rarely used in the social sciences, and as such, it is more relevant to look at a real world application of the method.

complex. Both models have the same dependent variable, but the number of independent variables differs. Testing the assumptions of regression analysis has only been done with regards to normality of the dependent variable and non-linearity in the parameters, as the remaining assumptions would require an immense amount of work when working with five different datasets, four of which by themselves contain a total of 120 imputed datasets. The goal of this thesis is not to find results that can be generalized to the population, and such tests would therefore be redundant. Any eventual problem would likely be present in all models, and as such comparing the models would still be valid.

Dependent variable

The dependent variable in both analysis models will be attitudes towards immigration. The variable is a scale made from six variables on attitudes towards immigration. The first three variables are formulated as: *allow few/many immigrants of: Same ethnic group/Different ethnic group/From poorer countries outside Europe.* The other variables are; *immigrants are good/bad for country's economy*, *cultural life is undermined by immigrants*, and *immigrants make country a worse or better place to live.* The first three variables were reversed so that all variables go in the same direction. High values indicate positive attitudes towards immigration. All variables are correlated between 0.3 and 0.8. Cranach's alpha for the scale is 0.800, and its value cannot be increased by deleting any of the items. All six variables are included in the imputation models, and were subsequently combined using their standard scores (z-scores) in the imputed datasets. Standard scores were used to compensate for differences in range between the first three and the second three variables. These standard scores were summed, creating a variable than span from about 0 through 42. The first group of variables span from 1 through 4, while the second group spans from 0 through 10. I created a scale to increase the variability in the variable and at the same time reduce any problems caused by non-normal distributions. The variable is normally distributed (see appendix figure 1).

Independent variables

Some of the variables included in the regression analysis have measurement levels that make it questionable to treat them as continuous variables in a regression. For example, TV watching is continuous only up to the last category (more than 3.5 hours), education only has three categories and could be treated as a set of dummy variables, and attitudes towards homosexuality only has four categories. However, when imputing the data, there are several advantages to treating the variables as continuous. In addition, recoding the variables after the

imputation will lead to a loss of information about the distributions. I have therefore decided to treat these variables as approximately continuous.

The first two models in both regressions will be similar. Both the simple and the complex regression models will include demographic variables; age, gender and centrality of municipality, in model 1. In model 2, I will control for the effect of education in both models. The simple regression will include variables for political satisfaction (how satisfied are you with the government), political trust (trust in parliament) and TV watching (how much TV do you watch on an average day) in model 3.

In model 3, the complex regression will include TV watching. Model 4 will include variables for measuring political trust and political competence. The variables included are *trust in country's parliament*, *satisfaction with government*, and *how often you find politics too complicated to understand* (referred to as *political confidence* from now on). To control for political placement and attitude, a variable for placement on left to right scale is included. Also, a variable measuring attitude towards homosexuality is included. To test for nonlinearity, I have included age squared in model 6. Since men traditionally have more negative attitudes towards both immigration and homosexuality, I have added an interaction between gender and attitude towards gays and lesbians[13]. The model is not largely based on relevant research, but is built for the purpose of testing different multiple imputations. This model was built specifically to include significant non linear relationships and interactions, and to thus be vulnerable to imputation model misspecification. Finding such interactions and non linear relationships was the result of trial and error, not because the variables necessarily are theoretically relevant

The imputation models

As mentioned under the assumptions of multiple imputation, I have created two imputation models, one simple and one complex. This will allow me to test my third hypothesis; *The substantial interpretation of a regression analysis will not differ between a simple and a complex imputation model.*

Simple imputation model

I have suggested that researchers pay close attention to bias in their datasets and that they take the appropriate measures when necessary. However, in practice, some researchers might not be overly familiar with the methods of multiple imputation. The literature on multiple

[13] This interaction was found after some trial and error, not because it is especially relevant theoretically.

imputation can be difficult to get an overview of. This might lead to some researchers performing "sloppy" imputations. Because of this, it is relevant to see how such a "sloppy" model compares to a proper model. To test this, I have made an imputation model that only contains the variables from the analysis models, which is the minimum to satisfy the assumption of a proper model. The number of imputed datasets, *m,* will be very moderate considering the high proportion of missing data (10 imputed datasets).

Complex imputation model

In many cases, survey data contains information about cooperation rates. My hope is that such information, in collaboration with register data, can be used to correct for nonresponse bias. The complex imputation model is built after an extensive exploration of the data. This includes a logistic regression where the outcome is; being a reluctant respondent (as opposed to cooperative). All relevant variables (that is, everything but such variables as "*how many employees does partner have"*, "*what is the age of the fourth person in the household*" etc.) were added in groups, using both the Forward Stepwise and Enter methods in SPSS. Variables that had a very large or a significant effect (or both) were put together in a regression to find a model that could best predict reluctance, and hence best be able to model the missing-mechanism. It is very interesting to see whether the results of a complex imputation are robust in the face of imputation model misspecification, and if such elaborate efforts to specify a correct model pays off. The model will also strictly treat the assumption of normally distributed variables. Variables that were not normally distributed were excluded from the model. This assures that the assumption of normally distributed variables is met. Because of the large proportion of missing data, the complex imputation is run with 50 multiple imputations, which is the maximum number needed according to some authors (Kenward and Carpenter 2007, Horton and Lipsitz 2001).

Statistical software

Multiple imputation was proposed to make analytical research easier for the researcher, but in the twenty years since the approach was first recommended, only a handful of the most advanced researchers have utilized the method. King (2001:65) blames this on the difficulties of using multiple imputation. The new generation of software however makes this substantially easier to do. I will be using both SPSS; a commercial statistical software package, and Amelia II; a free software for performing multiple imputation, to see if there are any differences across the results from different software packages and algorithms.

SPSS

SPSS is a commercial statistical software package. SPSS primarily uses the MCMC method to impute missing values. Before running an imputation in SPSS, there are a couple of important steps that have to be made. First, all variables must be set to their correct measurement level in the variable view. The procedure can only impute quantitative variables (scale or ordinal), not categorical/nominal variables (SPSS Missing Values Manual 2010:4). Nominal variables can however be used as predictors. Ordinal variables that measure an underlying continuous scale *can* be set to scale instead of ordinal. This is helpful in two ways. First, all imputation procedures assume that the variables are continuous, and by choosing the ordinal level, you lose information about the true variance in the data. Secondly, an ordinal variable with say, four categories, will require the specification on four parameters. In contrast, a continuous variable such as age will only require the specification of one parameter, even though it contains much more information. SPSS is by default limited to a model with 100 parameters. Thus, by setting ordinal variables to scale, it becomes easier to use several ordinal predictors. However, this limit can be exceeded by running the imputation in syntax and using the MAXMODELPARAM subcommand to set a higher limit of parameters.

In addition to setting the correct measurement levels, you need to decide how to handle data that is user missing. In many surveys, such as the European Social Survey, it is common to distinguish between different types of user missing, such as refusal to answer, not applicable and don't know the answer. It is a good idea to consider if imputing a value for the given variable is reasonable or not. It is reasonable to impute a value for attitude towards immigrants when a person refuses to answer, as there clearly is an underlying value to be measured. However, if the question is "*what is the age of your spouse or partner"* it does not make sense to impute a value for a person that is not in a relationship and has answered "*not applicable"*. Variables that were only applicable for a small portion of the sample (and thus limited the sample size) were not included in the imputation model.

In cases where the data might be missing completely at random, SPSS has the option of performing Little's MCAR test. The null hypothesis in the test is that the data is missing completely at random. A significant *p*-value for this test indicates that the data is not missing completely at random. In most social survey data, it is very unlikely that the data is missing completely at random. When running the test on the ESS data, I found, not surprisingly that the data is not missing completely at random ($p < 0.000$).

The method of imputation I used in SPSS is a Markov Chain Monte Carlo, with a specified maximum of 10 iterations. SPSS gives the option to limit the variable ranges. This will keep the data uniform and similar to original values, but this might not always be a good idea. If this option is used, SPSS will draw a new value until it draws a value that is within the range. This can lead to an underestimation of the variance. Without imposing limits, SPSS can draw negative values. But even if negative values are impossible in the data, it might still give a better representation of the uncertainty in the data, even if it might yield some nonsensical values, for example a negative value for a variable such income or amount of TV watched. Although such a value is not possible in the observed data, it might still provide a better basis for estimating standard errors. I have therefore not imposed any limits on allowed values.

After performing the imputation, one should check the iteration history. The SPSS Missing Values User Manual (2010) suggests looking for patterns in the values of coefficients and standard errors over the different imputations. Using the chart builder, one can see what values each iteration gave for each variable mean and standard error across all imputations. The convergence charts should look random if the imputation went well. The simple imputation model did not produce any unusual looking patterns. However, with 50 imputations, the complex imputation proved much more difficult to interpret, but all graphs seem appropriately random.

AMELIA

In a field dominated by product names that are mainly cryptic abbreviations, one software stands out. Gary King (Honaker, King and Blackwell 2010) appropriately named his software for handling missing data *Amelia*, after the famous pioneering female airplane pilot Amelia Earhart, who went missing after an attempt to circumnavigate the globe. The motivation behind creating the software seems to be the supposed slowness and computational requirements of other approaches such as MCMC. Amelia uses a much faster EMB-algorithm. This is combination of the traditional EM-algorithm with a bootstrap approach. The main computational problem in a multiple imputation is taking draws from the posterior. To simulate estimation uncertainty, Amelia bootstraps the data for each draw before running the EM-algorithm. The mode of the posterior from the bootstrapped data also provides the imputation with fundamental uncertainty (Honaker, King and Blackwell 2010:4-5). I am using Amelia II, the second version of the software, in this thesis.

Before running the imputation in Amelia, you need to make sure that all missing values are in fact missing from the data. Some software, such as SPSS, has the option of setting certain values, such as 999, to user missing. This allows the user to distinguish between different reasons for missing data, such as *refusing to answer, don't know* or *not applicable*. The problem is that Amelia does not recognize this as missing values, which will lead to some extreme outliers. Therefore, user missing (usually values such as 77, 88 and 99), such as don't know, refuse to answer and not applicable, will have to be recoded into system missing prior to running the imputation. If this is not done, these extreme values will be treated as true observations.

Amelia assumes that the complete data (both observed and unobserved) are multivariate normally distributed. Still, Amelia seems to work just as well as other more complicated methods, even when facing categorical or mixed data (Schafer 1997). Honaker, King and Blackwell (2010) recommend using transformations in the case of non-normally distributed values. As with all multiple imputation techniques, Amelia also assumes that the data is missing at random. As such, it is important to include any predictor related to the missing mechanism, and not just the variables intended for the analysis. This includes interactions and transformations. Even variables that would be problematic because of multicollinearity should be included as long as they increase the predictive power of the model (Honaker, King and Blackwell 2010:10). Ordinal values should as often as possible be allowed to be imputed as fractions and not just integers. This is also true for some dichotomous variables such as gender. Although a value of 0.67 on a dummy variable for male is nonsensical, is does carry more information about the distribution than a forced integer value. Nominal variables with more than two categories, such as county, on the other hand need to be properly specified (Honaker, King and Blackwell 2010).

Variables that are heavily skewed can be transformed using for example the natural logarithm, especially in the case of outliers. In the ESS data, most variables have a defined maximum and minimum value, and outliers are therefore usually not present, unless there is a coding error (or if one forget to recode user missing values). Logical bounds can be placed on the data to ensure that a value does not exceed a given value. However, this poses extreme restrictions on the model that might lead to lower variances than the imputation model implies. Generally, Honaker, King and Blackwell (2010:27) recommend not using bounds. As with the imputations in SPSS, I have opted to not put any bounds on the limits of the imputed values.

Amelia uses all variables in the dataset to impute missing values. Identification variables that are not be used in the imputation must therefore be marked as ID variables in the variable list. This method can also be used to select any variable in the dataset that you do not wish to use in the imputation process. But according to Honaker, King and Blackwell (2010) this is a waste of computer memory and will slow down the imputation process. It is recommended to delete unnecessary variables prior to running the imputation. Both the simple and the complex imputation datasets were reduced to only including the relevant variables before I ran the imputations.

When the data contains a high degree of missing data, very strong correlations between variables, or when the number of observations is small relative to the number of parameters, the choice of imputation model can highly influence your results. The ESS data suffers from at least the first two problems. In addition to the high degree of missing data, there are many variables that are quite collinear, such as the immigration attitude variables, the interaction and the squared age variable. A sign of danger is when the chain lengths for each imputation vary greatly (Honaker, King and Blackwell 2010:22). My imputation models initially had serious trouble converging. Some of my earlier attempts, even with a simple model, were aborted by me when the number of iterations reached more than 30 000 for one single imputation, which ran for hours (the first six imputations only took 20-30 iterations). When these chain lengths differ greatly, it is an indication that the EM algorithm is unstable. However, this problem can be solved quite easily by adding a ridge prior. A ridge prior is equivalent to adding *N* number of artificial observations to the data set that has the same means and variances as the observed data, but with zero covariances. Honaker, King and Blackwell (2010) recommend a starting value of between 0.5 and 1 percent of the number of observations, *n.* If the algorithm stays unstable, the ridge prior can be increased up to an upper bound of about 10 percent of *n.* I used ridge prior of 20, with total *n* being 2673. This solved the convergence problem.

Once the data has been imputed, Amelia II has a diagnostics function that can be used to compare the original and imputed distributions. If the distributions are very strange or very far from the original distribution, one can consider improving the imputation model. The imputed dataset's variable distributions were in accordance with the expected distributions.

Amelia creates *m* number of imputed datasets that have to be combined manually if you want to use another software, for example SPSS, for the analysis. This can be a daunting task

unless you are used to combining datasets using syntax. Amelia runs as an add-on to the free software R. Some knowledge about using R is therefore highly useful when using Amelia. Amelia/R does not have a simple interface for editing data and running statistical analysis. Because of this, most users will have a need for some additional software such as SPSS, SAS or STATA. Therefore, I recommend using SPSS (or another complete statistical software package) instead of Amelia unless you are used to working with R and combining datasets manually.

This leaves us with a total of four different multiple imputation datasets. To separate the two imputation methods, the Markov Chain Monte Carlo imputations will be referred to as MCMC while the Expectation Maximation imputations are referred to as EM. The complex datasets were run with 50 imputations, and will from now on be referred to as the EM50 and MCMC50 datasets. The simple regressions will be referred to as EM10 and MCMC10.

4. Results

After creating the four datasets, analysis were run using the SPSS software. First, let's have a look at the means of the imputed datasets. This will tell us if the imputations were successful in reducing bias. The variables from the population register were clearly biased, as education levels, age, gender, number of persons in household and centrality of municipality were significantly different between respondents and nonrespondents. In the original data, this information is lost due to listwise deletion. But in the imputed datasets, no data is lost. These variables will therefore be unbiased in the imputed data. The more interesting question is whether or not the additional variables have been affected. If the imputation was successful in reducing bias, the imputed means should tend to change towards the direction of the mean in the reluctant respondents. So how did the imputations fare in this respect? Below is a table comparing the means of the imputations to the means in the original data. The means of the reluctant respondents are shown to indicate the expected direction of change (since we assume that the nonrespondents should be closer to the reluctant respondents than to the cooperative respondents). I have also squared the difference from each imputed mean value and the original value (squaring the values assures that there are no negative values). The sum of the squared difference is presented in the bottom row.

Table 8. Comparison of means in different datasets

Variables	Original	Reluctant	MCMC50	EM50	MCMC10	EM10
Immigration	24.68	21.53	24.81	24.53	24.58	24.53
Age	45.37	48.46	46.21	46.21	46.21	46.21
Centrality	2.19	2.28	2.21	2.21	2.21	2.21
Education	2.11	1.86	2.00	2.00	2.00	2.00
TV watching	3.76	4.54	3.76	3.83	3.83	3.83
Politics too complicated	2.89	3.27	2.91	2.95	2.91	2.96
How satisfied government	4.77	5.08	4.78	4.75	4.75	4.75
Trust in country's parliament	5.65	4.78	5.65	5.58	5.59	5.59
Placement on left - right scale	5.25	5.31	5.25	5.08	5.59	5.23
Attitude towards homosexuality	1.95	1.92	1.99	1.81	1.96	1.98
Sum of squared difference			0.73	0.80	0.85	0.75

The mean value of the immigration variable changes in the expected direction in all but the MCMC50 data, where the change is in the opposite direction. However, these changes are truly minute, and overall, the values are very similar. Age is more substantially affected. Since the imputations use the entire dataset, the mean age in the sample increases from 45.37 to 46.21 years. Centrality is also increased, but by a much smaller margin. The mean educational level is reduced from 2.11 to 2.00 in the imputed datasets. These variables are complete variables from the population register, and as such, the means are identical across all imputations (since these values were not imputed). For TV-watching, the MCMC50 data has a very similar value to the original value (3.76), while the remaining three imputations have almost identical values (3.83) that increase in the expected direction (although the values seem identical here, they are different when looking at more decimal places). The political competence variable changes in the expected direction for all imputations. Here, simple and complex MCMC imputations produce very similar means. This is also the case for the complex and simple EM imputations. Satisfaction with government only changes in the expected direction for the MCMC50 data, however this change is very small, and the value is very close to the original data. The other imputations change slightly in the opposite direction. For political trust on the other hand, the MCMC50 data is the only one that does not change in the expected direction. Instead, it remains similar to the original value. For placement on left to right scale, the only imputation that changes in the expected direction is the MCMC10. The MCMC50 remains similar to the mean in the original data while the EM imputations change in the opposite direction. Finally, attitude towards homosexuality only changes in the

expected direction in the EM50 imputation. The other imputations change in the opposite direction. The MCMC50 data is closest to the original mean values (0.73). The most different dataset is the MCMC10 data (0.85), while the EM datasets lie somewhere in between (0.75 and 0.80).

There is no distinct pattern in the differences in imputed values. Some variables are remarkably similar across all imputations (TV watching, satisfaction with government and trust in parliament). Others vary between all imputations. The changes are however very small in most cases, and there is no point in overanalyzing the small differences found. Surprisingly, there is no pattern to suggest that either the complex and the simple regressions differ, or that the EM and MCMC method differs systematically in the estimated mean values. I therefore suspect that the small difference between the imputations are quite random differences caused by the random nature of the process of drawing imputed values. However, the variables from the population register do naturally change quite a bit. As such, the multiple imputations did not necessarily change the mean of the imputed variables substantially, but the inclusion of information otherwise lost due to listwise deletion does make some difference. It seems that bias was only reduced in variables that were complete for all respondents (the population register variables), not in the missing survey variables. What remains to be seen is how well each imputed dataset manages to maintain the covariance structure. Next, we will have a look at the simple and complex regression models and see how the imputations differ in their estimates.

The simple regression model

Table 9 shows the simple regression for all datasets. In the first model, age has a highly significant negative effect on attitude towards immigration. The effect is much smaller in the MCMC50 data (.028) compared to the other datasets (.045-.050). Older people seem to have more negative attitudes towards immigrants. The dummy variable for men does not have a significant effect, but men have a slightly more negative attitude towards immigrants than women. Both MCMC datasets have smaller b-coefficient and standard errors than the other datasets. Centrality of municipality has a strong positive effect in all datasets, showing that people who live in more central areas are more positive to immigration. The MCMC50 data has the lowest coefficients and standard errors.

In model two, I control for the effect of education. The effects of age and gender remain largely unaffected, although the b-coefficients are somewhat lower than in model 1. Centrality

of municipality is affected more heavily, with a reduced significance level in all models from <0.001 to <0.050. The b-coefficients are reduced by about 0.2 for all but the MCMC50 data, which is reduced by about 0.1. Education has a very large and highly significant positive effect. This suggests that while centrality does affect your stance on immigration, some of this effect can be explained by lower education in the less central regions. This is not surprising since a lot of the jobs requiring higher education are located in and around the largest city centers. The MCMC50 data has a much lower b-coefficient than the other data.

Table 9. Comparison of simple regressions

Model		Original data		MCMC 50		EM 50		MCMC 10		EM 10	
1	Constant	25.735	***	25.318	***	25.855	***	25.724	***	25.808	***
		(0.608)		(0.427)		(0.559)		(0.582)		(0.570)	
	Age of respondent	-0.047	***	-0.028	***	-0.049	***	-0.045	***	-0.050	***
		(0.009)		(0.006)		(0.008)		(0.008)		(0.009)	
	Dummy Male	-0.424		-0.316		-0.441		-0.421		-0.418	
		(0.325)		(0.227)		(0.321)		(0.361)		(0.349)	
	Centrality of municipality	0.612	***	0.420	***	0.524	***	0.523	**	0.562	***
		(0.151)		(0.109)		(0.142)		(0.160)		(0.134)	
2	Constant	19.697	***	21.577	***	20.071	***	20.365	***	20.03	***
		(0.759)		(0.539)		(0.735)		(0.756)		(0.746)	
	Age of respondent	-0.04	***	-0.023	***	-0.042	***	-0.038	***	-0.043	***
		(0.009)		(0.006)		(0.008)		(0.008)		(0.008)	
	Dummy Male	-0.336		-0.275		-0.378		-0.362		-0.354	
		(0.312)		(0.222)		(0.312)		(0.355)		(0.341)	
	Centrality of municipality	0.377	**	0.307	**	0.349	*	0.361	*	0.387	**
		(0.146)		(0.107)		(0.139)		(0.160)		(0.128)	
	Level of education	2.953	***	1.876	***	2.9	***	2.687	***	2.897	***
		(0.238)		(0.172)		(0.233)		(0.239)		(0.218)	
3	Constant	15.831	***	17.53	***	15.991	***	16.088	***	15.957	***
		(0.861)		(0.662)		(0.837)		(0.820)		(0.785)	
	Age of respondent	-0.036	***	-0.021	***	-0.037	***	-0.033	***	-0.038	***
		(0.008)		(0.006)		(0.008)		(0.008)		(0.008)	
	Dummy Male	-0.315		-0.254		-0.334		-0.36		-0.352	
		(0.286)		(0.208)		(0.290)		(0.310)		(0.306)	
	Centrality of municipality	0.274	*	0.224	*	0.237	†	0.243	†	0.267	*
		(0.134)		(0.101)		(0.125)		(0.140)		(0.123)	
	Level of education	2.002	***	1.326	***	1.993	***	1.881	***	2.004	***
		(0.228)		(0.167)		(0.227)		(0.225)		(0.206)	
	Satisfaction with government	0.679	***	0.591	***	0.676	***	0.688	***	0.682	***
		(0.082)		(0.063)		(0.079)		(0.103)		(0.084)	

Trust in parliament	0.698	***	0.635	***	0.703	***	0.678	***	0.692	***
	(0.076)		(0.058)		(0.082)		(0.083)		(0.074)	
TV watching	-0.344	***	-0.319	***	-0.317	***	-0.306	**	-0.314	***
	(0.083)		(0.065)		(0.083)		(0.095)		(0.077)	

† significant at p < 0.01, * significant at p < 0.05, ** significant at p < 0.010, ***significant at p < 0.001,

In model 3, variables for satisfaction with the government, trust in parliament and TV watching are included. This further reduces the effect of age by a small margin, but the variable remains highly significant. Gender is not affected by these variables and remains insignificant. The b-coefficients for centrality are reduced by about 0.08 for the MCMC50 data and 0.1 for the other datasets. The EM50 and MCMC10 data no longer have a significant effect because of the larger standard errors and reduced effects in these datasets. Education is heavily affected by these variables, and suffers a reduction in b-coefficient size by 0.9 to 1 for all but the MCMC50 data, which is only reduced by about 0.5. The effect is still substantial and highly significant in all models. Satisfaction and trust both have comparable positive effects on the dependent variable and are highly significant in all datasets. As usual, the MCMC50 data has a slightly smaller effect and standard error than the other datasets. Finally, watching a lot of TV is associated with negative attitudes towards immigration. This time however, the MCMC50 data does not stand out with lower b-coefficients, although the standard error is markedly lower than in the other datasets.

To sum up, there was only one place where the choice of imputation model had an effect in the substantial interpretation of the data; centrality of municipality was not significant in model 3 for the EM50 and MCMC10 datasets. Still, the variables showed a similarly sized b-coefficient, but had too large standard errors to be significant at a 0.05-level, but only just so (both were significant at a 0.10 level). And for the other datasets, the effect was not highly significant, but just barely within the margin, indicating that these differences could be due to the randomness in the imputation process. The simple regression did not give substantially different results between imputations. Centrality was not significant on a 0.05 level in model 3 in all imputation, but all the imputation models had quite similar significance levels. The differences were small, but just large enough for two of the imputation models to pass the threshold. But it is important to remember that the 0.05 significance level is a quite arbitrary cutoff, and one should not rely solely on the significance level in such cases. The interpretation of the results was not substantially different across the datasets.

The complex regression

Table 10 shows the results for all seven models in the complex regression analysis for the five datasets. The first two models only contain independent variables from the population register. Any variation in the results from the first two models is likely to be caused by differences in the imputed values in the dependent variable; *attitudes toward immigration*.

Table 10. Comparison of complex regression

Model	Original data		MCMC 50		EM50		MCMC 10		EM 10	
1 Constant	25.925	***	25.318	***	25.855	***	25.995	***	25.791	***
	(0.616)		(0.427)		(0.559)		(0.607)		(0.578)	
Age of respondent	-0.046	***	-0.028	***	-0.049	***	-0.043	***	-0.05	***
	(0.009)		(0.006)		(0.008)		(0.009)		(0.009)	
Dummy Male	-0.442		-0.316		-0.441		-0.500		-0.427	
	(0.327)		(0.227)		(0.321)		(0.323)		(0.358)	
Centrality of municipality	0.568	***	0.420	***	0.524	***	0.534	***	0.564	***
	(0.152)		(0.109)		(0.142)		(0.150)		(0.135)	
2 Constant	19.968	***	21.577	***	20.071	***	20.109	***	19.99	***
	(0.772)		(0.539)		(0.735)		(0.759)		(0.751)	
Age of respondent	-0.039	***	-0.023	***	-0.042	***	-0.036	***	-0.042	***
	(0.009)		(0.006)		(0.008)		(0.009)		(0.009)	
Dummy Male	-0.346		-0.275		-0.378		-0.418		-0.364	
	(0.314)		(0.222)		(0.312)		(0.310)		(0.350)	
Centrality of municipality	0.347	*	0.307	**	0.349	*	0.320	*	0.388	**
	(0.147)		(0.107)		(0.139)		(0.145)		(0.129)	
Level of education	2.884	***	1.876	***	2.900	***	2.859	***	2.909	***
	(0.240)		(0.172)		(0.233)		(0.236)		(0.218)	
3 Constant	22.028	***	23.358	***	22.045	***	22.063	***	21.933	***
	(0.862)		(0.612)		(0.873)		(0.845)		(0.798)	
Age of respondent	-0.033	***	-0.020	***	-0.036	***	-0.030	***	-0.037	***
	(0.009)		(0.006)		(0.008)		(0.009)		(0.008)	
Dummy Male	-0.369		-0.282		-0.391		-0.435		-0.381	
	(0.312)		(0.221)		(0.311)		(0.308)		(0.341)	
Centrality of municipality	0.328	*	0.291	**	0.331	*	0.300	*	0.367	**
	(0.146)		(0.106)		(0.136)		(0.144)		(0.130)	
Level of education	2.632	***	1.738	***	2.675	***	2.630	***	2.677	***
	(0.243)		(0.174)		(0.236)		(0.239)		(0.212)	
TV watching	-0.472	***	-0.425	***	-0.453	***	-0.455	***	-0.439	***
	(0.091)		(0.068)		(0.091)		(0.089)		(0.077)	
4 Constant	20.609	***	20.375	***	20.527	***	20.672	***	20.616	***
	(1.075)		(0.799)		(1.003)		(1.060)		(1.016)	
Age of respondent	-0.043	***	-0.024	***	-0.045	***	-0.039	***	-0.045	***
	(0.008)		(0.006)		(0.008)		(0.008)		(0.008)	
Dummy Male	-0.745	**	-0.443	*	-0.733	**	-0.802	**	-0.774	*
	(0.290)		(0.209)		(0.283)		(0.286)		(0.312)	
Centrality of municipality	0.197		0.195	†	0.190		0.169		0.209	†

		(0.134)		(0.100)		(0.123)		(0.132)		(0.126)	
	Level of education	1.587	***	1.176	***	1.661	***	1.596	***	1.67	***
		(0.232)		(0.168)		(0.223)		(0.228)		(0.215)	
	TV watching	-0.267	***	-0.283	***	-0.255	**	-0.254	**	-0.246	***
		(0.084)		(0.064)		(0.082)		(0.082)		(0.073)	
	Satisfaction with government	0.708	***	0.605	***	0.713	***	0.712	***	0.713	***
		(0.082)		(0.062)		(0.078)		(0.081)		(0.083)	
	Trust in parliament	0.630	***	0.589	***	0.625	***	0.616	***	0.613	***
		(0.077)		(0.058)		(0.082)		(0.076)		(0.073)	
	Politics complicated	-1.103	***	-0.752	***	-1.091	***	-1.101	***	-1.122	***
		(0.157)		(0.113)		(0.147)		(0.155)		(0.155)	
5	Constant	24.702	***	24.156	***	23.908	***	24.729	***	24.512	***
		(1.125)		(0.871)		(1.075)		(1.108)		(1.151)	
	Age of respondent	-0.025	**	-0.015	**	-0.03	***	-0.022	**	-0.027	**
		(0.008)		(0.006)		(0.008)		(0.008)		(0.009)	
	Dummy Male	-0.254		-0.185		-0.270		-0.310		-0.284	
		(0.286)		(0.205)		(0.276)		(0.282)		(0.307)	
	Centrality of municipality	0.189		0.193	*	0.176		0.158		0.205	†
		(0.130)		(0.099)		(0.120)		(0.128)		(0.123)	
	Level of education	1.380	***	1.063	***	1.581	***	1.397	***	1.486	***
		(0.226)		(0.164)		(0.217)		(0.223)		(0.219)	
	TV watching	-0.294	***	-0.289	***	-0.280	***	-0.277	***	-0.266	***
		(0.081)		(0.064)		(0.080)		(0.080)		(0.071)	
	Satisfaction with government	0.563	***	0.512	***	0.580	***	0.571	***	0.579	***
		(0.082)		(0.062)		(0.081)		(0.081)		(0.088)	
	Trust in parliament	0.666	***	0.597	***	0.655	***	0.651	***	0.644	***
		(0.075)		(0.057)		(0.081)		(0.074)		(0.077)	
	Politics complicated	-0.958	***	-0.693	***	-0.966	***	-0.960	***	-0.980	***
		(0.154)		(0.111)		(0.147)		(0.151)		(0.146)	
	Political placement	-0.435	***	-0.382	***	-0.407	***	-0.432	***	-0.420	***
		(0.071)		(0.058)		(0.068)		(0.070)		(0.076)	
	Homosexuality	-1.135	***	-0.913	***	-1.016	***	-1.140	***	-1.140	***
		(0.155)		(0.117)		(0.152)		(0.153)		(0.191)	
6	Constant	21.529	***	22.096	***	20.603	***	21.540	***	20.479	***
		(1.349)		(1.017)		(1.251)		(1.328)		(1.277)	
	Age of respondent	0.136	***	0.095	***	0.142	***	0.141	***	0.175	***
		(0.039)		(0.028)		(0.034)		(0.039)		(0.033)	
	Dummy Male	-0.333		-0.229		-0.346		-0.388		-0.376	
		(0.285)		(0.205)		(0.276)		(0.281)		(0.305)	
	Centrality of municipality	0.211		0.207	*	0.198	†	0.180		0.233	†
		(0.130)		(0.099)		(0.119)		(0.128)		(0.122)	
	Level of education	1.093	***	0.848	***	1.253	***	1.104	***	1.112	***
		(0.235)		(0.171)		(0.226)		(0.231)		(0.231)	
	TV watching	-0.256	**	-0.276	***	-0.250	**	-0.241	**	-0.223	**
		(0.081)		(0.064)		(0.079)		(0.080)		(0.072)	
	Satisfaction with government	0.567	***	0.516	***	0.584	***	0.573	***	0.587	***

		(0.081)		(0.062)		(0.080)		(0.081)		(0.087)	
	Trust in parliament	0.680	***	0.602	***	(0.668)	***	0.665	***	0.659	***
		(0.075)		(0.057)		(0.080)		(0.074)		(0.079)	
	Politics complicated	-0.936	***	-0.678	***	-0.94	***	-0.935	***	-0.947	***
		(0.153)		(0.111)		(0.145)		(0.150)		(0.145)	
	Political placement	-0.418	***	-0.374	***	-0.393	***	-0.415	***	-0.399	***
		(0.070)		(0.058)		(0.068)		(0.069)		(0.077)	
	Homosexuality	-1.082	***	-0.896	***	-0.969	***	-1.087	***	-1.068	***
		(0.155)		(0.116)		(0.151)		(0.153)		(0.192)	
	Age squared	-0.002	***	-0.001	***	-0.002	***	-0.002	***	-0.002	***
		(0.000)		(0.000)		(0.000)		(0.000)		(0.000)	
7	Constant	20.946	***	21.585	***	20.239	***	20.958	***	20.078	***
		(1.371)		(1.039)		(1.277)		(1.351)		(1.280)	
	Age of respondent	0.135	***	0.094	***	0.142	***	0.140	***	0.175	***
		(0.039)		(0.028)		(0.034)		(0.039)		(0.033)	
	Dummy Male	0.966		0.873	†	0.424		0.881		0.513	
		(0.635)		(0.487)		(0.586)		(0.627)		(0.540)	
	Centrality of municipality	0.214	†	0.207	*	0.199	†	0.184		0.234	†
		(0.130)		(0.099)		(0.119)		(0.128)		(0.122)	
	Level of education	1.107	***	0.858	***	1.261	***	1.116	***	1.120	***
		(0.235)		(0.171)		(0.226)		(0.231)		(0.230)	
	TV watching	-0.264	***	-0.280	***	-0.252	**	-0.247	**	-0.226	**
		(0.081)		(0.064)		(0.079)		(0.080)		(0.072)	
	Satisfaction with government	0.563	***	0.513	***	0.582	***	0.57	***	0.585	***
		(0.081)		(0.062)		(0.080)		(0.081)		(0.087)	
	Trust in parliament	0.674	***	0.599	***	0.667	***	0.659	***	0.657	***
		(0.075)		(0.057)		(0.080)		(0.074)		(0.078)	
	Politics complicated	-0.923	***	-0.669	***	-0.934	***	-0.923	***	-0.941	***
		(0.153)		(0.111)		(0.145)		(0.150)		(0.145)	
	Political placement	-0.417	***	-0.372	***	-0.393	***	-0.414	***	-0.399	***
		(0.070)		(0.058)		(0.068)		(0.069)		(0.077)	
	Homosexuality	-0.727	***	-0.62	***	-0.756	***	-0.740	***	-0.843	***
		(0.219)		(0.161)		(0.213)		(0.216)		(0.229)	
	Age Squared	-0.002	***	-0.001	***	-0.002	***	-0.002	***	-0.002	***
		(0.000)		(0.000)		(0.000)		(0.000)		(0.000)	
	Interacion man sexuality	-0.668	*	-0.553	*	-0.424		-0.653	*	-0.449	†
		(0.292)		(0.222)		(0.269)		(0.288)		(0.255)	

† significant at p < 0.01, * significant at p < 0.05, ** significant at p < 0.010, ***significant at p < 0.001,

Model 1 shows that age has a highly significant negative effect on attitudes towards immigration in all imputations. The EM models yield slightly higher coefficients (-.049 and -.050) than the original data (-.046). The MCMC10 imputation has slightly lower coefficients (-.046) while the MCMC50 has substantially lower (-.028) coefficients than the original data. All but the MCMC50 (.006) imputation have similar standard errors (.008-.009). But the

substantial interpretation of age remains the same in all models. Men have slightly more negative attitudes towards immigration, but the effect is far from being statistically significant in all imputations. The coefficients and standard errors show the same pattern; the MCMC50 imputation stands out with lower values on both, while the other imputations are very similar to each other and to the original data. Centrality of municipality has a positive effect on the dependent variable in all imputations. People in more central areas seem to have a more positive attitude towards immigration than those from less central municipalities. The effect is highly significant in all imputations. The same pattern repeats itself here, with the MCMC50 imputation giving lower coefficients and standard errors than the other data sets.

In model 2, I control for the effect of education. The effect of age is decreased slightly in all imputations, by .005 in the MCMC50 imputation and by between .007 and .008 in the other data sets. The standard errors remain, not surprisingly, unaffected. The effect of gender remains small and non-significant in all imputations, and the MCMC50 imputation still stands out with lower coefficients and standard errors. The effect of centrality is reduced by a fair margin when controlling for education. This indicates that the reason why people in less central municipalities are more negative towards immigration can be explained in part by differences in level of education among city-dwellers and people residing in more rural areas. However, the effect still remains statistically significant in all models. The MCMC50 and MCMC10 models have the lowest coefficients, of .307 and .320, while the original and EM50 data have coefficients of .347 and .349. The EM10 has the highest coefficient; .388. Education has a strong positive and highly significant effect on attitudes towards immigration. Again, the MCMC50 imputation has quite a bit lower coefficients, this time about a third lower than the other datasets (1.8 in the MCMC50 compared to about 2.9 for the other datasets). As usual, the standard errors are more similar across the other datasets, and lower in the MCMC50 data.

In model 3, I am controlling for the amount of TV watched. This is the first variable that was incomplete, as all the preceding variables (save for the dependent variable) were complete variables from the population register. The effect of age continues to drop as I control for TV watching, but only by about .003 for the MCMC50 data and .006 for the other datasets. Still, the effect of age remains highly significant for all models. After controlling for TV watching, the effect of being male is increased slightly in all models, but the effect is still non-significant. The effect of centrality is slightly reduced, but it remains significant at a 0.05 level in all models. The two MCMC imputations continue to have similar coefficients (.291

and .300). The original and EM50 data also have very similar coefficients (.328 and .331), while the EM10 data has by far the highest coefficient (.367). Education loses some of its explanatory power, and is reduced by about .2 to .3 points. As in the other models, the MCMC model stands out by having a lower coefficient and lower standard errors while the other models are remarkably similar, ranging from 2.630 to 2.677. TV watching has a strong negative and highly significant effect on the dependent variable across all datasets. The original data has the strongest effect, -.472, while the MCMC50 data has the lowest effect, -.425. The other imputed datasets have b-coefficients of between -.439 and -.455. The differences in the coefficients of the MCMC50 and the other datasets are not as large in this variable as in the previous variables. The standard errors are similar for the original data, the EM50 and the MCMC10 data (0.089-0.091). The EM10 data has a standard error of 0.077, while the MCMC50 has the lowest value, 0.068.

In model 4, I control for the effects of satisfaction with the government, political trust and political competence (how often you find politics too complicated to understand). This strengthens the effect of age by .004 for the MCMC50 data and .008 to 010 for the other data sets. Remarkably, the effect of the dummy variable for male is greatly increased in the MCMC50 data and nearly doubled in the other datasets. The effect is significant at a 0.05 level in all datasets, but is much smaller in the MCMC50 data (-.443), than in the other datasets (-.773 through -.802). When controlling for differences in attitudes towards the government and political competence, men are significantly less positive towards immigration than women. Men probably have higher mean values on these variables than women, but when these are held constant, men have a more negative attitude towards immigration.

The coefficients of centrality of municipality are greatly reduced, by about one third in all datasets. Centrality is almost significant in the MCMC50 data, with a t-value of 1.95 (the critical value is 1.96 for a 0.05 level of significance). This is the first model where the substantial interpretation between the datasets is somewhat different. Remarkably, the MCMC50 data does not have the lowest coefficient for centrality, but is quite similar (.195 compared to between .190 and .209) to most of the other datasets save for the MCMC10 data, which has a quite a bit lower coefficient (.169). The standard error of the MCMC50 data is lower than in the other datasets. The effect of education is greatly reduced, by about .6 in the MCMC50 data and more than a whole point in the other datasets, but remains highly significant. Higher educated persons thus seem to have higher trust and satisfaction with the government as well as political competence. But when controlling for these attitudes,

education has a reduced, but still very substantial effect on attitudes towards immigration. TV watching remains significant, but the coefficient is reduced by almost half in some models. While TV watching in the MCMC50 data had a bit smaller effect in the previous model, it now has the strongest effect by a fair margin (-.283 compared to between -.267 and -.246). This is the first variable where the MCMC50 model has the strongest coefficient. This suggest that while watching TV does seem to be correlated with negative attitudes towards immigration, the effect is not so great when controlling for other relevant attitudes. Being satisfied with the government predicts a more positive attitude towards immigration. The effect is large and highly significant, though .1 lower in the MCMC50 data. Trust in parliament also increases attitudes towards immigration by a comparably sized coefficient. The coefficient in the MCMC50 data is smaller, as usual, but not by a great margin (.589 compared to between .613 and .630). People who find it difficult to make up their mind about politics are more negative towards immigration. The effect is quite large and highly significant in all datasets, but substantially smaller in the MCMC50 data (.752 compared to 1.091 through 1.122) than in the other datasets.

Model 5 includes attitudes towards homosexuality and placement on left to right scale (where 0 is extreme left and 10 is extreme right). These are supposed to control for right wing conservative attitudes that are likely to be associated with negative attitude towards immigration. When controlling for these values, the effect of age is once again reduced in all datasets. Gender no longer has a significant effect in any dataset and the effect is greatly reduced. This might suggest that men are simply more likely to have a more right wing attitude than women. As such, the differences we saw in the last model were probably not because of gender, but because men are more likely to have more conservative attitudes towards homosexuality and immigration. When controlling for these differences, there is no significant difference between the sexes. Centrality of municipality has a just barely statistically significant effect in the MCMC50 data, but not in any of the other datasets. The effect of centrality is largest in the EM10 data (.205), but the standard error is too high for the effect to be statistically significant. Education has a slight reduction in effect, but remains significant in all models. This suggests that the effect of education cannot be explained by right wing conservative attitudes. This is not very surprising, as there is little reason to expect that people who place themselves to the right of the political spectrum are less educated than more left wing persons. The negative effect of TV watching is not substantially altered by these controls. The coefficients are slightly reduced, but not in a way that suggests any

interaction between TV watching and political stance. The MCMC50 data no longer has the strongest effect, but is just barely beaten by the effect in the original data (-.294 compared to -.289). Satisfaction with the government has a reduced effect in all models by .1 to .2 points. This might indicate that some of the effect of being dissatisfied with the government was actually the effect of belonging to the right wing part of the political spectrum. Since Norway had a left wing government at the time of the survey, it is not surprising that more of the right wingers were unsatisfied. Satisfaction with the government is thus not as important as the previous model initially suggested. As I have come expect, the MCMC50 data has a smaller coefficient and standard error than the other datasets. Trust in parliament on the other hand, has a slightly increased effect in all datasets, but this difference is not as great as to permit further discussion. As with the above variable, MCMC50 has a lower effect and standard error, while the other datasets give very similar numbers. It is noteworthy that satisfaction with government had a larger effect than trust in model 4, while trust has a slightly larger effect in this model (since both variables have the same range, a direct comparison of the coefficients is possible). Political competence is similarly reduced in effect, but still remains highly significant in all datasets, and the same pattern with the MCMC50 having lower values, is present. Placement on the left to right scale has a strong and highly significant effect on the dependent variable. This shows that people who place themselves towards the right wing side of the political spectrum are more negative towards immigration. Once again, the MCMC50 data stands out with lower values than the other datasets. Not surprisingly, being negative towards homosexuality strongly predicts negative attitudes towards immigration. The effect is highly statistically significant. The MCMC50 data has lower coefficients and standard errors, by between .1 and .2 compared to the other datasets, which are largely uniform.

In model 6, age squared is added to the regression. The addition of this variable gives a statistically significant increase in the model's explanatory power in all datasets. The effect of age is dramatically increased, showing that there is a curve linear relationship between age and attitudes towards immigration. Initially, the older the respondent is, the more positive he or she is towards immigration. But at a certain point, this trend turns, and people above a certain age tend to be more negative towards immigration. In other words, the elderly and the young seem to have more negative attitudes towards immigration that the middle aged. Age squared increases the effect of gender, but not to a high degree, and it is nowhere close to significant. The effect of centrality is similarly increased but the interpretation is the same as

in the previous model, with only the MCMC50 data showing a (barely) significant effect. The effect of education on the other hand is reduced quite a bit, by .2 to .3 points. The effect is still large and statistically significant in all models however. TV watching has a reduction of about .04 in all but the MCMC50 data, where the reduction is very small, only .013 points. As in the previous models, this leaves the MCMC50 data with the strongest effect of TV watching. The political control variables are by and large unaffected by the inclusion of age squared. This is also the case for placement on left to right scale. Attitude towards homosexuality is reduced very slightly, but not in a way that suggests any new interpretation. All in all, the inclusion of age squared improves the model, but does not change the interpretation of the variables compared to the previous model.

Finally, in model 7, I have included an interaction between gender and attitudes toward homosexuality. In the original data, this results in a significant improvement of the model. Of the imputed datasets, only the MCMC10 and MCMC50 datasets are significantly improved by including the interaction. For the original and the MCMC datasets, negative attitudes toward homosexuality have a significantly stronger negative effect on attitude toward immigrants among males than among females. The effect is still heavily present among women, but the effect is much larger among males. Age remains largely unaffected by this inclusion. Gender sees a greatly increased effect, but in most models the standard errors are also very large, and the net effect is not close to statistically significant. But for the MCMC50 data, the standard error is quite a bit lower than in the other datasets. The effect however, is high in the MCMC50 (-.896) as well as in the original data (-.966) and the MCMC10 (-.881) data. Both the EM datasets have much smaller coefficients (-.424 and -.513). The t-value for the MCMC50 data is 1.79, and is significant on a 0.10 level. This is an interesting result. All models have to a large degree managed to replicate the results of the original data, but when an interaction is added, the EM datasets seem to show their weakness. None of the other variables were affected by the interaction, and the interpretation remains the same as in the previous model for all the other variables.

So what are the major conclusions? First of all, all imputations have on average, lower standard errors than the original data. This leads to the test statistics being more efficient. But the original data did have efficient estimates to begin with, and the increase in efficiency does not influence the substantial interpretation of the results. I can imagine that such an increase in efficiency would be more valuable in a dataset that had a more limited number of valid cases. In a dataset of this size, efficiency, or rather lack of it, is not that big of a problem.

Among the imputations, the MCMC50 dataset stood out in many respects. Although the b-coefficients were smaller in most cases, there were a few instances where this model was the only one close to significant or statistically significant for some variables, such as centrality in several models and gender in the final model. Despite the low b-coefficients, the variables were equally significant because of the correspondingly low standard errors. But the substantial interpretation was is most cases the same across all the imputed datasets and the original data. There were a few exceptions where the MCMC50 data just barely had a significant effect, or where it was just below the 0.05 threshold, but nevertheless closer than any of the other datasets. Also, the MCMC50 data was the only dataset where the interpretation was different from that of the original data. This was only the case for centrality however. It is also worth mentioning that gender was close to being significant only in the MCMC50 data, which stands out as the one conclusion where the substantial interpretation is different to the other data.

So how does this result fit with my third hypothesis*; The substantial interpretation of a regression analysis will not differ between a simple and a complex imputation model.* In a simple regression model, the complex and simple imputations perform equally well. However, when the regression model is more complex, only the MCMC data manages to maintain a significant interaction. This suggests that the MCMC method is better than the EM-algorithm at maintaining complex structures and relationships between variables. As for whether or not a complex imputation model is better, I would argue that there is some evidence to suggest that this is the case, but the differences in these results are so small that the results can hardly be considered conclusive. At any case, such small differences could quite easily be explained by the randomness in the imputation process. Since the crucial factor here seemed to be imputation *method*, not imputation *model complexity*, my hypothesis is confirmed. Regression is such a robust method of analysis that complexity of imputation hardly matters (for this data at least), as long as the imputation model is proper[14]. It seems that multiple imputations' greatest virtue is being able to use all the information in the data. It also seems that this is the reason why multiple imputations have been shown to reduce bias. The information that is usually lost due to listwise deletion seems to be able to correct for nonresponse bias, but multiple imputation does not seem to be able to predict the correct values for the missing respondents on other variables in such a way as to reduce bias in these variables.

[14] Meaning the imputation models should be at least as complex as the analysis model.

5. Discussion

The main research question in this thesis is; *Can multiple imputation be used to correct for unit-nonresponse bias in survey data that contains only a limited amount of information about the nonrespondents?* Based on the approach I have taken, the answer is a cautious no. My approach was to use information about reluctant respondents to create an imputation model that could accurately model the nonresponse mechanism. Using logistic regression, I found the variables that could best predict being a reluctant respondent. These variables formed the basis of the complex imputation models, while the simple imputation models only included variables from the analysis models. The complexity of the imputation had little effect on the imputed means and subsequent regression analysis, while I found some differences between imputation methods. For the population register variables that were complete for all survey participants, the bias in these variables was reduced in all datasets. For the remaining variables however, the multiple imputation method did not manage to reduce bias substantially. There was evidence to suggest that at least the TV-watching and immigration variables were substantially biased in the original data. Both reluctant respondents and nonrespondents who participated in the follow-up study had values that deviated from the cooperative respondents mean on these variables. But the mean values of the imputed datasets did not change sufficiently to conclude that bias in these variables was reduced. Therefore, the conclusion is that nonresponse bias can be reduced in observed variables that are not included in analysis models due to listwise deletion, but that multiple imputation as an approach to reduced potential bias in variables that are unobserved in missing units does not work. In retrospect, there are several (some of them rather obvious) reasons why this approach failed.

1. Too many questionable assumptions have to be met

In order for this approach to work, several assumptions have to be made. First of all, we have to assume that the nonrespondents are so similar to the reluctant respondents that including variables that predict reluctancy is enough to correct for bias. In the ESS 2006 for Norway, the reluctant respondents seemed to be superficially similar to nonrespondents on the population register variables. When comparing variables found in the follow-up survey, the results were largely mixed and inconclusive, but I nevertheless found two variables that seemed to be clearly affected by nonresponse (attitudes toward immigration and TV-watching). Still, there were too many discrepancies to confidently assume that the reluctant respondents were an appropriate proxy for the nonrespondents in the ESS 2006 data for Norway.

The problem with the lack of homogeneity among the reluctant respondents and the nonrespondents are further exaggerated when considering applying the method to surveys in general. Assuming that reluctant respondents are part of the same group of people as the refusals might be valid in some countries, but not in others. The literature clearly shows that finding common characteristics of nonrespondents across countries is very difficult and the results are largely inconsistent. More research is needed on the applicability of this approach in general. More specifically, research is needed on nonrespondents and their prospective relationship with reluctant respondents. So far, the research on nonrespondents is not encouraging. Researchers seem to find that characteristics of nonrespondents differ across countries, and as such, a generally applicable model is very unlikely. It would be very interesting to find out if nonrespondents within countries differ across different surveys and through time. But this kind of research is not easy to do. It relies on getting information about respondents who do not want to participate in surveys. Refusal conversion strategies are useful for this field of research, but such an approach is both time consuming and costly. Some researchers find that such approaches are not worth the effort. For example, Teitler et al. (2006:135-136) found that for respondents that require a high level of effort, their final inclusion in the data had little impact on sample characteristics. Most importantly, this is because of the small number of cases gained. Secondly, the high-level-effort group closely resembled the moderate-level-effort group. This high level group contained few cases, and thus did not influence the results. When follow-up surveys are available, such as the case is with the European Social Survey 2006 data for Norway, there are still problems. Attitudes change over time. Initial refusals who are persuaded to *participate* in the follow-up might be different to the initial refusals that also *refused* in the follow-up survey. My results show that the means were similar between reluctant respondents and refusals on some variables. The variables where this was most clear were on level of education, age, TV watching and attitudes towards immigrants. On others, the refusals were more similar to the cooperative respondents (political interest, trust in politicians). And in most other cases, the differences in means were so small that they could not be considered significant. For this approach to work, the reluctant respondents and the nonrespondents would have to be remarkably similar to each other while clearly being separate from the cooperative respondents. The degree of similarity probably needed for this approach to work is probably far beyond what can be considered realistic, which brings us to the next point.

2. Multiple imputation is not suited for imputing missing values that are dissimilar to those in the observed data.

Multiple Imputation works so well for item imputation for precisely the same reasons why it *shouldn't* work for unit imputation. Multiple imputation gives missing items a plausible value that doesn't change the posterior distribution, so that information usually lost by listwise deletion can be included in the analysis. The inclusion of observed variables is what corrects for item nonresponse bias in this case, not the imputed missing values. When data is missing on almost all variables for the nonrespondents, there is little information that can correct for bias. Instead, we end up with something similar to a bootstrapping procedure for these variables; we are duplicating the already observed structures and thus increasing the sample size and statistical power, without disrupting the patterns in the data matrix. In retrospect, using multiple imputation to impute values that *differ* from the observed values is contrary to the strengths of the method. In hindsight, this is something I *could* have recognized prior to testing the application, but I think it is fair to say that I would not have come to this understanding without the my experiences with trying it. In cases with so little information about the missing units, this method is not applicable for correcting for bias in unobserved variables. It is possible that information from the follow-up survey would be able to reduce bias more substantially. This would of course be a valuable approach to assessing bias in European Social Survey 2006 data for Norway specifically. But the goal of this thesis was to find a generally applicable approach, not one that is only suitable for the rare cases where follow-up information is available. In addition, there were problems with the reliability of the follow-up survey that would make such an approach questionable (although one could argue that the information from the follow up would be more accurate that no information at all).

So what are there other approaches to handling unit nonresponse other that multiple imputation? Weighting the data would create more correct means of the population register variables in a univariate statistic, but since the covariance structure would be the same as in the original data, results of a regression analysis would be the same as for the imputed and original datasets. What would a valid approach require in terms of the method for estimating plausible imputed values? First of all, we would have to be more certain about the similarities between groups with observed variables, such as the reluctant respondents, and groups with mostly unobserved variables, such as refusals. Even if we found strong similarities, these similarities would likely not be strong enough to clearly separate cooperative respondents from reluctant respondents in such a way as to affect the posterior distribution of the data. But

if there was solid evidence to suggest that reluctant respondents are part of the same group as refusals, one could use the refusals alone as the basis for the posterior distribution on which to draw imputed values for refusals (or other similar groups of nonrespondents). This could be done by separating the reluctant respondents in another data file without the cooperative respondents and running the imputation. In this data however, the rate of missing to observed values would be absurdly large, with 79 more or less complete cases and 735 to 1000 missing units (dependent on how similar other groups on nonrespondents, such as *unable, do not have the time* or *language problem* are to the reluctant respondents). The imputation would be based on the posterior distribution of only the reluctant respondents, and thus be quite different to the cooperative respondents. In a sense, this would be a form of selective bootstrapping, since we would duplicate the posterior distribution observed in reluctant respondents, thereby in a sense increasing the size of the reluctant group to compensate for missing units. But since there are only 79 reluctant respondents, the standard errors would in this particular case become too large, making it difficult to find any significant relationships. Because of this, such an approach would only be valid with a more moderate rate of nonrespondents to reluctant respondents. The response rate in the ESS 2006 for Norway was only 65 per cent. The lower the response rate is, the more likely it is that the data is biased. But the lower the response rate is, the more difficult it is to correct for bias using the multiple imputation approach. This leaves us with a Catch-22 scenario. Low rates of nonresponse could probably be successfully imputed, but with low rates of nonresponse, there is little need for correcting for bias (as the low rates of nonresponse seldom manage to bias the data substantially). The workload of doing 10-50 imputations, and manually combining these with the complete cases from the cooperative respondents, is quite substantial. However, a skilled programmer could write the appropriate macros to make such an approach practical. But this approach suffers from much of the same problems as my approach; the inclusion of several questionable assumptions on the similarities of groups of respondents and the uniformity of the missingness mechanism, as well as a high rate of unobserved to observed values. It is difficult to say if such an approach would be any more useful than simply weighting the sample or using listwise deletion. In any case, the approach could hardly be considered generally applicable, as the assumptions of such an approach working really start to become insurmountable. Proper handling of unit nonresponse remains a difficult problem. As of now, there are no cure-all methods. The best approach might be to utilize a combination of multiple imputation to correct for item nonresponse and weighting to correct for unit nonresponse bias.

This might not prove to be a generally applicable method according to Rubin (1996), but it may nevertheless be the best option available at this moment.

In addition to my main research question, I came across two relevant subquestions after reviewing the literature. My second research question was; *Is there a continuum of resistance or is the classes of nonparticipants model more fitting to explain the missingness mechanism?* My results indicate that the classes of nonparticipants model can better explain the missingness mechanism in the ESS 2006 data for Norway, than the continuum of resistance hypothesis. Previous research shows that the different models seem to fit with different datasets. While the classes of nonparticipants fit better in this case, I have no reason to believe that my results can be applied to other surveys, or even other rounds of the European Social Survey for Norway. Nevertheless, knowing which models fits the data can help create a better model for nonresponse.

My third research question was related to the different methods of multiple imputation; *Do MCMC and EM produce different results in terms of imputed values and subsequent regression analysis results?* Although the results were largely uniform across all datasets, there were subtle differences between the two different imputation methods. In my experience, the proposed advantage of the EM-algorithm, namely that it is supposed to be faster than the MCMC approach, is no longer valid. The computing power available today makes the differences in calculating time negligible. When it comes to the differences in results, the MCMC method outperformed the EM-algorithm when facing a complex regression analysis with interaction variables. Based on this, I would recommend using the MCMC method when available. In the cases where the EM imputations did not produce significant results, this was mostly due to the EM data having larger standard errors than the MCMC data. But I must stress that these differences were marginal, with the MCMC data having just large enough t-values to pass the 0.05 significance test. As such, the interpretation should be quite similar in both cases. Nevertheless, the MCMC data were better at maintaining the interactions between variables. If the analysis includes interactions or non linear relationships, the MCMC approach should be used.

Conclusion and further research

The approach of using multiple imputation to correct for unit nonresponse bias using information about reluctant respondents was not very successful. Looking back, this is no

surprise, as my approach asked multiple imputation to do the opposite of what the method is good at; which is maintaining the observed structures in the data. This does to some degree answer the question of why this approach has received almost no attention in the published literature. Researchers who are knowledgeable about multiple imputation understand why the approach is not feasibly to begin with, and don't have to try it out in practice to realize that it won't work.

I have proposed a different approach to using multiple imputation and information about reluctant respondents to correct for unit nonresponse bias. However, I believe such an approach would suffer from too many of the same problems to be worth pursuing. Therefore, I cannot recommend that more research is spent on this particular approach. If the approach of using reluctant respondents is to be useful, a different method of predicting the values of missing respondents should be used. Once again, if such an approach is akin to a single imputation, this leads to problems with correctly estimating the variance. The evidence from research on nonrespondents demonstrates quite well how inconsistent nonresponse is across countries. Because of this, further research into cross-country similarities seems to be a dead end.

For further research, I would recommend that researchers instead focus on examining how serious unit nonresponse bias can affect results of analysis. For example, one could replace missing values with values from the follow-up survey for the initial nonrespondents before running a multiple imputation. If the results are not affected, that might suggest that robust forms of analysis such as regression are not heavily biased by unit nonresponse and that research based on this data remains valid. A more frightening result would of course be if the results were substantially different. Such a result could raise questions on the validity of research that uses the European Social Survey. But I do not find the latter scenario plausible. Nevertheless, such studies could give valuable insight into how big of a problem unit nonresponse is in terms of the validity of previous and future research.

References

Andrieu, Christophe, Nando de Freitas, Arnaud Doucet and Michael I. Jordan (2001) *An Introduction to MCMC for Machine Learning*. Amsterdam:Kluwer Academic Publishers

Berglund, Patricia A. (2010) *An Introduction to Multiple Imputation of Complex Sample Data using SAS® v9.2.* SAS Global Forum, paper 265 2010. Downloaded from http://support.sas.com/resources/papers/proceedings10/265-2010.pdf 05.07.2010

Babbie, Earl. (2007) *The Practice of Social Research. 11th ed*. Wadsworth, Belmont, CA

Bethlehem, Jelke G. (2002) "Weighting Nonresponse Adjustments Based on Auxiliary Information". In *Survey Nonresponse*, ed. R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little, p. 41-54. New York:Wiley.

Billiet, Jaak, Phillipens, M., Fitzgerald, R. and Stoop, I. (2007). "Estimation of Nonresponse Bias in the European Social Survey: Using Information from Reluctant Respondents". *Journal Of Official Statistics*, vol. 23: 135-162.

Billet, Jaak, Hideko Matsuo, Koen Beullens and Vasja Vehovar (2009) Non-Response Bias in Cross-National Surveys: Designs for Detection and Adjustment in the ESS. In *ASK. Society. Research. Methods. (ASK. Społeczeństwo. Badania. Metody)*, issue: 18 / 2009, pages: 3-43

Curtin, R., Presser, S., and Singer, E. (2000). "The Effects of Response Rate Changes on the Index of Consumer Sentiment". In *Public Opinion Quarterly, 64, 413–428.*

De Leeuw, Edith and Wim de Heer (2002) "Trends in household Survey Nonresponse: A Longitudinal and International Comparison". In *Survey Nonresponse*, ed. R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little, p. 41-54. New York:Wiley.

Dellaert, Frank (2002) *The Expectation Maximation Algorithm.* College of Computing, Georgia Institute of Technology. Technical Report number GIT-GVU-02-20 February 2002. Downloaded from citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.9 05.05.2010

Diaconis, Persi (2008) "The Markov Chain Monte Carlo Revolution". In *Bulletin (new series) of the American Mathematical Society*. Volume 46, Number 2, April 2009, Pages 179–205

Gelman, Andrew and John B. Carlin (2002) "Poststratification and Weighting Adjustments". In *Survey Nonresponse*, ed. R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little, p. 41-54. New York:Wiley.

Groves, Robert M. (2006) "Nonresponse Rates and Nonresponse Bias in Household Surveys". In *Public Opinion Quarterly*, Vol. 70, Number 5. Pp. 646-675

Groves, Robert M., Floyd J. Fowler Jr., Mick P Couper, James M. Lepkowski, Eleanor Singer and Roger Tourangean (2004) *Survey Methodology*. Wiley:Hoboken, New Jersey, USA.

Honaker, James, Gary King and Matthew Blackwell (2010) *Amelia ll: A program for missing data.* Downloaded from http://gking.harvard.edu/amelia/docs/amelia.pdf 12.05.2010

Horton, N. J. and Lipsitz, S. R. (2001), "Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables," *The American Statistician*, 55, 244–254.

Kalton, G. and D. Kasprzyk (1986) "Handling Wave Nonresponse in Panel Surveys". In *Journal of Official Statistics, 2.* Pp 303-314

Kenward and Carpenter (2007) “Multiple Imputaion. Current perspectives”. In *Statistical Methods in Medical Research.2007; 16; 199-218.*

King, Gary, James Honaker, Anne Joseph and Kenneth Scheve (2001) “Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation”. In *American Political Science Review. Vol 95, No. 1, March 2001; 49-69*

Lynn, Peter, Paul Clarke, Jean Martin, and Patrick Sturgis. (2002) “The effects of extended interviewer efforts on nonresponse bias”. In *Survey nonresponse*, edited by Robert M. Groves, Don A. Dillman, J. L. Eltinge, and R. J. A. Little. p. 103-120. New York: Wiley.

Lynn, Peter (2003) “PEDAKSI: Methodology for Collecting Data about Survey Non-Respondents”. In *Quality & Quantity Volume 37, Number 3, 239-261*

Marker, David A, David R. Judkins and Marianne Winglee (2002) “Large-Scale Imputation for Complex Surveys”. In *Survey nonresponse*, edited by Robert M. Groves, Don A. Dillman, J. L. Eltinge, and R. J. A. Little. p. 329-343. New York: Wiley.

Metropolis, Nick (1987) The Beginning of the Monte Carlo Method. In *Los Alamos Science. Special Issue.*

Montaquila, J.M., and Jernigan, R.W. (1997). “Variance estimation in the presence of imputed data”. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 273-277.

Nielsen, Søren Fedor (2003) “Proper and Improper Multiple Imputation”. In *International Statistical Review*, Volume 71, Number 3 (2003), 593-607.

Rässler, Susanne and Rainer Schnell (2003) “*Multiple Imputation for Unit-Nonresponse versus Weighting including a comparison with a Nonresponse Follow-Up Study*”. Friedrich-Alexander-University Erlangen-Nuremberg, Chair of Statistics and Econometrics Discussion Papers, Number 65/2004. Downloaded from http://ideas.repec.org/p/zbw/faucse/652004.html#provider 10.04.2010

Robert, Christian P. and George Casella (2004) *Monte Carlo Statistical Methods*. New York:Springer

Rubin, Donald B. and Little (1987) *Multiple Imputation for nonresponse in surveys.* New York:Wiley

Rubin, Donald B. (1996) “Multiple Imputation After 18+ Years”. *In Journal of the American Statistical Association, Vol 91, No. 434 (June, 1996) pp. 473-489*

Schafer, J.L. (1999) Multiple Imputation: A primer. In *Statistical Methods in Medical Research. 1999; 8: 3-15*

Schafer, J.L. (1997) *Analysis of Incomplete Multuvariate Data.* Great Britain:Chapman & Hall

Stoop, Ineke (2004) “Surveying nonrespondents”. In *Field Methods, Vol 16, No 1, February 2004.*

Stoop, Ineke (2005) *Hunt for the last respondent.* The Hague: Social and Cultural Planning Office of the Netherlands.

Teitler, J.O., R.E. Reichman, and S. Sprachman (2003) “Costs and Benefits of Improving. Response Rates for a Hard to Reach Population”. In *Public Opinion Quarterly 2003, 26, 126–138*.

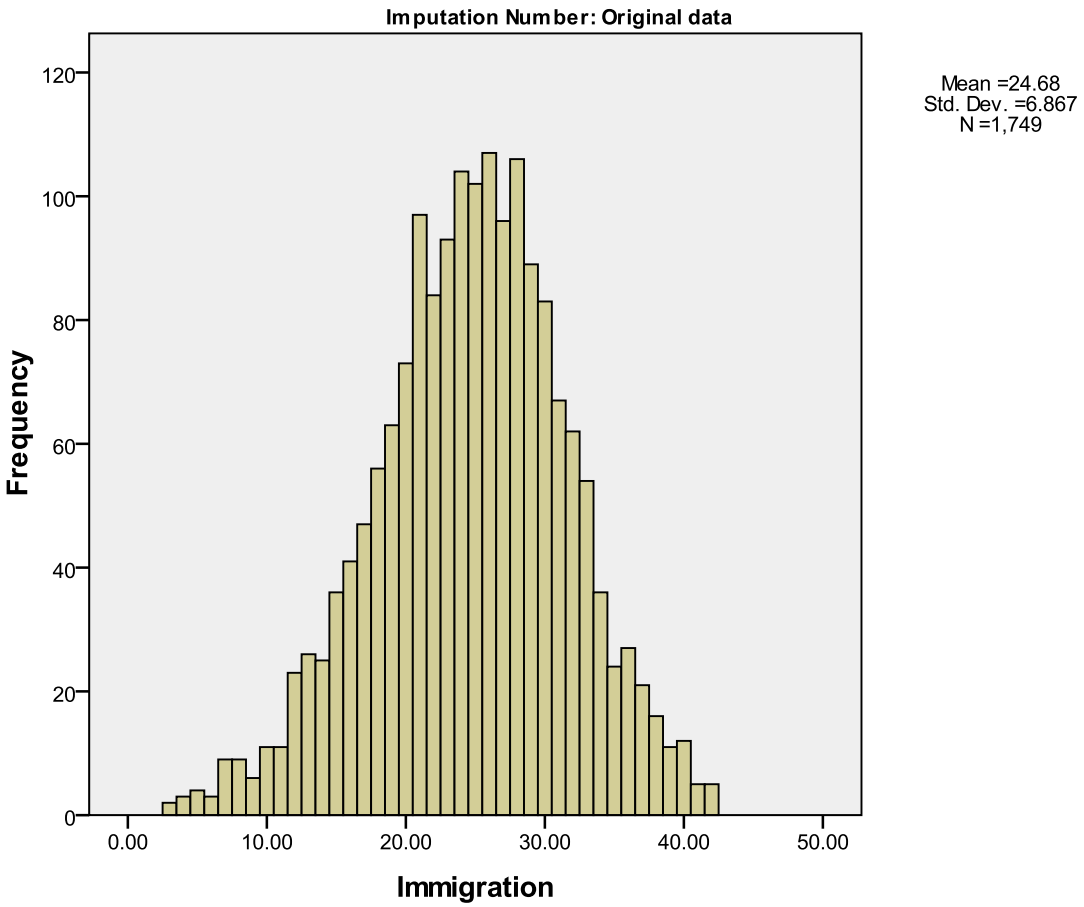
Voogt, Robert (2004) *"I am not interested". Nonresponse bias, response bias and stimulu effects in election research.* Doctoral thesis. Downloaded from http://dare.uva.nl/document/74218 01.04.2010

Vehovar, V. (2007) "Non-response Bias in the European Social Survey". In G. Loosevelt, M. Swyngedouw and B. Cambré (eds.) *Measuring Meaningful Data in Social Research.* Leuven: Acco, pp 335-356

Zhang, Paul (2003) "Multiple Imputation: Theory and Method". In *International Statistical Review / Revue Internationale de Statistique*, Vol. 71, No. 3 (Dec., 2003), pp. 581-592

Appendix

Appendix Table 1. Histogram of the dependent variable



Appendix Table 2. Correlation between original and follow-up survey

Variable	Correlation	
TV watching	0.673	**
Take part in social activities	0.526	**
Feeling of safety of walking alone in local area after dark	0.600	**
Trus in people	0.532	**
Political interest	0.684	**
Satisfaction with democracy	0.544	**
Trust in politicians	0.510	**
Immigrants make country worse or better place to live	0.653	**
Involved in work for charity	0.561	**

** significant at p >0.010.

Appendix table 3. Descriptive statistics of variables

	N	Minimum	Maximum	Mean	Std. Deviation
Immigration scale	1749	0	42	24.68	6.867
Age of respondent	2673	15	101	46.21	18.92
Centrality of municipality	2673	0	3	2.21	1.05
Level of education	2673	1	3	2.00	0.67
TV watching, total time on average weekday	1750	0	7	3.76	1.79
How satisfied with the national government	1733	0	10	4.77	2.02
Trust in country's parliament	1743	0	10	5.65	2.24
Politics too complicated to understand	1747	1	5	2.91	0.99
Placement on left right scale	1702	0	10	5.25	2.04
Gays and lesbians free to live life as they whish	1745	1	5	1.96	0.96