Nr. 73
Materialien aus der Bildungsforschung

Sigrid Wehner

Exploring Trends and Patterns of Nonresponse: Evidence from the German Life History Study

Berlin 2002

Max-Planck-Institut für Bildungsforschung
Max Planck Institute for Human Development

TABLE OF CONTENTS

List of Tables and Figures

Acknowledgements

1. Introduction to the Nonresponse Problem 4
1.1 The Ideal World: Mathematical Theory 4
1.2 The Practical World: Typology of Errors in Gathering Survey Data 6
1.3 The Dark Chapter: The Nonrespondents 11
1.4 Consequences of the Nonresponse Problem 12
2. Exploration of the Nonrespondents in the East German Life History Study 16
2.1 The Sample Development 17
2.1.1 The First Wave 21
2.1.2 The Second Wave 21
2.1.3 The Nonresponse Study 22
2.2 Reasons for Sample Drop-Out 24
2.2.1 The Cohort Composition of the NRS File 24
2.2.2 Gender Proportions 32
2.2.3 The Interview Method 34
2.3 Exploring the Nonresponse Data 37
2.3.1 Strategy for the Comparison of Nonresponse and EGLHS Main Study 37
2.3.2 Finding out Discriminating Variables 40
2.3.3 Summary of the Nonresponse Data Exploration 44
3. Predicting Nonresponse 54
3.1 Predicting Nonresponse for Cohort 1930 55
3.2 Predicting Nonresponse for Cohort 1960 60
4. Estimating Bias in Multivariate Relationships 66
4.1 Weighting Adjustments 66
4.2 The Heckman Sample Selection Model 67
4.3 Application of Correction: An Example With an Intentionally Biased Sample 68
4.4 The Heckman Correction in a Regression Model for Women of Cohort 1960 74
5. Final Discussion 79

References

Appendix A: Codebook for Used File <<nrdata>>
B: Summary Graphics per Cohort with Nonresponse Deviations
C: Formula

LIST OF TABLES AND FIGURES

Tables

TABLE 1: Sample Development in the East German Life History Study 20
TABLE 2: Reasons for Loss of Target Persons per Cohort - Initial Sampling 1991/92 26
TABLE 3: Reasons for Loss of Target Persons per Cohort - Wave 2 1996/97 28
TABLE 4: Reasons for Loss of Target Persons per Cohort - Nonresponse Study 1996/97 30
TABLE 5: Topics and Variables for NRS Data Examination 37
TABLE 6: Working Hypotheses on Profile of Nonresponse Study 40
TABLE 7: Inspection of Significance Patterns - Chi-Square and t-Tests 41
TABLE 8: Logit Models Predicting Being in the Nonresponse Study for Cohort 1930 56
TABLE 9: Predicted Probabilities for Nonresponse in Cohort 1930 / Initial Sampling 1991 58
TABLE 10: Predicted Probabilities for Nonresponse in Cohort 1930 / Wave 2 1996 59
TABLE 11: Logit Models Predicting Being in the Nonresponse Study for Cohort 1960 61
TABLE 12: Predicted Probabilities for Nonresponse in Cohort 1960 / Wave 1 1991 62
TABLE 13: Predicted Probabilities for Nonresponse in Cohort 1960 / Wave 2 1996 63
TABLE 14: Logit-Model to Predict Selection for the Biased Sample 70
TABLE 15: Linear Regression Model for Household Income 1996 / Three Different Calculations 71
TABLE 16: Logit-Model to Predict Panel Participation against Nonresponse Study 75
TABLE 17: Heckman Sample Selection Model Using Panel and Nonresponse Study 76

Figures

FIGURE 1: The Sampling Process and Possible Errors 6
FIGURE 2: Trend of Refusal Rates in German Surveys 11
FIGURE 3: Example of Nonresponse Bias in Linear Regression 13
FIGURE 4: Sample Development in the East German Life History Study 23
FIGURE 5: Cohort Composition of Wave 1 and 2 and Nonresponse Study 24
FIGURE 6: Percentage of Successful Interviews after Low and High Number of Contacts 27
FIGURE 7: Reasons for Loss of Target Persons before Field Beginning 31
FIGURE 8: Reason for Systematic Loss after Field Beginning - Comparison of Cohorts per Study 31
FIGURE 9: Gender Proportions 32
FIGURE 10: Gender Proportions per Cohort and Study 33
FIGURE 11: Cohorts 1930 and 1960: Percentage of Men in EGLHS and Census Data 33
FIGURE 12: CAPI/CATI Differences for Percentages of Neutral Loss per Cohort and Study 34
FIGURE 13: CAPI/CATI Differences of Refusal Rates per Cohort and Study 35
FIGURE 14: Percentage of Higher School Degrees in CATI and CAPI Interviews 36
FIGURE 15: Percentage of Unmarried Persons 1991 per Cohort and Gender 44
FIGURE 16: Percentage of Never Married Persons 1991 per Cohort and Gender 45
FIGURE 17: Educational Degree in School for Cohort 1930 - Nonresponse versus Main Study 46
FIGURE 18: Level of Vocational Training for Cohort 1930 47
FIGURE 19: Average Number of Jobs until 1991 per Cohort and Gender 48
FIGURE 20: Average Number of Jobs 1989-1996 per Cohort and Gender 49
FIGURE 21: Percentage of Unemployed Persons 1991 per Cohort and Gender 49
FIGURE 22: Party Membership in the SED per Cohort and Gender 50
FIGURE 23: Percentage of Persons Not Members of Any Party 1989 - Cohorts 1930 and 1960 51
FIGURE 24: Percentage of "Don't know" and "Refused" in the Party Attitude Scales 52
FIGURE 25: Histogram of Household Income in the Original and in the Biased Sample 69
FIGURE 26: Position of Estimated Regression Coefficients within Confidence Intervals 73
FIGURE 27: Bias Correction of Coefficient Estimate for Educational Variable 74
FIGURE 28: Pattern of Regression Results in Panel Study, Heckman Model, and Panel + Nonresponse 77

Acknowledgements

The following text is my research project report for the M.A. in Social Science Data Analysis at the University of Essex. I want to thank several persons for their support.
In the first place, I have to name two persons who mainly motivated my work.
Leonie Huddy was a stimulating instructor in sampling methods at the Essex summer school. I have greatly profited by the participation in her course and by the opportunity to work as her teaching assistant. I would like to express my sincere thanks for her encouragement and for her theoretical guidance through this research project.
I am also especially grateful to Karl Ulrich Mayer, the director of the German Life History project at the Max-Planck-Institute for Human Development in Berlin, for his permission to use the data of the East German Life History Study and for all the support and resources he made available to me.

I would also like to thank Heike Solga and Anne Goedicke from the Max-Planck Institute. They spent their time and their knowledge about the East German society and were open to discuss many sociological interpretations.
In addition, I want to express my thanks to Elinor Scarbrough for guiding me through my studying years at the Essex summer school and to Janet Brightmore for managing any kind of questions in the summer school's office.
I am also grateful to Kathryn Townend for reading the text and improving my English language.
Finally, I want to thank my husband Hans Wehner for helping me to keep my patience.
I have to add, that any remaining errors are my own responsibility.

Sigrid Wehner
Humboldt University Berlin
April 2001

1. Introduction to the Nonresponse Problem

This thesis concerns itself with the question to what extent do surveys suffer from nonresponse and which strategies are suitable for necessary corrections of nonresponse bias.

As an introduction to this question, the first chapter discusses the conflicting principles between the imaginary idealised sample (as assumed in the mathematical theory of probability) and the real world of survey research with the many possible sources of errors and focuses upon the nonresponse problem. The second chapter examines a nonresponse follow-up study which was conducted as part of the East German Life History Study. The exploration of the nonrespondents has two aims: firstly, to obtain knowledge about the data and to collect robust descriptions and secondly, to isolate variables as candidates in order to explain nonresponse. On the basis of the exploratory chapter, hypotheses about nonresponse behaviour shall be formulated. The third chapter presents a model to predict nonresponse. The fourth chapter discusses possible corrections for nonresponse in multivariate relationships: weighting strategies and the Heckman sample selection model which will be illustrated by an example. The final chapter gives a brief review of the findings.

1.1 The Ideal World: Mathematical Theory

In the social sciences, surveys are used to generalise information obtained from a concrete collection of observations which build <u>the sample</u> to an abstract <u>total population</u>. Kish (1965) calls it *"a hypothetical infinite set of elements generated by a theoretical model"*. The mathematical theory of probability forms the basis of our statistical inferences. It was formulated by Kolmogorov[1] in 1933. He transferred the intuitive idea of uncertainty about what will happen to the set theoretic axioms of probability theory. Kolmogorov constructed the concept of an <u>event</u> as a possible result of an experiment and the concept of <u>probability as a measure.</u> Guided by this approach to probability, two theorems are important for understanding the idea of sampling: the law of large numbers (there are a weak law and a strong law) and the central limit theorem[2]. I will not cite them as formularies as e.g. given and proven by Feller (1950). To understand their implications on the concept of sampling, I will concentrate on the following simplified summary :

---

[1] Kolmogorov, A. N. (1956). Foundations of the Theory of Probability. Second English Edition. See Axioms, pp.2,14-16.

[2] The correct mathematical formulations as well as the proofs can be found in: Feller, William (1950). Probability Theory and Its Applications, Volume I. pp.155-156,191-196. See also Kolmogorov (1956:61-64,66-68).

The Strong Law of Large Numbers

In a random sample, the observed average of a variable tends to the true population mean by growing sample size.

This is the basis for taking the value of a realised sample as an estimate for the true value in the population.

The Central Limit Theorem

Given infinitely many random samples, the obtained sample averages group around the unknown population mean following a normal distribution.

This is the underlying mathematical theory when usual confidence intervals are calculated to obtain a range for the discrepancy between "true" and "observed" values.

Obviously, both of the above mentioned theorems use the abstract idea of an ideal random sample, although it is more than that: of infinitely often repeated experiments. Mathematicians do not have problems with abstract assumptions. For each proposition they write them down and thus stay in their ideal world. Their main scientific exercise is to prove that the proposition, the theorem, or the law follows the rules of logic - based on the assumptions!

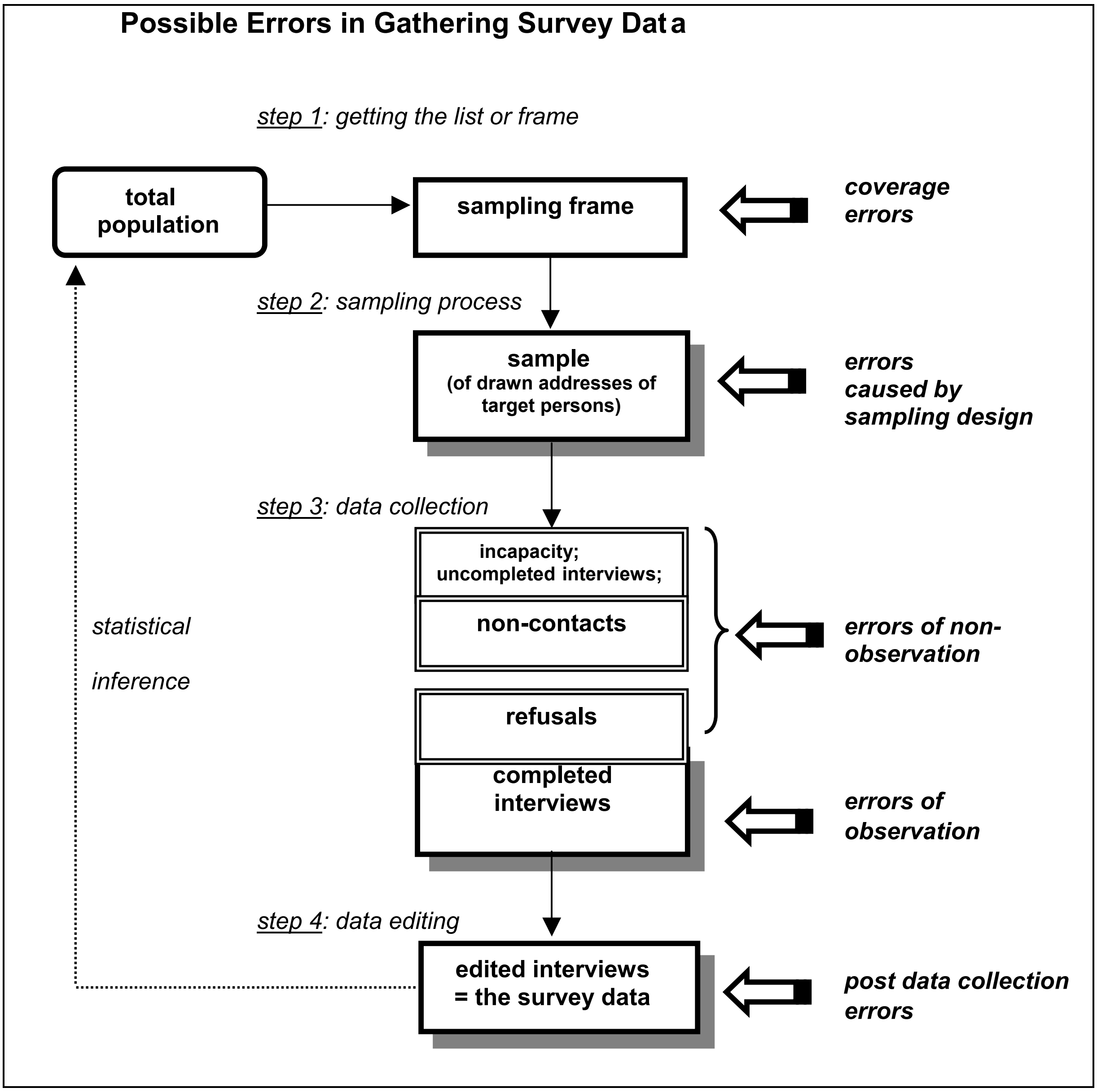
The theory does not say anything, however, about whether these assumptions are fulfilled in a real world application or not. The social scientist who uses statistics is confronted with a number of deviations from the ideal. Feller (1950) states in the introductory chapter of his book about probability theory: *"The history of probability (and of mathematics in general) shows a stimulating interplay of theory and applications, and each new application creates new theoretical problems and influences the direction of research."*

We can never establish a survey with infinitely many repetitions. Taking a sample from a concrete population is not as easy as flipping a coin several times. We cannot avoid a lot of method-immanent errors when collecting concrete survey data. Although I will focus on the problem of nonresponse, it is necessary to see it in the context of the whole sequence of research steps. In the next step, I will, therefore, present a definition of different error types in the sampling process.

1.2 The Practical World: Typology of Errors in Gathering Survey Data

FIGURE 1 below presents four steps in the procedure from the total target population to a completed survey data set. It is based on the terminology given by Kish[3] (1965) and Groves[4] (1989).

FIGURE 1: The Sampling Process and Possible Errors



Whereas Kish as well as Groves both work out the terminology concept as a taxonomy, the graph demonstrates the character of a process through which one has to go in order to obtain the survey data.

[3] Kish, Leslie (1965). Survey Sampling. Chapter 1, Introduction. see: "A Taxonomy of Survey Units and Concepts", pp. 6-8.
[4] Groves, Robert M. (1989). Survey Errors and Survey Costs. Chapter 1, pp. 6-47.

According to the mathematical theory of probability discussed above, only random samples are assumed: a random sample of N persons is regarded as the realisation of N independently repeated observations[5]. This independence[6] of one target person's information from another's is the basic requirement for our statistical theory. Jean Converse (1987) gives a broad historical overview on the development of sampling in the United States and points out the debate on quota versus probability sample. A historical example for the superiority of random over quota sampling is given by the American presidential election polls in 1948[7]. The pollsters failed to predict Truman's re-election using quota techniques whereas a national probability sample supplied the right forecast. Nowadays, scientific survey research in general accepts and applies the probability concept.

I will now explain the several steps shown in FIGURE 1.

Step 1:

At the beginning of each intended survey, the researcher has to formulate the fundamental research question and the corresponding definition of the target population. As a consequence, it is necessary to first find some kind of physical list of all possible sample candidates. This is called the sampling frame. The question of whether frames are fundamentally accurate depends upon many national conditions e.g. whether a survey is planned in a developing country or in an industrial nation. In countries with elaborated national census practice (USA) or with laws regarding the registration of the inhabitants' place of residence (Germany), we assume that the error due to out-of-date lists will be relatively small compared to other potential disturbances. Problems arising out of incorrect lists are discussed by Kalton[8] (1983: 56-57), Kish (1965: 53:59), and Verma[9] (1998: 5.2-5.9). Besides questions such as:

- How old is the last update of the frame?
- Do elements of the target population appear more than once? (duplication)
- Are there persons in the list who do not belong to the target population? (overcoverage)

---

[5] In mathematical context, this is usually notified as "i.i.d." random variables and means independent and identically distributed observations. Whereas e.g. the central limit theorem in the Lindeberg-Feller version does not necessarily demand identical distributions, the assumption of independence of observations is usually essential.

[6] The concept of mutual independence see in: Kolmogorov (1956:8-12).

[7] See: Jean Converse (1987) Survey Research in the United States. The 1948 presidential election forecast is mentioned on pp. 360-361, the quota sampling debate also on pp. 202-211. A general historical overview of the development of survey research (not only in sampling techniques, but also in knowledge of e.g. scaling and measurement) see in chapters 1-4. Gallup's failure in 1948 is also mentioned in the German book about methods of empirical research in the social sciences by Schnell/Hill/Esser (1995:40).

[8] Kalton, Graham (1983). Introduction to Survey Sampling.

[9] Verma, Vijay (1998). Sampling Methods. Course Notes Part 1 and 2, Essex Summer School 1998.

we also have to worry about whether a serious proportion of persons do not appear in the list. This might be the case for persons who do not live in private households (like inhabitants of hospitals or prisons) or who do not have official or permanent addresses (for instance homeless people). So in step 1 we are confronted with possible coverage errors, particularly the latter error of undercoverage. The noncoverage rate can sometimes be assessed. For example, in telephone surveys we could estimate it by looking up statistics on households with a telephone. There are also cases in which the amount of undercoverage error is not easy to measure, however, for example for the rate of homeless people or for persons residing in a country illegally. Other external information or comparative studies would be needed to do this. Kish (1965:531) writes that coverage errors are seldom reported. He suspects that there is a noncoverage rate[10] of more than 10% for national samples.

Step 2:

Once a sampling frame has been accepted, one has to decide on a sampling design. As it is very often impossible to draw a simple random sample out of a complete list, a lot of strategies such as stratification, one-stage and multi-stage clustering or mixed designs have been developed. I will not discuss the different possibilities in detail because this topic constitutes a large separate field of knowledge. Explications of all the techniques are given by Kish[11] (1965, esp. chapters 3-6,10), Kalton (1983:8-56), and Verma (1998:1.1-1.40, 3.1-3.20, 4.1-4.18). Each design has advantages and disadvantages and a compromise between data accuracy and survey costs has to be found. The aspects of costs are demonstrated by Groves (1989:49-80). Contrary to the situation in step 1, the amount of errors caused by the sampling design can be well assessed. There are formulas to calculate the so-called design effect[12]. Examples are given by Groves (1989:265-267, 271-279) and practical exercises by Verma (1998:1.14-1.21, 1.28-1.35). For this step, it has to be emphasised, that the sampling errors are in fact a known quantity[13]. The magnitude can be predicted by our knowledge of the sampling design and probability theory.

---

[10] One should keep in mind that since Kish's book in 1965, social changes in a society might have occured. An increase in social problem groups (e.g. homeless or criminal persons) could produce even higher noncoverage rates.
[11] The complete book consists of more than 600 (!) pages dealing with sampling techniques.
[12] Design effect is named "Deff", Kish (1965).
[13] For example, we can calculate weighting corrections for stratified and cluster samples. See: Kish (1965:77-82, 148-166) or Kalton (1983:69-81).

Step 3:

When the list of the drawn sample (this may be a set of addresses or telephone numbers) is completed, the fieldwork begins. Interviewers start to contact target persons[14] and collect interviews. This data collection step inherits two different disturbing influences: errors of observation and errors of non-observation.

The errors of observation are caused by circumstances and instruments of the data collection itself. This might be

- the situation and the place where a person is interviewed: e.g. at home or on the street?
- the presence of third persons: e.g. could other family members influence the target person?
- the interviewer's personality as well as the interviewer's training with the questionnaire[15]
- the interview method: e.g. personal interview or telephone interview or is it a mail questionnaire?
- wording of the questions
- used categories and scales for the questions: e.g. to reduce non-attitude or acquiescence problems.

Of course this type of error can never be excluded perfectly, because it has arisen out of the research situation itself. It is possible, however, to refer to collected knowledge[16] and research in this field to ensure that as much care as possible is taken when constructing an empirical project.

The **errors of non-observation** - on which this thesis concentrates - are caused by persons who could not be interviewed successfully. This group consists of 3 types of missing people:

1) Incapacity and other reasons

   Sometimes target persons are ill, temporarily in hospital or incapable of answering questions for other reasons (e.g. psychological or mental). This group[17] is usually

[14]Although surveys on households or firms do not consider a person as the relevant target element, I use the word "person" for the final sampling unit because in the following chapters I use the data of the East German Life History Study which is based on persons.

[15] With the German life history data I refer to standardised questionnaires and interviews. Of course several of the mentioned problems are also important in qualitative interviews, however.

[16] Errors of observation are discussed in: Schnell/Hill/Esser (1995) Methoden der empirischen Sozialforschung, pp.297-351. Groves (1989) gives detailed chapters about: the role of the interviewer's behaviour in personal and CATI interviews (see pp. 360-406); the role of the respondent's personality, esp. non-attitudes, recall problems, social desirability effetcs (see pp. 407-447); and influence of the questionnaire itself by wording, order of the questions etc. (see pp. 449-484). See also Converse (1987) pp. 54-104,137-144,189-201.

[17] I also count here if: a) started interviews could not be completed substantially for any of the above mentioned reasons, b) persons finally could not be found etc.; see Kish (1965:534).

expected to be small. If the research question itself is not related to special topics of incapacity (for instance, a health study on people in old age), there is no great cause for concern.

2) Non-contacts

A greater influence on the representativeness of a survey is produced by persons who could never be contacted although several attempts were made and the addresses were checked. Some people never seem to be at home when the interviewer appears or when the telephone rings. Kish calls them the "not-at-homes". We have to think carefully about whether this loss of target persons occurs randomly or whether these non-contacts share special characteristics and, therefore, differ from the rest of the sample.

3) Refusals

The third group of missing people are those who were contacted successfully, but who refused (for whatever reasons) to give interviews. This is an increasing problem for many surveys.

All these types of missing persons are together called the **nonrespondents**. It is easy to calculate the amount of nonresponse in terms of the nonresponse rate. It is given as the percentage of the realised interviews out of the total set of sample elements. Kish (1965:532-535) recommends keeping an account of the attempts to contact and to tabulate the source of nonresponse. He regards it as being necessary for a prediction in future surveys and argues that *"...reporting the extent of nonresponse has become an accepted responsibility for better surveys."*

There is, however, neither an easy way to assess the magnitude of the consequences caused by nonresponse nor is there an accurate solution for correction (as was mentioned above for corrections due to the sampling design). The situation is more tractable in panel studies if a correction for panel attrition is necessary. Since we have existing data for the panel drop-outs from former waves, possibilities exist to describe the panel persons who refused or who were unable to be contacted. On the basis of this knowledge, weighting corrections for panel mortality can be calculated.

The extremely difficult problem is the initial nonresponse, because, besides the information from the sampling list, there is no data about the nonrespondents.

Step 4:

When the field work is over, data editing forms the last stage in the process which results in the survey data set. Data consistency checks are done, obvious coding errors have to be corrected, coding of possible open questions will be performed and finally the codebooks will be produced. These last possible error influences are called <u>post data collection errors</u>. Although they can never be totally eliminated, careful data editing can contribute a great deal to high data quality. Computer-assisted interview techniques (CATI[18] and CAPI) combine the data collection and the data editing stage. Integrated coding and filtering procedures reduce possible errors.

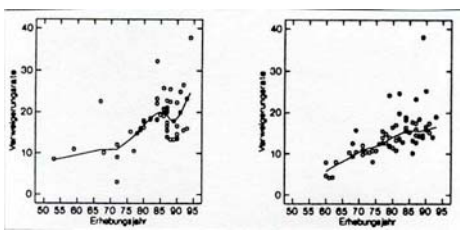
1.3 The Dark Chapter: The Nonrespondents

After the general description of possible errors in surveys, I will now focus on the <u>problem of nonresponse.</u> There are two main reasons why this requires attention.

The first reason is provided by the <u>total amount of nonresponse</u>. It is not marginal and is a problem if the nonrespondents are a special selection. Brehm (1993:16-19) reports nonresponse rates of approximately 33% for academic organisations and refusal rates (as a special subtype of nonresponse) of about 50% for commercial organisations. Trends of increasing nonresponse rates during the years 1960 to 1990 are mentioned. Schnell (1997) examined survey data and the corresponding field reports available at the Central Archive for Empirical Social Research, University of Cologne. His book presents a broad overview of the development of nonresponse rates between 1954 and 1994. For 1990-94, he reports refusal rates of between 12% and nearly 40% for academic surveys. The following figure is a citation out of Schnell's book and shows the refusal rates of German surveys until 1995, separately for academic and commercial surveys.

FIGURE 2: Trend of Refusal Rates in German Surveys

academic commercial



The graph is cited from:
***Schnell, Rainer (1997)**. Nonresponse in Bevölkerungsumfragen, p.87.*

(y-axis: refusal rate; x-axis: year of survey)

We see a trend of increasing refusal rates which means less cooperation of sample persons. Also Groves/Couper (1998:166) mention a declining trend[19] for survey cooperation in the USA or *"... at least it is becoming harder to maintain the same level of cooperation..."*.

---

[18] CATI=<u>c</u>omputer <u>a</u>ssisted <u>t</u>elephone <u>i</u>nterview; CAPI=<u>c</u>omputer <u>a</u>ssisted <u>p</u>ersonal <u>i</u>nterview

[19] This is a very general summary. The authors discuss in more detail that response rates depend on many circumstances e.g.: long running survey or not, census data or not, national differences, survey organisations etc. See Groves/Couper (1998:157-173).

The second reason is the uncertainty about the consequences of nonresponse. There is generally practically no information about the typology of the sampled but not interviewed persons. Brehm[20] (1993) calls them *"the phantom respondents"* (which is also the title of his book). Naming the problem is trivial: we do not know much about these nonrespondents because they did not answer our survey questions. Assessing the impact of this loss of information is anything but trivial, however. Groves/Couper[21] (1998:49) argue that the *"biggest drawback in attempting to study nonresponse"* is the fact that the people we are interested in are exactly the nonrespondents. Above all the question we are interested in is: Do the nonrespondents really differ from the respondents? If they appear quite similar, then the problem is negligible. Often we do not know this, however. I, therefore, call the nonresponse problem "the dark chapter".

1.4 Consequences of the Nonresponse Problem

There are two important consequences of considering the impact of nonresponse:

1) Bias in univariate statistics

Point estimates as means or proportions calculated on the basis of the respondents might be over- or underestimated. The bias problem could be ignored if the total nonresponse rate is low and negligible and the nonrespondents share the same characteristics as the participants (i.e. one can assume them to be missing at random). The bias problem is larger if one of these conditions is destroyed. The nonresponse bias is at its highest if we have both: a high nonresponse rate and nonrespondents differing from survey people. Groves/Couper (1998) demonstrate several examples of bias[22] of means. Brehm (1993:93-96) explains the bias contribution in the formula for the population mean.

Estimates of variance based on the respondents' sample will generally underestimate the population variance which also means that statistical inferences might be incorrect. The mathematical formulas are given by Brehm (1993:97-100).

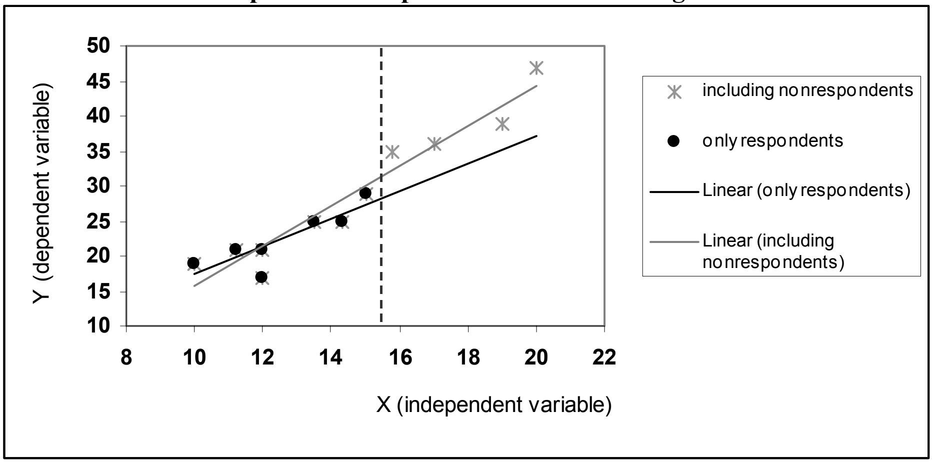
2) Bias in multivariate relationships

In multivariate relationships, there is a danger of biased regression coefficients and underestimation of confidence intervals. The following graph shows a simple regression

[20] Brehm, John (1993) The Phantom Respondents: Opinion Surveys and Political Representation.
[21] Groves, R.M./Couper,M.P. (1998) Nonresponse in Household Interview Surveys.
[22] Lower and higher bias situations are visualised in Groves/Couper (1998), pp.4-5.

model with only one independent variable X.

FIGURE 3: Example of Nonresponse Bias in Linear Regression



------------ marks the region where the sample is truncated

The example demonstrates what happens if the nonrespondents are a systematic selection. In our case, the observations at the high end of the scale are assumed to be the missing ones. The black regression line is estimated after taking the respondents' observations only, the grey line is the true regression, taking respondents and nonrespondents together. We see that the estimated slope in the observed sample is too small. Brehm (1993:100-101) shows that there is also a danger of complete misspecification of the multivariate model. He gives an example of a linear relationship in the observed sample, which is nonlinear, when adding the nonrespondents' information.

This gives rise to the question as to whether nonresponse corrections are possible. Whereas the correction of item nonresponse (e.g. missing values for income) can be done by sophisticated imputation[23] techniques; whereas for the correction of partial unit nonresponse (as in the case of panel mortality) we have at least some initial variables for the later drop-outs, the correction of unit nonresponse in the initial stage of a survey is more complicated, however. Some ideas are, therefore, required regarding what kind of people are not part of the data. Brehm (1993:20) writes: *"... if we know the reasons why people choose or refuse to participate in surveys, we may be able to reduce or correct nonresponse"*.

[23] The theoretical background for multiple imputation is given by Little/Rubin (1990). For the practical work STATA has implemented an "impute" command and also enables a correction of standard errors when using imputed data.

Evidence from the German Life History Study: The Nonresponse Study

In the following chapters, I will examine a special nonresponse study which forms part of the East German Life History Study[24]. The complete German Life History Study[25] is a retrospective longitudinal data base and contains data of the life course of persons from different birth cohorts. The East German data is a two-wave-panel study (1991,1996) and offers a combination of panel data and retrospective longitudinal data. After the field work of wave 2, the life history project succeeded in establishing a follow-up study on nonrespondents[26] from the first wave. There are several reasons why this data is interesting to analyse:

- Firstly, the study has a general aspect in that it includes <u>additional information about the "dark chapter persons"</u> who do not usually appear at all.
- The more specific aspect can be found, however, in the <u>high comparability of nonresponse and main study</u>. The wave 2 and the nonresponse interviews were collected by the same instruments (CAPI and CATI interviews), by the same survey institute and by the same interviewers. The retrospective design allows one to assign corresponding events of a person's life to the time of the initial sampling, so that equivalent information for the first wave is also available. It also has to be mentioned, that the wording of the questions, the categories and the rules for data edition were identical.
- These advantages allow one to describe and compare nonrespondents and participants of the main study, not only on an aggregate level of information. The individual data enables us to instead <u>consider nonrespondents and respondents together in a multidimensional view</u>. This could not be done with aggregate comparisons. The multivariate context of the nonresponse study offers the opportunity of checking the influence of nonresponse bias and the efficiency of bias corrections.
- This insight into the nature of the nonrespondents in the EGLHS is of course particularly useful for the life history study itself. The <u>specific historical situation</u> makes it interesting from a more general point of view, however. The initial sampling took place in 1990, only shortly after the end of the GDR. It cannot be assumed that the reason for increasing nonresponse is an inflation of surveys[27]. East Germans did not have experience with

---

[24] In the following abbreviated as "EGLHS", directed by K.U. Mayer, Max Planck Institute for Human Development, Berlin.

[25] Data collection in West Germany took place in the 1980s, in East Germany in the 1990s. More information about the German Life History Study can be found on the Internet:
http://www.mpib-berlin.mpg.de/BAG/Studien/ZA-Info38/Wagner.htm (for the West German data)
http://www.mpib-berlin.mpg.de/BAG/Studien/ZA-Info38/solga.htm (for the East German data)

[26] Abbreviated: "NRS"=nonresponse study.

[27] Schnell (1997:30) mentions this for industrial nations with a longer tradition in surveys.

surveys under the old regime. Instead, it is of especial interest as to whether specific socialisation and career patterns and experiences in the life course explain typical groups of nonrespondents.

Taking the advantage of comparable individual data for nonrespondents and respondents, the next chapters will first explore and describe the nonrespondents using the collected data as well as the information from the field reports. Having worked out an imagination what type of people had been missed by the main survey, I will then introduce regression models to predict the participation in the nonresponse study. Following that, I will present strategies to correct for nonresponse bias in multivariate models - especially, an example for correction with the Heckman sample selection model.

2. Exploration of the Nonrespondents in the East German Life History Study

The basic examinations of the nonresponse data are guided by the approach of exploratory data analysis as established by Tukey (1977)[28]. The relationship between exploring data and subsequently confirming hypotheses is important and mutually inspiring (see discussion by Erickson/Nosanchuk[29], 1992) just as graphical presentations are an essential tool. Excellent principles of visualisation are shown by Tufte[30] (1990). The exploratory approach means generating as many views and aspects of the data as possible. This chapter presents tables and graphics for a fundamental comprehension of the data whilst two conflicting principles have to find a balance:

- compression of information: on the one hand, it is necessary to combine as much information as possible in order to compare groups and to detect structures;
- simplification of information: on the other hand, graphical presentations have to show a great deal of simplicity in order to guide the spectator's eye to the essential part.

Before detailed findings of the NRS data analysis can be discussed, however, it is necessary to reflect and to understand the special nature of the nonrespondents' sample. The two initial sections of this chapter deal, therefore, with the sampling and the field phase of the East German Life History Study (EGLHS).

Firstly, I will explain the general design of the study and consider the sample development as a whole, by pursuing it in absolute numbers of target persons. Secondly, I will concentrate on more specific reasons for sample drop-out as reported by the survey institute. It is important to focus on problems concerning the ease of contacting individuals as well as whether persons are likely to refuse or not in order to discover what kind of nonrespondents the NRS file contains.

Finally, after these two parts, the concrete findings of the data analysis of the nonresponse study will be discussed.

---

[28] Tukey's ideas were already published in the 1940s and 1950s. The year 1977 refers to the edition of his book: Tukey, John W. (1977) Exploratory Data Analysis. Addison-Wesley Publishing Company. Reading, Mass.

[29] Erickson, B.H. and Nosanchuk, T.A. (1992) Understanding Data (Second Edition). Open University Press, Buckingham, see pp. 3-11.

[30] Tufte, Edward Rolf (1990) Envisioning Information. Graphics Press. Cheshire, Connecticut. See chapters: "Layering and Separation", pp. 53-66 and "Color and Information", pp. 81-96.

2.1 The Sample Development

The Main Study

The German Life History Study[31] (GLHS) begun in 1981. It started with a retrospective collection of life course data of persons from the birth cohorts 1929-31, 1939-41, 1949-51, 1959-61 in West Germany, with the birth cohorts 1919-21, 1954-56, and 1959-61 being later added (total number of persons in the West German data base: N=5591).

After German reunification, the study was extended to inhabitants of the former GDR. The data which I will analyse here is the East German part of the life history study and stems from birth cohorts 1929-31, 1939-41, 1951-53, 1959-61[32]. The first wave of the main study in East Germany was in 1991. The basis was a master sample from the central list of inhabitants of the GDR, October 1990. The sampling units were persons and a separate sample was drawn for each cohort (infas (1995:2)). The target persons were personally interviewed in paper-pencil interviews during the field phase 1991/92 which resulted in N=2330[33] completely utilisable interviews.

A second wave was already intended at this time and was actually realised in 1996/97. The method changed from paper-pencil to computer-assisted interviews. Computer-assisted telephone interviews (CATI) could be used because the number of households with a telephone[34] in the former GDR had increased over five years. Only the individuals, who still had no telephone or who had refused to be interviewed by telephone but had agreed to a personal contact, were interviewed by computer-assisted personal interviews (CAPI). The wave 2 data consists of N=1394 completed interviews which is a panel success of 59.8% (counting the wave 1 interviews as 100%).

In the same methodological manner as in the first wave, the questions of the CATI/CAPI programme were presented in several thematic modules concerning e.g.

- personal information
- places of residence, household
- jobs and retirement
- education
- membership in organisations
- activities and social network
- marriages, partners, children.

[31] directed by Karl Ulrich Mayer, Max Planck Institute of Human Development, Berlin.

[32] The following abbreviations for the birth cohorts are used: "1930"=1929-31; "1940"=1939-41; "1950"=1951-53; "1960"=1959-61.

[33] After data edition: N=2331 persons. TABLE 1 refers to N=2330 from the field report. The data analysis was done on the basis of the edited file, which later investigated unclear cases and finally resulted in N=2331 utilisable interviews.

[34] 45% households with telephone in wave 1; 63,5% with telephone in wave 2. See field report: infas (1998:9). Ostdeutsche Lebensverläufe im Transformationsprozeß. Methodenbericht zur Hauptstudie (Panelkohorten).

The event oriented spell design for the longitudinal information offered the possibility that the target persons did not have to construct their complete life course ordered by the time axis. To reduce possible recall errors[35], persons were guided time by time within relevant topics of their life. Examples for spell data in the life history study are job-spells, episodes of education, spells for the places of residence, or marriages and episodes of living together with partners. The East German study now combines retrospective longitudinal data with a panel design. This provides a rich source for many sociological research questions (e.g. about educational qualifications and job careers in a transformation society).

The Nonresponse Study

A special nonresponse study was conducted in 1997 in which the Max Planck Institute and the data collection institute infas[36] cooperated. There were two main intentions for this study. The first one was to clear up the methodological problem of the relationship between the nonresponse rate and the sample representativeness. The second one was to find out whether such interviews gathered by special effort can be considered as additional cases of the main study without worsening the quality of the sample.

The nonrespondents whose wave 1 contact protocol showed one of the following reasons were chosen:

- person was not at home
- person could not be reached
- person absent for a longer period
- illness
- person did not cooperate because of lack of time
- person refused information in principle
- other household members pretended that person was not at home
- interview was prevented by third persons
- person refused to give interview.

The last group (the refusals) were not all elected as candidates for the nonresponse study: only those persons whose contact protocol could "justify a new attempt of contact"[37] - as is formulated in the field report. This means that the hard core refusals had been excluded from the beginning. Unfortunately, it is neither reported which proportion of the NRS data is given by initial refusals or by initial non-contacts; nor is the refusal/non-contact information available on an individual level. This disadvantage places all the more importance on the inspection of the aggregate field information (which follows later in the next section). The final pool selected in this manner for the new gross sample of the nonresponse study consisted

[35] For a discussion on the aspect of retrospection and data quality in the German life history data see in: Wehner, S. (1999:14). Exploring and Visualizing Event History Data.
[36] infas Institut für angewandte Sozialwissenschaft GmbH in Bonn Bad Godesberg.
[37] This criterion is cited from the infas field report: infas (1998), see nonresponse study on pp. 26-30.

of N=1246 persons out of 2131 nonrespondents from the initial sample (this means that the NRS gross sample contains (1246/2131)*100=58,5%[38] of the wave 1 nonrespondents). From this gross sample, 282 persons had to be subtracted due to out-of-date addresses, problems of finding persons, etc. This is reported as a whole under the title of "neutral loss" in the field report. The final actualised gross sample of nonrespondents consisted of N=964 persons.

In addition to an announcement letter, an incentive of DM 50,-- was offered to all selected nonresponse target persons. Although the majority of them once again did not participate (69,5% refused again), 201 nonrespondents finally gave complete interviews in which CATI/CAPI programmes[39] similar to the ones in wave 2 were used. We can define the response rate of the NRS study in three ways: taking all the initial nonrespondents of wave 1 as 100%, a response rate of only 9,4% is calculated (201 out of 2131). Taking the pool of the NRS gross sample, we end up with a response rate of 16,1% (201 out of 1246). Defining the actualised gross sample as 100%, we calculate a final response rate of 20,9% (201 out of 964). The last version is given in the infas report.

The Nature of the Nonresponse Sample

Summing up the preconditions of the nonresponse study, we can recognise: independently from the kind of calculation of the response rate, the NRS sample obviously cannot be assumed to be a representative selection of all nonrespondents. Thus generalisations of the findings will be limited. The sample is in principle biased as most of the refusals were again missed out and so it consequently appears to be a filtered extract of special "difficult" target persons.

In the next step, we will take a look at the sample development for all parts of the study in more detail. It was shown above that numbers for percentages depend strongly upon the basis chosen as 100%. For this reason TABLE 1 follows the absolute number of target persons from the initially drawn sample until the final survey data set and lists the reasons why people dropped out.

[38] rounded to one decimal
[39] With slight modifications concerning additional questions about jobs, marriages, partners and education before 1989.

TABLE 1: Sample Development in the East German Life History Study
- Number of Persons in the Target Population -

EGLHS wave 1 1991/92	N=	in %
initial gross sample	4750	100,0%
minus: neutral loss	281	5,9%
remaining actualised gross sample	**4469**	**100%**
minus: systematic loss		
refusals		
person refused to give interview	1463	32,7%
person refused any information	148	3,3%
	1611	36,0%
no contact to household or person	204	4,6%
(not at home or person was absent for a longer period)		
other reasons		
person was ill	84	1,9%
person had no time	124	2,8%
interview prevented by other persons	42	0,9%
person pretended not to be at home	45	1,0%
no information	21	0,5%
	316	7,1%
total systematic loss	2131	47,7%
remaining interviews	**2338**	**52,3%**
minus: not utilisable interviews	8	0,2%
remaining utilisable interviews	**2330**	**52,1%**
final total response rate		52,1%
refusal rate		36,0%

Non Response Study 1996/97	N=	in %
total systematic loss of wave 1	**2131**	
excluding hardest refusals:		
remaining pool of nonrespondents	1246	100,0%
minus: neutral loss	282	22,6%
actualised gross sample	**964**	**100%**
minus: systematic loss		
again refused	670	69,5%
no contact to household or person	61	6,3%
person was ill	29	3,0%
remaining interviews	**204**	21,2%
minus: not utilisable interviews	3	0,3%
remaining utilisable interviews	**201**	**20,9%**
final total response rate		20,9%
refusal rate		69,5%

EGLHS wave 2 1996/97	N=	in %
initial gross sample	2330	100%
minus: panel mortaliy	171	7,3%
minus: neutral loss	261	11,2%
actualised gross sample	**1898**	**100%**
minus: systematic loss		
refusals	328	17,3%
no contact to household or person	92	4,8%
person was ill	40	2,1%
remaining interviews	**1438**	75,8%
minus: not utilisable interviews	44	2,3%
remaining utilisable interviews	**1394**	**73,4%**
final total response rate		73,4%
refusal rate		17,3%
panel success		59,8%

All figures in this table are taken from the methodological reports given by the sampling institute infas:
wave 1: *infas (1995) Lebensverläufe und historischer Wandel in der ehemaligen DDR, Methodenbericht der Hauptstudie. Bonn,* p.19.
wave 2 and nonresponse study: *infas (1998) Ostdeutsche Lebensverläufe im Transformationsprozeß, Methodenbericht zur Hauptstudie. Bonn*, pp.18-19 (wave 2), pp. 27-28 (nonresponse study).

General remark: According to the notation of German software versions, the comma-character is used as a decimal point in all tables and numbers of this thesis.

2.1.1 The First Wave

In 1991/92 the initial gross sample started with N=4750 addresses of target persons that had been drawn up (see left column: "EGLHS wave 1" in TABLE 1) .

Neutral loss (N=281) was caused by out-of-date/incorrect lists. The following reasons are given in the field report: street or house number could not be found, apartment had no inhabitants, not a private household, person was unknown, had new address or did not belong to the target population. In the methodological report, this kind of loss is considered to be non-systematic. The percentages are, therefore, calculated on a new 100% basis for the updated gross sample (N=4469).

The total amount of systematic loss is distributed across the following reasons:

1) refusals

Here we have persons who were contacted but who had refused to give an interview (N=1463). In addition, the field report for wave 1 lists a second type of refusal: persons who in principle refused to give any kind of information to anybody (N=148). If we add both groups together, we have a total refusal rate of 36%.

2) non-contacts

This group (N=204) consists of persons who, despite several attempts at contact, were never at home or who were absent for a longer period.

3) other reasons

Here we find: ill persons (N=84); people who said that they had no time for an interview (N=124); people who pretended (through other household members) not to be at home (N=45); people who were prevented from answering by third persons (N=42); and lastly a very small group with no information concerning the reasons (N=21).

Finally, N=2338 interviews were conducted. There was a marginal number of N=8 *non utilisable interviews* as the interview had been interrupted or errors had been caused by the interviewer. At the end, the wave 1 data set of the EGLHS consists of N=2330 interviews which presents a total response rate of 52,1%.

2.1.2 The Second Wave

In 1996, the initial gross sample of wave 2 consisted of N=2330 respondents from wave 1 (see lower right column "EGLHS wave 2" in TABLE 1). Over a 5 year period, the problem of inaccurate addresses had logically increased. The rate of *neutral loss* (N=261) is now 11,2% of the starting sample whereas it was only 5,9% in wave 1. We are now additionally confronted with the problem of *panel mortality* (N=171) which means that we have a loss of 7,3% of the utilisable wave 1 data set. This panel mortality appeared because persons could not be traced or because persons now refused to participate in wave 2. Unfortunately, both reasons (inaccurate addresses and refusals before interview attempt) are not reported separately but counted together in the methodological report. The neutral loss (N=261) and the panel mortality (N=171) are both reported as "non-systematic" and then subtracted from the initial gross sample. The remaining sample consists of N=1898 cases and is again calculated as 100%. Although I am arguing that it is not obvious to consider the complete panel mortality as a "neutral" loss of target persons, in TABLE 1 I decided to follow the criteria provided by the infas report. I did this because the survey field reports are the empirical researcher's basic information and must be taken as they are, especially when the researcher is using data that has already been gathered. Subsequent to TABLE 1, I will present a visualisation of the sample development which only contrasts absolute numbers and avoids discussible fixings of 100%. Schnell (1997:23,26-27,72-76) also discusses the problems of defining "neutral" loss and calculating response rates on the basis of the actualised gross sample[40] as 100%. This is usually done and obviously improves the survey institutes' success rate.

The reasons for *systematic loss* are not given in the same detailed categories as in wave 1. Now we only have to subtract one group of refusals (N=328), the group of non-contacts (N=92) and a group of ill persons (N=40). After the subtraction of 44 non utilisable interviews, we end up with a wave 2 data set of N=1394 persons. On the basis of the updated 100% sample (N=1898), this is a response rate of 73,4%. The refusal rate[41] of 17,3% is lower than in wave 1. The panel success (based upon the final wave 1 and wave 2 interviews) is 59,8%.

2.1.3 The Nonresponse Study

The nonresponse study was conducted in 1997 after the field work for wave 2 (see upper right column "Non Response Study 1996/97" in TABLE 1). It started with an initial gross sample (subset of wave 1 nonrespondents with adamant refusals excluded) of N=1246 addresses. The *neutral loss* (N=282) had to then be subtracted. (Again the methodological report updates the initial target group and sets it to 100% (N=964). This gives rise to the same doubts over whether it is a "non systematic" loss as discussed in the case of the panel mortality.) After losing target population through contacted persons who had once again refused (N=670), non-contacts (N=61), ill persons (N=29), and non utilisable interviews (N=3), the final nonresponse study consists of N=201 interviews. On the basis of the updated gross sample (N=964) the response rate is 20,9% (see earlier discussion about calculating this), and the

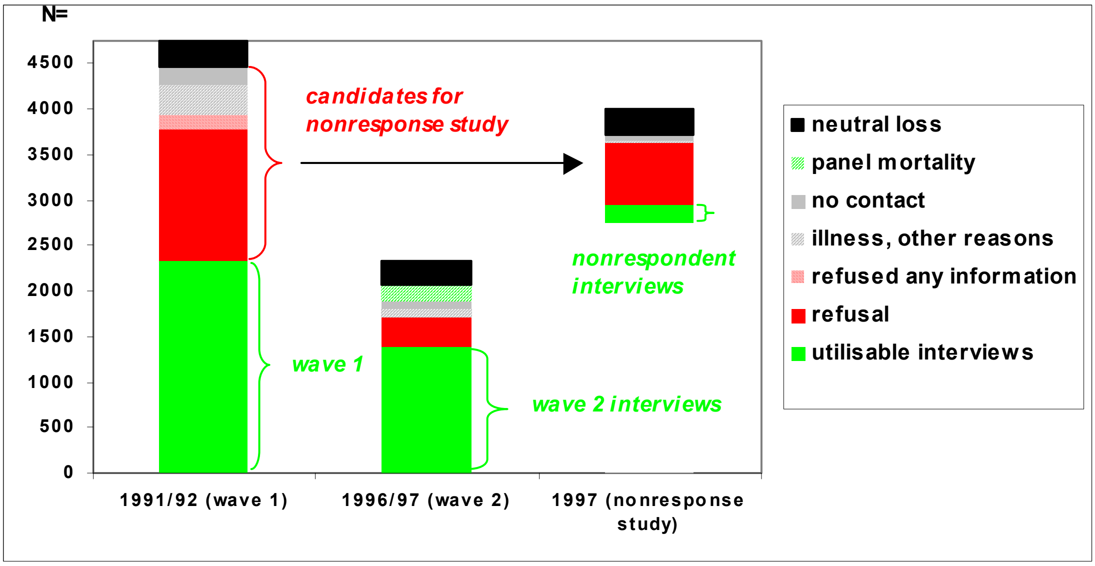
[40] in German: bereinigte Bruttostichprobe

[41] An examination of the panel persons (only cohorts 1940,50,60) revealed that a positive answer to the wave 1 question "whether a person would be willing to give another interview in a later study" is a significant indicator of participation in wave 2. See Wehner, S. (2000).

refusal rate is 69,5%. This seems to be a poor result[42] if we restrict ourselves to looking only at these percentages. We have to keep in mind, however, that this is a study of a special group. We have got information about some "dark chapter persons" who usually do not appear in surveys at all.

FIGURE 4 below provides a visual image of the sample development. Colour attributes are used especially to enable a graphical contrast of the proportion of completed interviews versus nonrespondents.

FIGURE 4: Sample Development in the East German Life History Study
- Nonrespondents versus Completed Interviews -



The left bar is a stack of the initial drawn sample: its total length represents N=4750 persons. The neutral loss is given in black, non-contacts in grey, refusals of any kind of information (only wave 1) in light red, illness together with other reasons (in wave 1) and non utilisable interviews are given in grey stripes. The green bar indicates the final amount of utilisable interviews. In wave 2 we also see the number of persons lost by panel mortality (green stripes). In all three studies, the most obvious blocks are the refusals (in red) and the completed interviews (in green). In wave 1 about one third of the initial sample are refusals. This proportion is clearly much lower in wave 2, because only the participants were contacted 5 years later. We see that the wave 2 stack contains only a small red band.

[42] Groves/Couper (1998:50) mention that special nonresponse studies are often confronted with low response rates.

The stack of the nonresponse study contains these initial nonrespondents[43] who were selected for the study (in other words, the harshest refusals were excluded). The stack of the nonresponse study shows a high proportion of persons who again refused to be interviewed. The small green band indicates the successfully converted nonrespondents who form the basis of the following explorations.

We see graphically in FIGURE 4 the proportion of successful response (green) in comparison to the complete initial sample. We can also already discern a considerable amount of loss (black, grey, red) at the initial sampling time. This is a picture that every survey should draw as it enables one to visualise the starting and final situation of the sample. As already mentioned, however, this is often not the point of view of the sampling institutions.

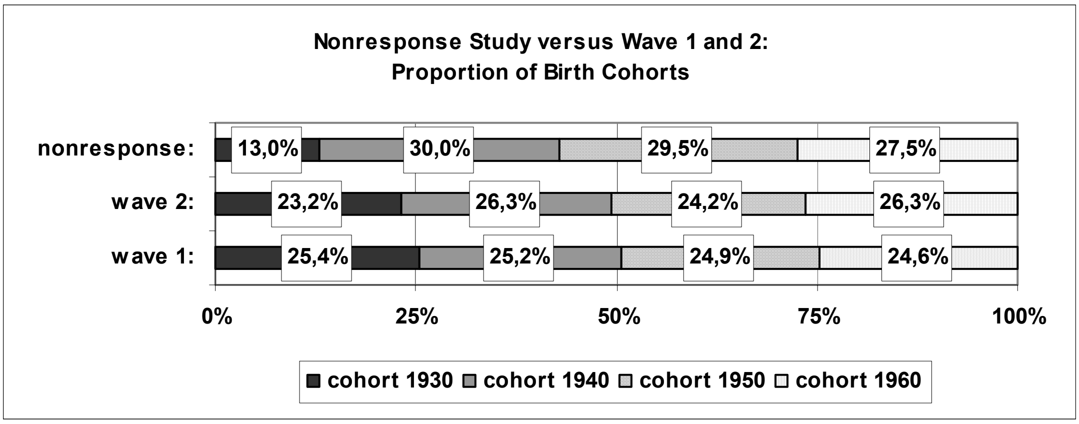
In the following section, I will examine the different reasons for loss of target persons more intensively. The sources of information are the methodological reports by infas (1995 and 1998) which also list the field situation divided into cohorts. In order to cite the numbers, I had to adopt the practice of calculating rates and percentages on the basis of the actualised gross sample. The respective tables indicate the basis for the corresponding 100%.

2.2 Reasons for Sample Drop-Out

2.2.1 The Cohort Composition of the NRS File

The first noticeable attribute of the nonresponse data is an overproportional lack of cohort 1930 as can be seen in the next graph.

FIGURE 5: Cohort Composition of Wave 1 and 2 and Nonresponse Study



[43] As explained in the chapter about sources of error, the real problem is the initial nonresponse. The wave 2 panel mortality is, therefore, not part of the nonresponse study.

The wave 1 sample started with cohort groups of equal size[44], due to the intentional cohort design of the life history study. In wave 2, five years later, the cohort composition changed slightly but not that much.

Only 13% of the NRS persons belong to cohort 1930 whereas the starting percentage in wave 1 was 25,4%. This large difference cannot be explained by natural mortality in the oldest group. Given this, an assumed mortality influence should also have occurred in wave 2 in which the interviews were performed only several months earlier. When one takes wave 1 as the basis for comparison, the percentage of cohort 1930 in the NRS file is only 50% of the expected proportion (13% NRS versus 25,4% wave 1), whereas the cohort reduction factor in wave 2 is only 91% (23,2% wave 2 versus 25,4% wave 1). I additionally checked this with official census data[45]. When one takes the percentage for cohort 1930 of the microcensus 1991 as the basis (=19,5%) and the corresponding percentage in 1995 from the Statistical Yearbook (=18,3%), we get a similar reduction factor of about 93,8%.

To assess whether the different cohorts in the NRS file might represent different types of nonrespondents, I collected the reported reasons for loss of sample persons per cohort.

Reasons for Loss in the Initial Sampling

Firstly, we will look at the situation during the initial field. TABLE 2 below lists the percentages of loss by different reasons.

---

[44] This is of course not the representation of the German population pyramid. Due to the fact that I will not use the total initial sample as a whole for inferential statistics about the population, however, possible cohort stratification weights are neglected. The intention is to instead detect the differences or similarities of the NRS nonrespondents.

[45] The same birth cohorts as in the EGLHS data were selected, i.e. 1930=1929-31, 1940=1939-41, 1950=1951-53, 1960=1961-63. The four groups altogether counted as 100%, so that the percentage e.g. 19,5% for cohort 1930 indicates the relative proportion to the pool of the four cohort groups. Only the East German population was taken. It has to be mentioned that the census data counts the population by a regional concept. Therefore, East German population might contain West-East movers. In contrast to which, the life history study sampled using the concept of origin. This means that the sample only contains persons who were citizens of the GDR in 1989. The sample does not, therefore, contain original West German citizens, although target persons might have moved to cities in West Germany. It is necessary to be aware of these conceptional differences so that margin distributions can only offer an approximate check. Sources for census information are:
Mikrozensus 1991, 70% subsample (scientific public use file) of the complete census which is a 1% population sample. Neue Bundesländer, Berlin only the former Eastern part.
Statistical Yearbook 1997, Federal Statistical Office Wiesbaden. Population data for 1995 on page 62.

TABLE 2: Reasons for Loss of Target Persons per Cohort - Initial Sampling 1991/92

		cohorts			
	total	**1930**	**1940**	**1950**	**1960**
starting with the initial gross sample (N=4750):					
percentage of total neutral loss	5,9%	3,1%	4,1%	6,6%	9,5%
with main reasons given as:					
person unknown	3,0%	1,1%	1,6%	4,0%	5,1%
person has new address	0,2%	0,0%	0,3%	0,1%	0,5%
person died	0,7%	1,0%	1,2%	0,2%	0,2%
now taking the actualised gross sample as 100% (N=4469):					
systematic loss given by:					
person had no time	2,8%	1,8%	2,4%	4,0%	2,9%
interview prevented by other persons	0,9%	1,0%	1,0%	1,0%	0,8%
person pretended not to be at home	1,0%	0,7%	1,1%	1,2%	1,0%
no information	0,5%	0,3%	0,7%	0,5%	0,4%
illness	1,9%	3,4%	2,0%	0,9%	1,4%
no contact	**4,6%**	**3,1%**	**3,6%**	**5,1%**	**6,4%**
refusal rate	**36,0%**	**40,9%**	**37,5%**	**35,1%**	**35,5%**
(as refusal counted both: person refused to give interview or refused any kind of information in principle)					
total response rate (counting remaining utilisable interviews)	52,1%	53,5%	51,6%	52,0%	51,4%
now taking the realised interviews as 100% (N=2330):					
number of contacts only 1	3,2%	4,8% (m)	1,7% (m)	2,1% (m)	2,3% (m)
		5,0% (w)	3,8% (w)	2,8% (w)	2,9% (w)
2	42,2%	46,6% (m)	35,5% (m)	40,7% (m)	40,9% (m)
		49,3% (w)	39,0% (w)	41,3% (w)	44,3% (w)
3	27,9%	28,6% (m)	33,1% (m)	26,6% (m)	26,0% (m)
		24,8% (w)	30,8% (w)	29,0% (w)	25,6% (w)
4-7 and more	26,5%	20,0% (m)	29,7% (m)	31,0% (m)	31,7% (m)
		20,9% (w)	26,4% (w)	26,9% (w)	27,5% (w)
number of persons in the realised sample (utilisable interviews) N=	**2330**	**592**	**586**	**577**	**575**

(m) = men only (w) = women only

(Categories for reasons of loss and numbers were taken from: infas methodological report (1995) overviews 5 and 12, see also TABLE 1 above. Information for number of contacts is cited by gender as given in the report.)

The total response rates do not differ very much across the cohorts (the lowest is 51,4% for cohort 1960, the highest 53,5% for cohort 1930). If we consider the structure of sources of loss, however, we see that the 1930 and the 1960 people are at the end of contrasting scales. On the whole, the reasons for loss of target persons are described as follows:

- The chances of finding and contacting people are best for the oldest cohort 1930 and decline across the cohorts. The younger the target persons are, the more likely it is that difficulties will arise concerning out-of-date addresses and chances of contacting individuals .

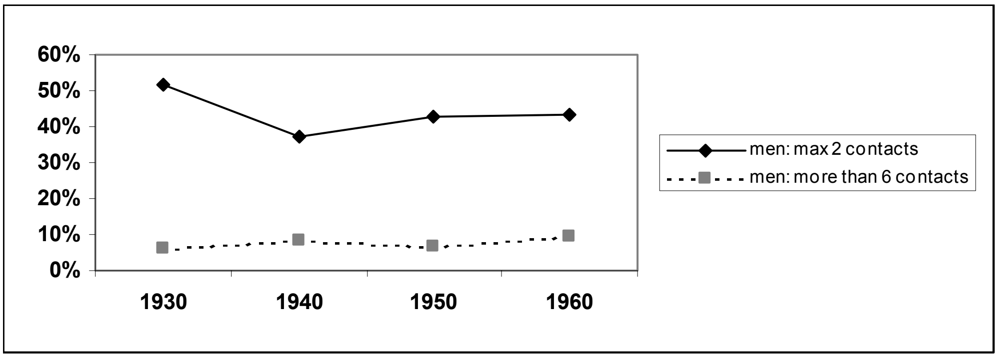
- The chances of convincing a person to give an interview, given contact could be established, is lowest for older people and highest for the young cohort 1960. The refusal rates increase from cohort 1960 to cohort 1930. This might nor merely be an effect of age, but also reflects different GDR specific experiences in the life course. For cohort 1930, the interview happened after a complete phase of active work in the GDR. Persons of cohort 1960 had, on the one hand, experienced their education and the beginning of their careers in the old regime. On the other hand, the political change concerned them to a larger extent as it occured in their years of establishing a family and a career.

Some single results from TABLE 2 are:

The total "neutral loss" is lowest for the oldest persons (cohort 1930: about 60 years old at field time) with 3,1% and highest for the youngest persons (cohort 1960: about 30 years old) with 9,5%. The predominant category is "person unknown" (5,1% for cohort 1960; 1,1% for cohort 1930).
The category "illness" does not have very high percentages, although this rises from younger to older persons.
The non-contact rate for the youngest group is double (6,4%) the rate of the oldest persons (3,1%). This is also consistent with the examination of the number of necessary contacts (see last section of TABLE 2). In cohort 1930, 51,4% (4,8%+46,6%) of the men could already be interviewed with a maximum effort of 2 contacts, whereas only 43,2% (2,3%+40,9%) of the 1960 men could be reached after a maximum of 2 contacts.

The following figure demonstrates how cohort 1930 was easier to contact. It shows the percentage of men who were interviewed after a maximum of 2 contacts (full line) and the percentage for those with 6 or more necessary contacts (dashed line). We see that cohort 1930 is an exception when contrasted with any other group: the "low effort" group (maximum of 2 necessary contacts) is highest with over 50% (versus 43,2% for 1960), the "high effort" group (6 or more contacts) is lowest with about 6% (versus 9,2% for 1960).

FIGURE 6: Percentage of Successful Interviews after Low and High Number of Contacts Men in Wave 1 1991



100%=all completed interviews

The corresponding numbers for women are very similar, as they are generally easier to contact: the "low effort" curve for women will be above the men's line and the "high effort" curve under the line for men.

Reasons for Loss in Wave 2

The next table lists the reasons for sample drop-out in the panel study.

TABLE 3: Reasons for Loss of Target Persons per Cohort - Wave 2 1996/97

			total	cohorts 1930	1940	1950	1960
situation before starting wave 2 (N=2330):							
field report lists loss of wave 1 participants under the title ***"panel mortality"*** reasons: "person could not be found" and "target person had refused panel participation in a written statement" (reasons given together in <u>one</u> category)			7,3%	7,4%	7,8%	7,5%	6,6%
now counting the wave 2 pool with above panel mortality subtracted as 100% (N=2159):							
percentage of total neutral loss			12,1%	14,4%	8,6%	11,8%	13,4%
with main reasons given as:							
person unknown			4,2%	3,6%	2,4%	5,4%	5,2%
person has new address			3,7%	3,1%	3,6%	3,0%	5,2%
person died			2,0%	5,5%	1,5%	0,4%	0,6%
now taking the new actualised gross sample as 100% (N=1898):							
systematic loss given by:							
illness			2,1%	5,8%	1,8%	0,8%	0,0%
no contact			**4,8%**	**4,5%**	**4,6%**	**4,7%**	**5,6%**
refusal rate			**17,2%**	**19,0%**	**17,2%**	**18,9%**	**14,0%**
(systematic loss is only given in these three categories, not in the same detail as in wave 1)							
total response rate (counting remaining utilisable interviews)			73,4%	69,0%	74,2%	72,0%	78,4%
final panel success (percentage of realised wave 2 interviews from wave 1 interviews)			59,8%	54,4%	61,6%	57,7%	63,2%
now taking the realised interviews as 100% (N=1394):							
number of contacts (given only for CAPI interviews)		only 1	10,1%	7,0% (m)	15,0% (m)	7,0% (m)	5,3% (m)
				11,9% (w)	15,4% (w)	11,3% (w)	8,8% (w)
		2	35,9%	55,8% (m)	20,0% (m)	30,2% (m)	28,1% (m)
				54,8% (w)	26,9% (w)	35,8% (w)	38,2% (w)
		3	30,9%	25,6% (m)	37,5% (m)	34,9% (m)	22,8% (m)
				19,0% (w)	38,5% (w)	37,7% (w)	30,9% (w)
		4-7 and more	20,9%	11,6% (m)	27,5% (m)	27,9% (m)	43,8% (m)
				13,8% (w)	19,2% (w)	15,1% (w)	22,1% (w)
number of persons in the realised sample (utilisable interviews)		N=	**1394**	**324**	**366**	**339**	**365**

(m) = men only (w) = women only

(Categories for reasons of loss and numbers were taken from: infas methodological report (1998:10-19), see also TABLE 1. Information for number of contacts is cited by gender as given in the report.)

The total "panel mortality rate" (under which both are reported: either person/ address could not be found or person refused panel participation by a written statement) does not differ very much across cohorts (6,6% for 1960; 7,4% for 1930). Generally, the total neutral loss is higher than in wave 1 because, at the time of the first wave in 1991, the addresses were relatively up-to-date, whereas in wave 2 in 1995, possible residential mobility during the years had been taken into account.

The highest loss due to the task of actualising the gross sample is given for the two contrasting cohorts. Again they present a structure with opposing aspects: finding persons and addresses correctly is most difficult for the younger persons ("person unknown" + "new address" =10,4% for 1960), whereas problems increased in the category "person died" for cohort 1930, who were about 65 years old.

This pattern is also given in the reasons for systematic loss where the participation of cohort 1930 suffers most through "illness" (5,8%) in contrast to the other cohorts.

The no contact rate is at nearly the same level as in wave 1 and reveals the same pattern across cohorts (4,5% lowest for cohort 1930, 5,6% highest for cohort 1960).

The refusal rate is generally much lower than in wave 1. This is plausible since we only consider persons who had already agreed to give the first interview. 86% of the wave 1 persons had agreed to give a second interview later (see: infas (1995:23)). The overall pattern of the refusal rates produces the same trend across cohorts as wave 1. So in conclusion, we find that wave 2 confirms the finding:

- the oldest cohort 1930 is the "refusal cohort"
- the youngest cohort is the "hard to reach cohort".

Reasons for Loss in the Nonresponse Study

We will now investigate whether the same characterisation can be found in the nonresponse study. The following table also lists the reported reasons for loss per cohort in this study.

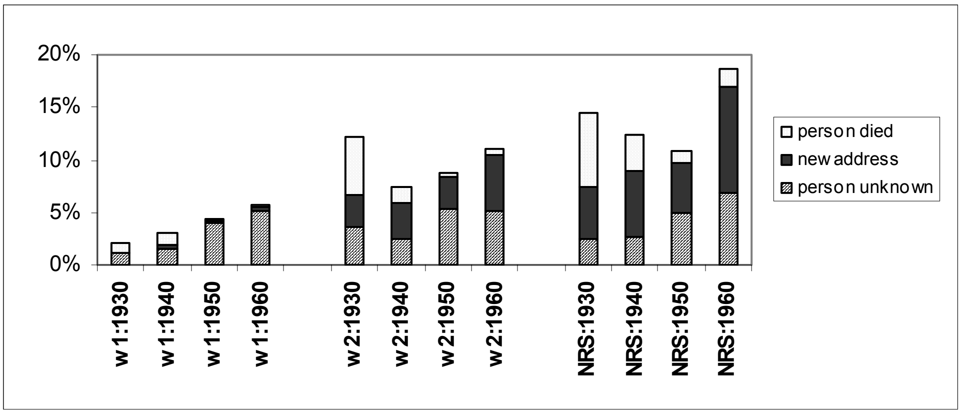
TABLE 4: Reasons for Loss of Target Persons per Cohort - Nonresponse Study 1996/97

		cohorts			
	total	1930	1940	1950	1960
situation before starting NRS study:					
% of persons from the initial nonrespondents who were selected for the NRS study (N=1246 out of 2131) / no information about proportion of refusals/non-contacts	58,5%	55,3%	63,5%	60,6%	54,3%
now counting these candidates for theNRS sample as 100% (N=1246):					
percentage of total neutral loss	22,6%	25,1%	18,3%	20,5%	27,4%
with main reasons given as:					
person unknown	4,2%	2,5%	2,6%	5,0%	6,8%
person has new address	6,5%	5,0%	6,3%	4,7%	10,1%
person died	3,3%	7,0%	3,4%	1,2%	1,7%
now taking the new actualised gross sample as 100% (N=964):					
systematic loss given by:					
illness	3,0%	6,6%	2,5%	1,6%	1,9%
no contact	**6,3%**	**2,8%**	**5,6%**	**8,2%**	**8,4%**
refusal rate	**69,5%**	**78,2%**	**70,4%**	**66,3%**	**63,5%**
total response rate (counting remaining utilizable interviews)	20,9%	12,3%	21,5%	23,1%	25,7%
now taking the realised interviews as 100% (N=201):					
no information about the number of contacts; no individual information who was a refusal and who was a non-contacted person in the initial sampling 1991					
number of persons in the realised sample (utilisable interviews) N=	**201**	**26**	**61**	**59**	**55**

(Categories for reasons of loss and numbers were taken from: infas methodological report (1998:28-29), see also TABLE 1.)

The generally high refusal rates highlight the special character of the NRS data. What could finally be gathered from this study appears to be the hard-to-reach persons, who have already been mentioned. However, we also find that cohort 1930 has the highest refusal rates (78,2% against 63,5% for 1960), whereas cohort 1960 suffers most from mobility problems (6,8%+10,1%=16,9% person unknown or new address; 8,4% no contact rate). The following two figures show the composition of reasons for loss. FIGURE 7 gives the reasons for loss before the field started, i.e. during the phase of checking the addresses. We can recognise increasing problems in obtaining correct addresses for younger persons. The mobility is highest for cohort 1960.

FIGURE 7: Reasons for Loss of Target Persons before Field Beginning Comparison of Cohorts per Study

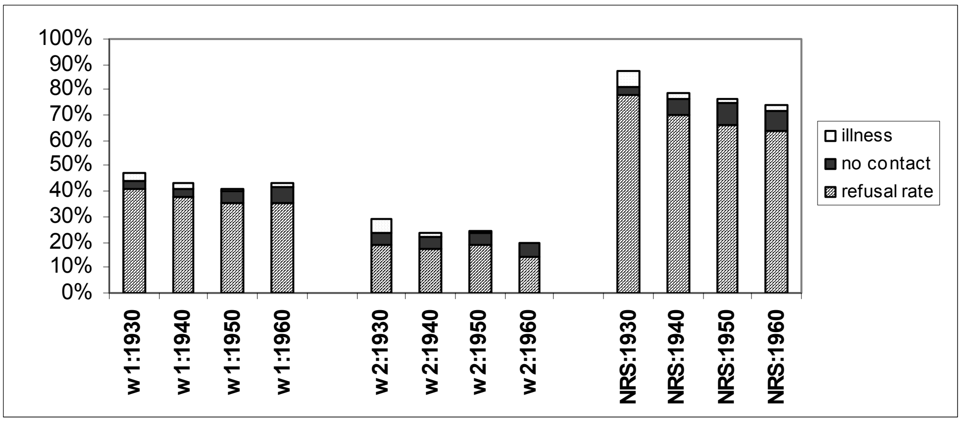


w1=wave 1 1991/92
w2=wave 2 1996/97
NRS=nonreponse study 1996/97

100%=gross sample before checking addresses

FIGURE 8 shows the systematic loss after the gross sample had been actualised. For cohort 1930, we have a stable pattern of rather low non-contact rates in comparison to the refusal rates for the youngest cohort, which are constantly the highest.

FIGURE 8: Reason for Systematic Loss after Field Beginning - Comparison of Cohorts per Study



w1=wave 1 1991/92
w2=wave 2 1996/97
NRS=nonreponse study 1996/97

100%=actualised gross sample

Summary of the Cohortwise Inspection:

There is a stable trend for birth cohorts in all three parts of the study with the following characterisation at both ends:

- The oldest birth cohort 1930 has the highest refusal rates, but fewest problems concerning being located and contacted. A clear part of the loss of these persons is caused by higher rates of mortality and illness. In the context of sample participation, I will name the **cohort 1930: "the refusing generation".**
- The youngest birth cohort 1960 is obviously the most problematic with regards to being located and contacted. Given a successful contact, however, the least effort is then

required to get interviews (lowest refusal rates, least number of contacts). I will name this **cohort 1960: "the mobile generation".**

2.2.2 Gender Proportions

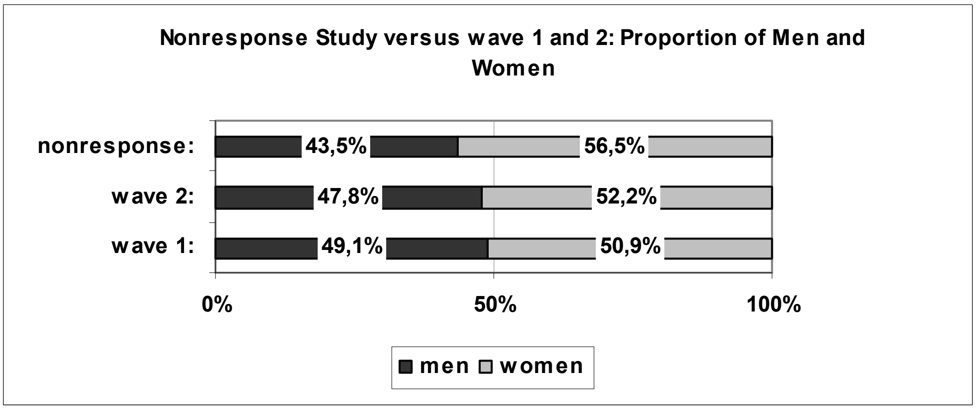
Looking at the loss of people due to invalid addresses, the no-contact rates and the necessary number of contacts, wave 1 has a better sample participation for women. On the other hand, women tend to be more reluctant than men. The following results are reported for wave 1 (see: infas (1995), overview 11)):

	men	**women**
total neutral loss in wave 1	7,9%	4,1%
category: "person unknown"	4,1%	2,0%
no contact in wave 1	5,3%	3,9%
refusal rate in wave 1	32,2%	38,7%

The analysis of the number of contacts also revealed earlier success in interviewing women (see above the text following FIGURE 6).

The next figure presents the proportion of men and women in the realised samples.

FIGURE 9: Gender Proportions



We see that the men start to disappear in the panel study and even more so in the NRS file. In general the picture is the same when inspecting the gender proportion per cohort (see next figure). If we concentrate on both ends of the cohorts, there is a contrary result once again. Concerning the survey participation, the youngest people show virtually no gender differences with 46,5% men in the beginning, 43,8% in wave 2 and 45,5% in the NRS study. The oldest group has an exceptional position: wave 2 contains more men than the initial sample(52,2% in wave 2; 49,0% in wave 1) whereas the NRS file has lost men.

FIGURE 10: Gender Proportions per Cohort and Study

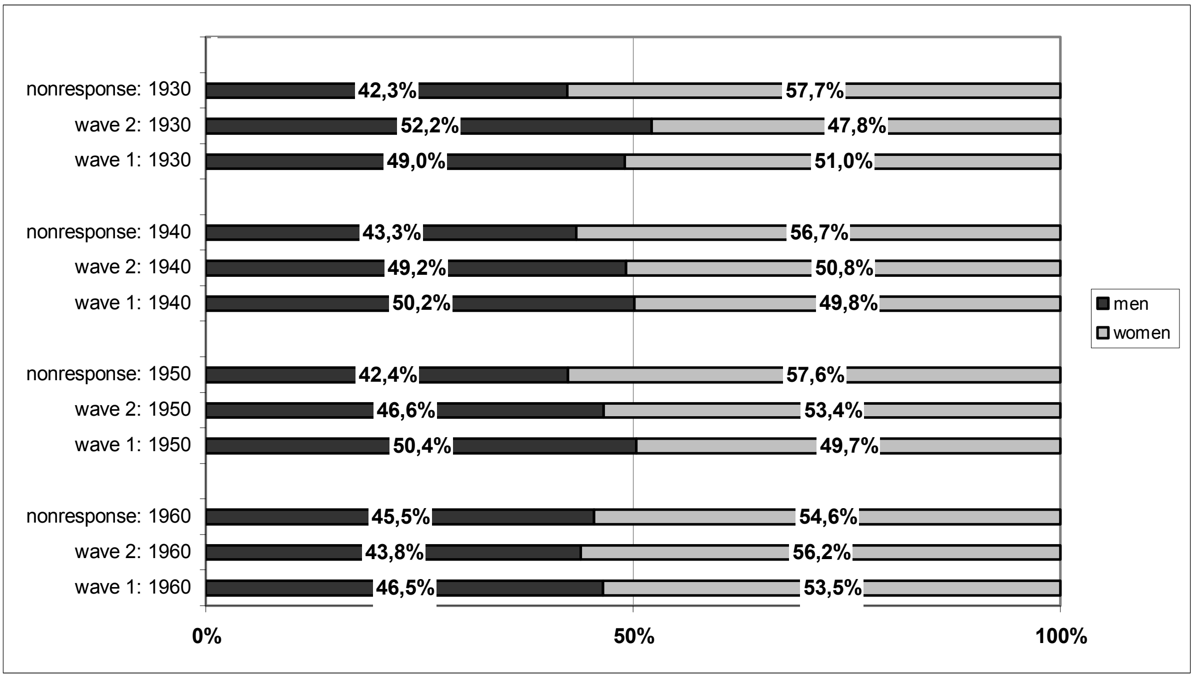
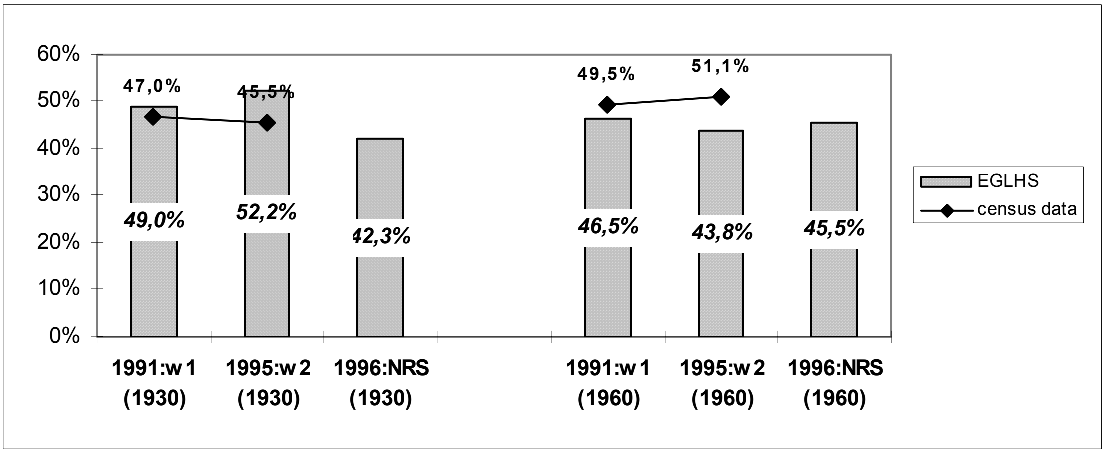


FIGURE 11 below shows the proportion of men indicated by grey bars in the two contrasting cohorts. The corresponding census data percentages for 1991 and 1995 are symbolised by dots. In the young group, men are underrepresented from the beginning and this is maintained over time. This reflects that the NRS study reached men and women to the same extent which is obviously not the case for the oldest cohort. The strangest effect is that of obtaining more men in the panel.

FIGURE 11: Cohorts 1930 and 1960: Percentage of Men in EGLHS and Census Data



Summary:

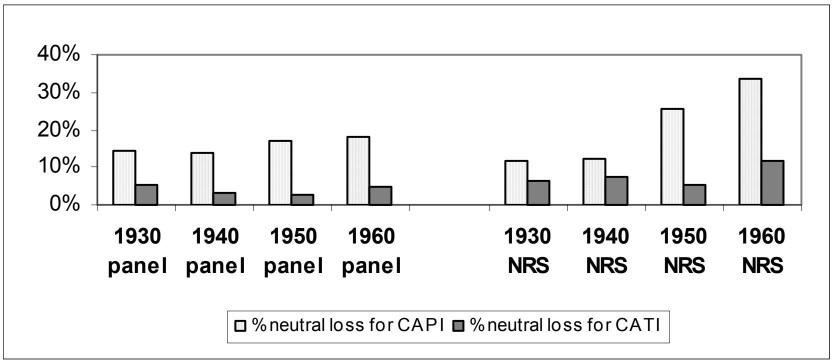
The final result is that a tendency towards a lack of men emerges, which is not consistent for all groups. Whereas in the "mobile generation" 1960, the gender proportion is retained, the "refusing generation" 1930 shows larger differences throughout the studies. This result requires attention throughout the analysis of the nonresponse data.

2.2.3 The Interview Method

The final part of this section about the field situation is concerned with aspects related to the interview methods, particularly with differences between personal (CAPI) and telephone (CATI) interviews. After the decision to switch from paper-pencil to computer-assisted interviews in wave 2, a combination of CAPI and CATI methods was necessary. The aim was to recruit as many persons as possible by telephone because the investigation of mobile persons is easier by telephone. Since the telephone density in East German regions was still at a level of 63,5%, however, personal interviews were also necessary.

It turned out that the telephone survey was much more effective with regards to being able to contact individuals (checking and actualising addresses, finding the target persons). The costs of several attempts[46] at contact are more favourable in telephone surveys than in personal interviews. We can recognise the telephone survey as being more efficient in FIGURE 12. The graph shows the percentage of neutral loss for CATI and CAPI .

FIGURE 12: CAPI/CATI Differences for Percentages of Neutral Loss per Cohort and Study



neutral loss here means: "persons unknown" + "new address"

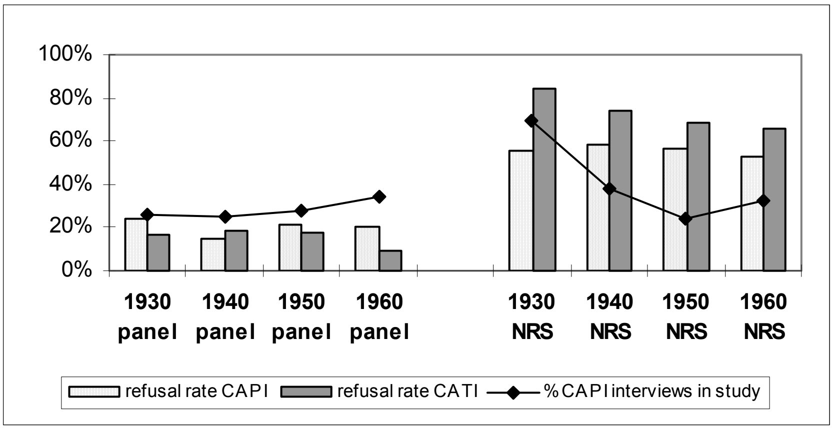
100%=gross sample before checking addresses

We can see for all cohorts and also for both panel and nonresponse study, that the rates of loss are clearly lower in the telephone field. In the nonresponse study (NRS), problems with addresses increase over cohorts and are highest for CAPI target persons in the mobile generation 1960.

The next figure focuses on the refusal rates.

[46] There was no limit to the number of attempts to find and check the address status for CATI persons whereas the CAPI field was limited to a maximum of 5 contacts (with even more in some individual cases).

FIGURE 13: CAPI/CATI Differences of Refusal Rates per Cohort and Study



100%=actualised gross sample

numbers taken from: infas (1998:18-19,29)

In addition to the bars, which mark the refusal rate for the CAPI and the CATI field, the dots connected by a line indicate the total percentage of realised CAPI interviews in the corresponding study.

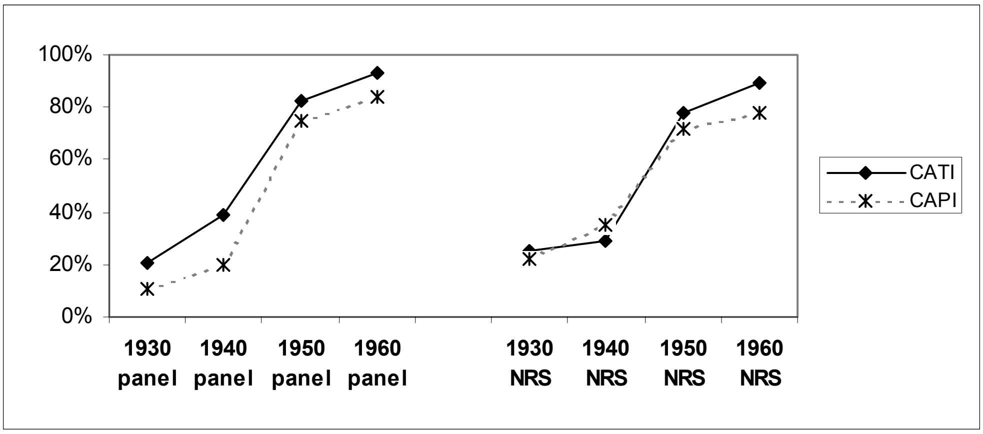
In the panel study, the refusal rates over cohorts do not differ greatly across the interview methods (exception: cohort 1960, where telephone clearly has lower refusal rates). In the nonresponse study, however, the CAPI method could reduce refusal rates. In all NRS cohorts the refusal rates for telephone are higher[47]. We can once again see the striking position of the "refusing generation" 1930 with the highest refusal rates when the telephone was used. We also find in general that cohort 1930 is exceptional with regard to interview method use: CAPI was used in 69,2% of the NRS interviews; in wave 2 it was only in 25,8%. It turned out that this cohort was again the most difficult to convince to participate.

At the end of the reflections about the interview method, it has to be mentioned that the telephone field in East Germany tends to favour higher education and higher income groups. This emphasises all the more the fact that the life history study had to decide on personal interviews as well and, as a consequence, accept higher survey costs. FIGURE 14 compares the participation of the highest educational levels in school under the CATI/CAPI aspect.

[47] Obviously it is better to convince potential survey participants personally than over the telephone. Looking at e.g. the reported refusal reasons in wave 2, the statement "I am not interested" was mentioned in 36,7% of the reasons for CATI, but in only 22,4% for CAPI. See infas (1998:20).

In the main panel study the telephone field recruited more better educated people. This is stable across all cohorts. The nonresponse study only gives this result for the youngest cohorts 1950 and 1960. Bearing in mind that cohort 1930 has the extremely high CAPI proportion, a special selection mechanism seems to have occurred here that cannot be explained by mere socio-economic differences.

FIGURE 14: Percentage of Higher School Degrees in CATI and CAPI Interviews



Percentages were counted together for the educational levels: "grade 10" + "Abitur", which are the two highest levels out of four.

100%=all completed interviews in panel or nonresponse study

Remark:
The use of a line diagram has the purpose of visualising the trend of cohort and interview methods. It should NOT be interpreted as linear interpolation between the 10 years cohort gap. The increase from 1940 to 1950 reflects the historical change in the school system. Whereas cohorts 1930 and 1940 were at school before or during the World War (usual case: 8 years of school), the 1950 and 1960 cohort experienced their educational careers under the reformed GDR school system (usual case: 10 years of school).

The next section presents particular results of the analysis of the nonresponse data.

2.3 Exploring the Nonresponse Data

The guiding question of this chapter is:

What typology of original target persons is represented by the collection of NRS interviews?

I suggest the following structured approach to the answer.

The first step is to introduce a selection of appropriate variables, to define the underlying point of view for the descriptive work and to formulate working hypotheses.

The second step is to find out which of the variables concerned serve as discriminating attributes of the nonrespondents. This task will be done by statistical tests for group comparisons.

Finally, these variables will be presented in more detail by describing the quality and the direction of differences.

2.3.1 Strategy for the Comparison of Nonresponse and EGLHS Main Study

Selection of Variables

In principle only the variables which provide analogue information for the nonresponse and main study[48] were inspected. TABLE 5 gives an overview of the selection.

TABLE 5: Topics and Variables for NRS Data Examination

topic	variables
basic demographics	gender, cohort
family	marital status 1991 and 1996, number of children 1991 and 1996
residence and household	number of residence changes, number of city changes between December 1989 and interview, months living in actual residence 1996, household size 1996, single household 1996? monthly net household income 1996, monthly net household income 1996 per person
education	highest school degree by December 1989, highest degree of vocational training by December 1989
job and career	employed 1991 and 1996, unemployed 1991 and 1996, in retirement 1991 and 1996, position in job 1991 and 1996, number of jobs up until December 1991, number of jobs between December 1989 and interview
political and psychological variables	party membership by December 1989, self esteem items (e.g.: "I have a positive attitude towards myself.")
methodological variables	interview method (CATI/CAPI)

(For the complete codebook of the used file with percentage distributions per group see: appendix A.)

[48] According to the stand of data edition October 2000. For this reason, the scales for attitudes towards parties (1996) were not chosen in this first selection. However, they were inspected later and will be discussed at the end of the section about political variables.

The reasons for the selection of the variables are as follows:

- ***gender and cohort*** as the basic demographic and design variables; following the results of the field inspection, special attention will be paid to cohorts 1930 and 1960 (see section above)
- ***variables which are often considered to be related to nonresponse behaviour***
  family: marital status and number of children; mobility in residence and household variables, such as the number of persons and household income; education: degrees in school and vocational training. Brehm (1993:23-38) gives examples from U.S. data: e.g. surveys tend to undercount men; telephone surveys as opposed to face-to-face interviews might underestimate lower educational categories; older persons are easier to contact but are more likely to refuse. Also see Schnell (1997:198-209, 219-223), Kish (1965:533)
- ***job and career variables***[49] because they are relevant for the special research interest of the German Life History Study; Kish (1965) states that the nature of refusals might differ by social class or profession
- ***political and psychological variables*** (available: party membership in 1989 and self-esteem items from the 1996 study) because it is argued that nonrespondents might differ from survey participants in their opinions, as shown by Burton[50] (1999:219-224). Burton discovered through British data, that reluctant persons tend to refuse as a result of their concern about privacy and confidentiality. They are also less interested in politics and more pessimistic; whereas hard-to-contact people are more willing to give answers to questions about income or savings. Psychological characteristics might also play a role as to whether a person refuses or not (see critical discussion in Schnell (1997:190-193) about "hard core refusals")
- ***methodological variables*** (equivalently available for both studies: CATI or CAPI interview) as research immanent information.

Most of the variables are time-related. They were, therefore, constructed for 1996 (wave 2 and NRS persons) and 1991 (wave 1 and NRS). With regard to problems of retrospection, the best comparison is the one between the NRS and wave 2 since both groups were interviewed within a short time-span (1996/97). The comparison between NRS and the original wave 1 sample has to take a principle disadvantage into account: the 1991 variables were up-to-date information for the wave 1 persons, but were retrospective for the nonresponse persons. For this reason, only sufficiently clear questions were chosen (e.g. only official marital status

[49] In the subsequent analysis only for cohorts 1940,50,60. Cohort 1930 was not interviewed about jobs in wave 2, because most of the target persons were already in retirement.
[50] Burton, Jonathan (1999). Public Attitudes and Responses to Survey Research. University of Essex. PhD thesis.

instead of living with partners which was also asked). It is important to recognise, however, that all the listed variables of both studies were collected under exactly the same conditions: by the same survey institute and in the same interview situation; by the same wording of the questions; the same answer categories and rules for data edition. The fact that some of these topics cannot be compared is often a big drawback when a validation of survey data is intended by similar studies or by census data.

Basic Group Comparisons

The study inherent structure allows for several group comparisons:

1) NRS persons versus all persons of the original 1991 sample. This is possible for variables related to 1991/92.
2) NRS versus those persons of the initial sample who did not participate in the panel study (not-panel-persons). This is only possible for variables from 1991/92.
3) NRS versus panel persons. This is possible for comparable variables collected in 1996/97 and of course also for variables from 1991/92.

Overview of the possible comparisons:

compare	*to*[51]	*comparison time*[52]
nonresponse study (N=201)	**EGLHS wave 1** **all persons** (N=2331)	**August 1991**
	EGLHS wave 1 **only not-panel-persons** (N=951)	**August 1991**
	EGLHS wave 2 **panel persons** (N=1380)	**March 1996**

Working Hypotheses

Exploratory work does not usually intend to test or to confirm hypotheses, but rather to "see into the data" (Ericksson/Nosanchuk, 1992). This is the first unprejudiced phase of data analysis. Adopting this approach, the following alternative hypotheses about the possible profile of the nonrespondents are intended to be working hypotheses. They shall act as a guideline for filtering out meaningful variables so that one can better distinguish between interesting and trivial information.

[51] Number of cases differ slightly from TABLE 1 (status of field report) due to subsequent data edition.
[52] Differing from the interview time, time-varying variables were taken for the earliest possible point in time available for all persons to be compared.

As was already discovered by the analysis of the field situation, the NRS study mainly contains an extract of problems of non-contact, but it also distinguishes the "1930 = refusing" from the "1960 = mobile" cohort. We are, therefore, interested in seeing which combination of variables characterises the two groups. It might even occur, however, that even under these conditions no differences in any of the analysed variables will be detected. The next table summarises the working hypotheses. I focus on the cohort category as this resulted from the field description. Additionally, gender differences will be discussed if necessary.

TABLE 6: Working Hypotheses on Profile of Nonresponse Study

	formulation of hypothesis	*what the data exploration would have to detect*
1	The NRS persons do not differ from the survey participants. They are considered to be a **random selection.**	No obvious differences from the original sample, neither for panel nor for not-panel persons. As a consequence, N=201 more interviews could simply be added to the EGLHS.
2	With a response rate of only 21%, having left out the strongest refusals once again, the NRS profile is **characterised by hard-to-contact people**.	A stable pattern of deviating variables can be found across all cohorts. They explain problems of contacting mobile persons.
3	**Mixture hypothesis:** NRS persons are not a homogeneous group; they are instead characterised by both: refusals and hard-to-contact problems. In particular, cohort 1930 contains refusals and cohort 1960 contains mobile persons.	Data exploration has to isolate a complex of variables and subgroups, which clearly separate cohorts 1930 and 1960, and has to explain different selection mechanisms.

2.3.2 Finding out Discriminating Variables

Although we cannot draw any conclusions from the pure comparison of marginal distributions to the structure of multiple interdependencies[53] (neither in the case of existence of differences nor in the case of non-existence), it is nevertheless a useful exploratory tool. It will be used in the group comparisons, which were introduced above, with respect to the two time points of the study. We need to look at the data from as many angles as possible. It is important to inspect whether there are once again gender differences and also whether a stable group of variables can be found to characterise the "mobile" versus "refusing" generation spectrum. I

[53] Schnell (1997:135-139) discusses this problem of the external validations of surveys with census data.

performed chi-square and t-tests[54], testing nonrespondents against persons of the main study in four ways: nonrespondents versus all wave 1 persons; nonrespondents versus wave 1, including only the non-panel persons; nonresponents versus wave 1, including only the panel persons; and finally nonrespondents versus wave 2 variables (panel persons). Each test was performed for the whole group as well as cohortwise and also for men and women separately. TABLE 7 shows the compressed results of the significance tests as coded lists.

TABLE[55] 7: Inspection of Significance Patterns - Chi-Square and t-Tests

nonrespondents versus		*wave 1 - all persons*							*wave 1 - only non-panel-persons*							*wave 1 - only panel-persons*						
1991 variables		all	c30	c40	c50	c60	m	w	all	c30	c40	c50	c60	m	w	all	c30	c40	c50	c60	m	w
family status 1991	*chi*	**	---	**	---	*	**	**	**	---	*	*	---	**	**	**	*	**	---	**	**	**
#children 1991	*chi*	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	**	---	---
"	*t*	**	---	---	---	**	*	*	**	---	---	---	---	---	*	**	---	---	---	**	*	---
school , highest degree	*chi*	**	**	**	---	---	---	**	**	**	**	---	---	*	**	**	**	**	---	---	---	**
vocational training	*chi*	**	**	*	**	---	*	**	**	**	**	---	---	*	**	**	**	---	*	---	---	**
#jobs until 1991	*t*	**	**	**	*	**	*	**	**	*	*	*	*	**	**	**	**	**	*	**	*	**
employed 1991	*chi*	*	---	---	---	---	---	---	**	---	---	---	---	*	**	---	---	---	---	---	---	---
unemployed 1991	*chi*	*	---	---	---	*	---	*	*	---	---	---	---	---	---	*	---	---	---	**	---	*
retired 1991	*chi*	**	---	---	---	---	**	**	**	---	*	---	---	**	**	**	---	---	---	---	*	*
employment status 1991	*chi*	**	---	---	---	---	*	*	**	---	---	---	---	*	**	**	---	---	---	*	---	---
position in job 1991	*chi*	**	---	---	---	**	**	---	*	---	*	---	---	---	---	**	---	---	---	**	*	---
party membership 1989	*chi*	**	**	---	---	**	**	**	**	---	---	---	---	---	*	**	**	---	---	*	**	*

		wave 2 - panel persons						
1996 variables		all	c30	c40	c50	c60	m	w
family status 1996	*chi*	**	**	**	---	*	**	**
#children 1996	*chi*	---	---	---	---	**	---	---
"	*t*	*	---	---	---	**	*	---
#residence changes 1989-96	*chi*	---	---	---	---	---	---	---
"	*t*	---	---	---	---	---	---	---
#changes if city 1989-96	*chi*	---	---	---	---	---	---	---
"	*t*	---	---	---	---	---	---	---
months in actual residence	*t*	---	---	---	---	---	---	---
#persons in household 1996	*chi*	*	---	*	---	**	---	*
"	*t*	---	---	*	---	**	---	---
single household 1996	*chi*	**	---	**	---	**	**	---
household income 1996	*t*	**	---	*	*	*	---	**
income per person 1996	*t*	---	---	---	---	---	---	---
#jobs 1989-96	*chi*	---		---	---	**	---	---
"	*t*	---		---	---	---	---	---
employed 1996	*chi*	---		---	---	---	---	---
unemployed 1996	*chi*	---		---	---	---	---	---
retired 1996	*chi*	---		---	---	---	---	---
employment status 1996	*chi*	---		---	---	---	---	---
position in job 1996	*chi*	---		---	---	---	---	---
self-esteem items: sestem5	*chi*	*	---	---	*	---	---	---
"	*t*	---	*	---	*	---	---	---
sestem2	*chi*	---	---	---	---	---	---	---
"	*t*	---	*	---	---	---	---	*
sestem8	*chi*	---	---	---	---	*	---	---
"	*t*	---	---	---	---	---	---	---
sestem9	*chi*	---	---	---	---	---	---	*
"	*t*	---	---	---	---	---	---	---
sestem10	*chi*	---	---	---	---	---	---	---
"	*t*	---	---	---	---	---	---	---
cati/capi interview method	*chi*	*	**	*	---	---	*	---

legend:

chi = chi-square-test
t = t=test (two-tailed)

c30 = comparison only within cohort 1930
c40 = only within cohort 1940
c50 = only within cohort 1950
c60 = only within cohort 1960
m = men only (all cohorts)
w = women only (all cohorts)

* = 5% significance
** = 1% significance
--- = no significance (lower than 5%)

job variables 1996 only for cohorts 1940,50,60

stable differences

bold: significant in at least 2 cohorts

no difference in any group

[54] For continuous variables. Count information and ordinal scales (level of education, self-esteem items) were treated both, as discrete and continuous variables.
[55] Imagining this table on a computer screen interactive highlighting of rows or columns would assist the data inspection. The table was produced "by hand" after calculating the tests in STATA. The exploratory tool of compressing grouped information into a condensed view is not yet enough implemented in standard software packages.

The list allows a simultaneous multiple comparison for each subgroup.
Horizontally, we are able to inspect whether differences are stable: e.g. the variable "jobs until 1991" differs significantly throughout all comparisons. In contrast to this, we see a clear complex with no significant results in any of the subgroups, e.g. "residence changes between 1989 and 1996". The horizontal inspection also shows that the significant variables do not indicate a homogeneous trend: e.g. the chi-square-test for the highest degree in school is significant for the complete group; yet the detailed inspection only shows this result for cohort 1930 and 1940. The significance pattern also differs by gender. Vertically, the pattern table allows one to find out the relevant variables separately within cohort or gender.

The following overview groups the results according to the stability of discrimination. Afterwards, in the final section of this chapter, the main findings will be presented from the qualitative aspect.

Non Discriminating Variables

residence[56]:

- number of residence changes
- number of city changes between 1989 and 1996
- duration of residence in present home 1996

do not show any significant differences in any subgroup. Apparently, the NRS persons do not differ in their residential mobility during the transformation years (on the basis of the realised interviews).

income:

- household income is significant, but this is not the case for
- household income per person[57].

It can be assumed that "number of persons in the household" / "single household 1996" are the controlling variables.

job and career:

For 1996 we cannot detect differences in the labour market variables

- number of jobs between 1989 and 1996[58]
- being employed or unemployed
- in retirement
- employment status (combination of employment/retirement/not working)
- position in job (worker, employee, self-employed).

The tests for the 1996 job variables in TABLE 7 were only done for cohorts 1940,50,60.

[56] Because the 1996 interview only requested variables about changes in residence, the actual place of residence is not available. Thus, no rural/urban comparison was possible.
[57] Calculated as household income divided by number of persons.
[58] The only significance is in cohort 1960 using the chi-square test.

self-esteem items[59]:
There are only few significant results on a 5%-level for some subgroups. Chi-square tests were done, taking the 7-point-scale as categories; t-tests assuming it as the interval scale. The overall impression is that the nonrespondents are similar to the main study (wave 2) persons. So far the nonrespondents do not show a deviant profile in the psychological variables.

Stable Differences

The variable "number of jobs until 1991" (reported job-spells) is significant in each subgroup.

Cohort and Gender Specific Differences

Generally, the least amount of deviations is given for cohort 1950. It turns out that the marital status is important for each cohort in both observation years. We end up with a cohortwise list of differing variables.

Cohort 1930:
- marital status 1991 and 1996
- level of education in school
- level of vocational training
- number of jobs up until 1991
- party membership 1989
- interview method (personal as opposed to telephone interview)

Cohort 1940:
- marital status 1991 and 1996
- single household 1996; number of persons in household 1996; household income 1996
- level of education in school
- level of vocational training
- number of jobs up until 1991
- position in job 1991
- retirement 1991
- interview method

Cohort 1950:
This cohort shows the fewest differences in the nonrespondents. There are only:
- marital status 1991
- number of jobs 1991
- level of vocational training
- household income 1996 (but not per person).

Cohort 1960:
- number of children 1991 and 1996
- number of persons in household; single household 1996; household income 1996
- unemployment 1991; employment status[60]
- number of jobs up until 1991 and between 1989 and 1996

[59] Items like e.g. "I wish I could have more self-respect" using a 7-point scale ranging from "totally disagree" to "totally agree".For the complete wording of the items sestem5,2,8,9,10 see the codebook, appendix A.
[60] Contains unemployment as a category.

- position in job 1991
- party membership 1989.

It has to be emphasised that the list also contains variables which differ in only one test. One should not over-interpret the level of significance, particularly for cohort 1930. With only N=26 persons in cohort 1930 (out of a total N=201 nonrespondents) a chi-square test could sometimes switch from being significant to not significant[61] even if it was only a single person who changed the cell. It is more important to give meaningful substantial interpretations for the nonrespondents' profile.

As supplementary information, appendix B contains a cohortwise collection of descriptive graphs for the listed variables. In the following part, I will concentrate on the more general findings.

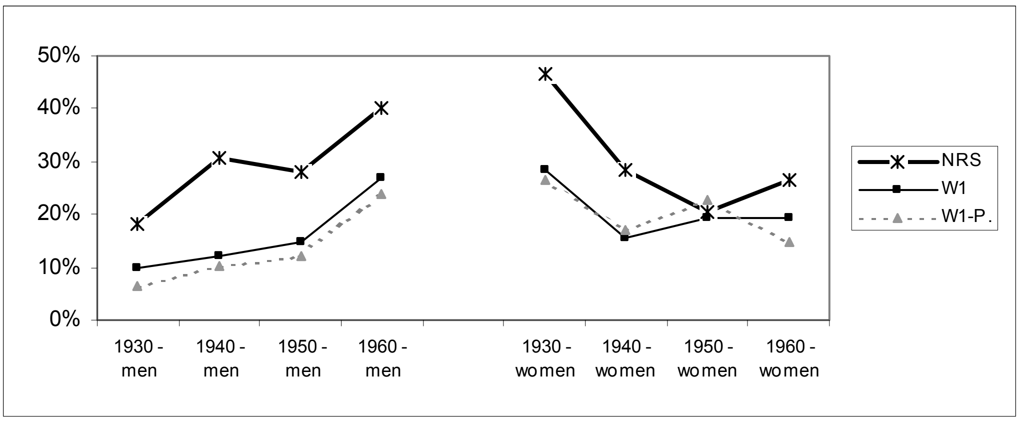
2.3.3 Summary of the Nonresponse Data Exploration

According to the thematic groups of the analysed variables, I will describe the nonrespondents according to the topics family, education, labour market and politics.

Family

The most stable result across cohorts and gender is the overproportional part of unmarried persons, mainly never married. The structure of deviations in the marital status is not the same for men and women. FIGURE 15 shows the proportion of unmarried persons across cohorts as well as for men and women separately with respect to the time of the initial sampling 1991. Two lines are given for the main study: for the total wave 1 sample and for the wave 1 subset containing only later panel participants.

FIGURE 15: Percentage of Unmarried Persons 1991 per Cohort and Gender



unmarried = never married + widowed + divorced

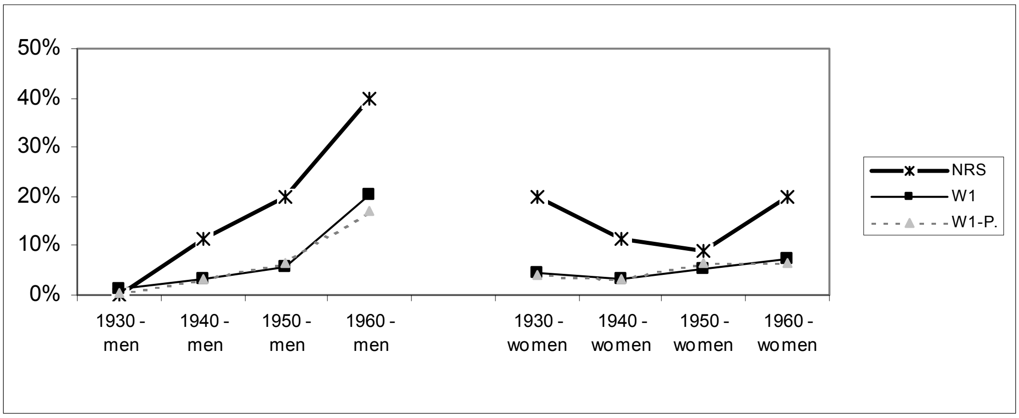
---

[61] Taking at least a 5% level.

We see for men that the proportion of unmarried[62] nonrespondents is clearly higher in all cohorts. The panel line is slightly lower, which reflects that it is easier to contact married men than the unmarried. The structure for women is different. The 1950 women are almost the same in both studies and the panel participation does not generally reveal the correlation with the "unmarried" status.

The next figure restricts the marital status to the category "never married".

FIGURE 16: Percentage of Never Married Persons 1991 per Cohort and Gender



Except for the 1930 men, all subgroups of the nonrespondents have more never married persons. We witness once again that the change across cohorts is gender specific. There is an increasing percentage of never married men over the generations, whereas a decreasing/increasing line for the women with a point of inflection for 1950. There are no remarkable deviations in the panel line. It can be clearly recognised, however, that the nonrespondents retain the global trend across cohorts, although at a generally higher percentage of never married persons. The analysis of the marital status 1996 produces the same picture[63]. Due to the fact that we only compare panel participants to the nonrespondents, the discrepancy is even higher.

For the household size 1996 we have a pattern similar to the graph for the unmarried persons: there are more single nonrespondents. Again the 1950 cohort does not differ.

[62] Being "unmarried" does not imply that all of the persons live in single households. For the youngest group in particular, we have an increasing proportion of persons living together with partners, but who are not married. The life history study also asked for these forms of cohabitation. To overcome possible problems concerning retrospection or comparability across cohorts, only the stable official information about marriages/divorces/death of spouse was employed here.

[63] Therefore, no extra figure here.

The number of children is an important criterion for the distinction of younger people. What's more, however, 1960 nonrespondents have no children (32% NRS men vs. 11,3% panel men/ 16,7% NRS women vs. 4,9% panel women).

Summary: Nonrespondents tend to be unmarried (mainly never married), single and the younger persons have no children.

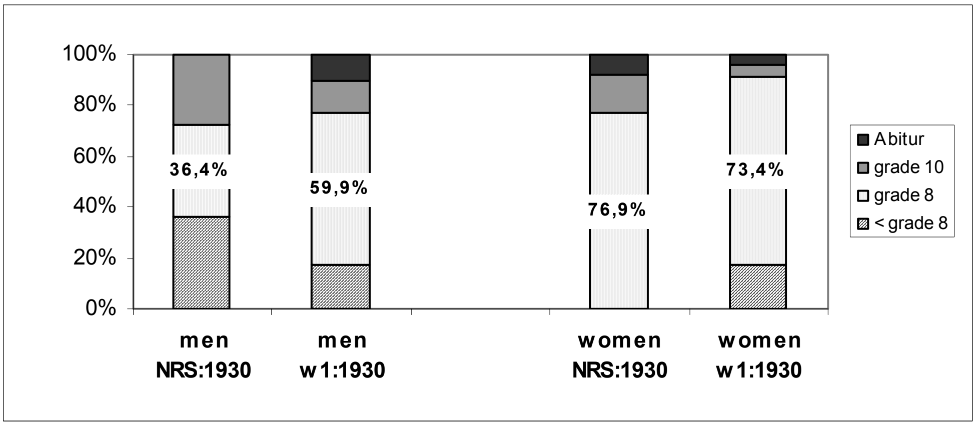
Education

The degree of education in school and the level of vocational training distinguishes nonrespondents from survey participants in both of the oldest groups, although this is no longer the case for the younger generation. Of course this reflects that a historical change occurred with regard to receiving better education. Whereas the older persons experienced school and vocational training during or shortly after the Second World War, younger persons could profit from better education opportunities: this was particularly the case for women[64]. As a consequence, the nonrespondents' educational profile cannot be described as one general trend.

The degrees in school and vocational training[65] do not differ remarkably for either men or women in the younger cohorts 1950 and 1960. The 1940 cohort shows diverging results: the NRS men are better qualified, whilst the 1940 NRS women have lower qualifications.

The NRS profile is a particularly special one for the 1930 cohort. I will discuss it in more detail as a result of the connection to the refusal hypothesis. FIGURE 17 shows the structure of the school variable for men and women.

FIGURE 17: Educational Degree in School for Cohort 1930 - Nonresponse versus Main Study



[64] The historical change of the educational structure and the career system in the GDR is described in: Huinink/Mayer/Trappe (1995:106-111) Bildungs- und Berufsverläufe. In: Huinink/Mayer (1995). Kollektiv und Eigensinn.

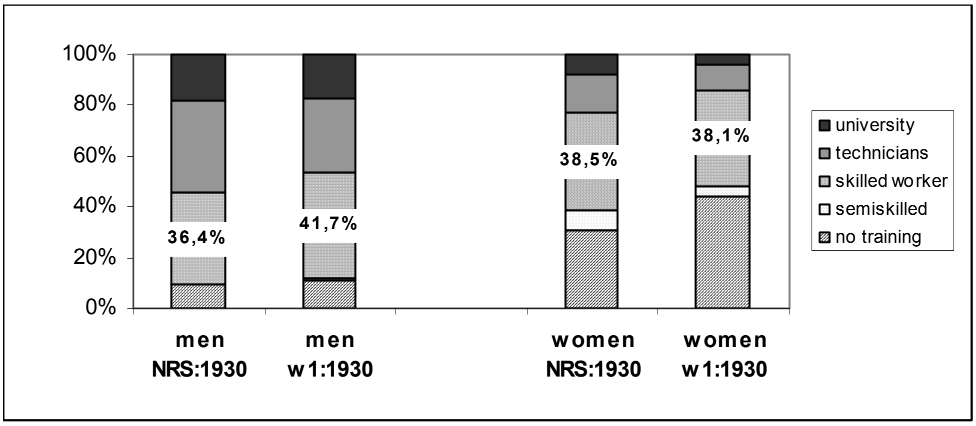
[65] A weak trend towards a lower level of vocational training in the 1950 cohort.

The percentages for the category "grade 8" (=8 years of school having been successfully completed) are marked by numbers. This degree was the standard case during the historical time of the generation in consideration. We see that the 1930 NRS women are generally better qualified. This is due to the absence of the lowest category and the greater part of the highest levels.

The picture of the NRS men is characterised by a reduction of the "standard" category (grade 8) from both directions: there is an overproportional part in the lowest category, whereas the highest level is missing. It is remarkable, however, that the 10 year level (which was still not the standard case) has a higher proportion. I, therefore, doubt that these findings can be interpreted as a generally lower level of education of the 1930 men. If we assume, in accordance with the "refusing generation hypothesis", that at least some information about soft refusals can be discovered in the data, this might have another explanation. The fact that the NRS sample missed the highest levels of education indicates that indeed the hard-core refusals stem from this group (academic and system-stable jobs). The survey refusals of the 1930 cohort, having lived their entire working life in the former conditions, could not now expect new conditions for their careers. Thus they did not see any use in participating in a survey. The realised NRS sample, though missing the highest school degrees, contains converted refusals who also show the trend to better qualification (i.e. higher percentage of grade 10).

This is purely speculative reasoning since we do not in fact actually possess this individual information. Regarding the level of vocational training, however, we see that the NRS men, in spite of the school differences, are relatively similar to the main study (see the next figure). The NRS women are generally better trained (more university and technical degrees with fewer women in the "no training" category).

FIGURE 18: Level of Vocational Training for Cohort 1930

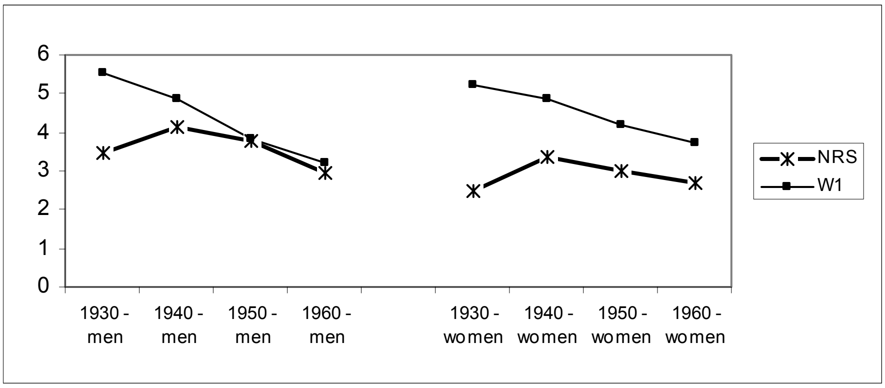


Summary: *All in all we can describe the 1930 cohort as the well qualified stable "middle". It indicates that those people who had benefited from the old system and who had completed their active working life, were more likely to refuse to participate in the survey in the new unstable situation. The hardest refusals appear to be the group with the highest educational degrees.*

Labour Market

The number of jobs up until 1991 was found to be a stable discriminating variable for all cohorts (see significance pattern in TABLE 7). The following graph allows a qualitative look at this result. It shows the average numbers of reported jobs in the careers of men and of women. In general, the nonrespondents have fewer jobs, counted as job-spells up until 1991 (which is the censoring time for the initial sample). The lines for NRS and main study (wave 1)[66] converge across cohorts; for men 1950 and 1960 they come together. The variables are the most distinctive for the 1930 nonrespondents. This is, on the one hand, a consequence of longer education periods; on the other hand it reflects more stability in the careers which emphasises this cohort's stable position in society.

FIGURE 19: Average Number of Jobs until 1991 per Cohort and Gender



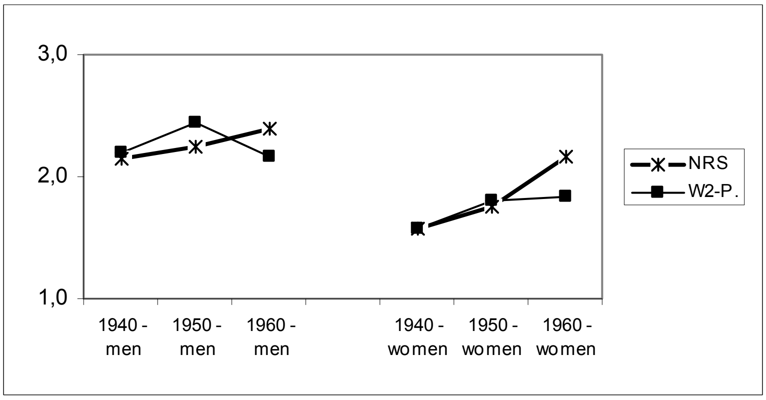
I additionally checked this variable in more detailed subgroups, namely defined by cohort, gender and school degree. The average number of jobs up until 1991 in general was lower for the nonrespondents, the t-tests, however, were significant for only few subgroups ("men, cohort 1930, lowest degree in school"; "women, cohort 1950, highest degree in school"; "women, cohort 1960, highest two degrees"). But I have to add that the absolute number of cases is decreasing when breaking down the sample into subgroups. Another interpretation for the difference of the number of jobs might be given by the aspect of retrospection. The main study (wave 1) had inquired the job spells up until 1991 at the actual interview time 1991,

[66] The wave 2 information is left out as it is very similar.

whereas the NRS study gathered this information five years later. Without doubting fundamentally on the data quality[67] of this retrospective study, the unstable transformation years between 1991 (wave 1 interviews) and 1997 (NRS interviews) might have influenced which kind of jobs were properly recalled and reported.

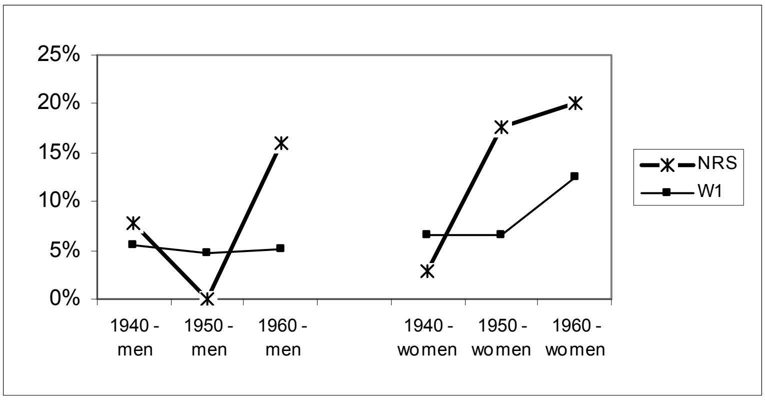
To the same extent that educational and stability explanations lose importance for the younger cohorts, we can also recognise a trend towards greater influence of labour market conditions on younger people. Contrary to the better situation of the 1930 nonrespondents, however, the 1960 group is characterised by more unfavourable conditions. I will concentrate on two main findings: the number of jobs between 1989 and 1996 and unemployment for cohorts 1940 to 1960[68]. If we inspect the average number of job-spells between the years 1989 and 1996, we find that the youngest persons had more jobs during this period. They had to adopt new economic and labour market conditions.

FIGURE 20: Average Number of Jobs 1989-1996 per Cohort and Gender



It turns out that the missing mobile target persons might have moved, maybe only temporarily, in order to find a job. This is reinforced if we inspect the percentage of unemployed persons for 1991 in the following figure.

FIGURE 21: Percentage of Unemployed Persons 1991 per Cohort and Gender



67 See arguments above about event oriented design, data editing,etc.
68 1930 persons are excluded because most of them went into retirement in the beginning of the 90s.

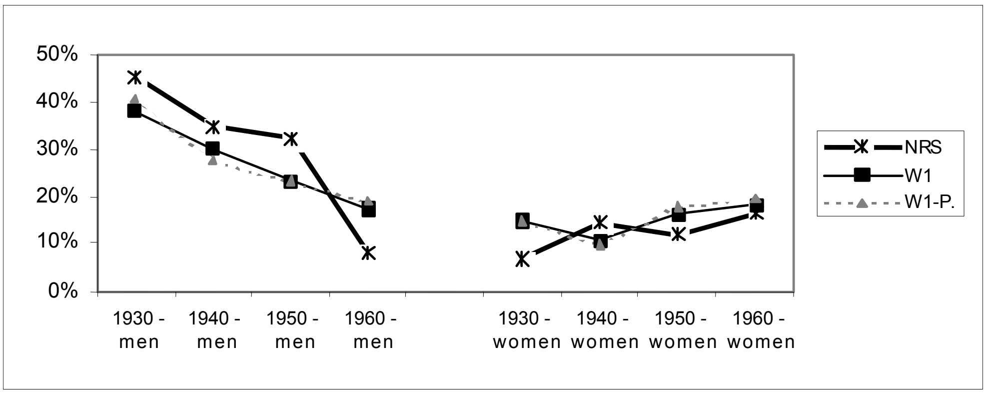
It shows the highest percentage of unemployed nonrespondents in the youngest group. Being unemployed already in 1991, in the early phase of the transformation, obviously forced younger persons, to a great extent, to be mobile[69]. It is also generally obvious, that younger women experienced career disadvantages in comparison to men. Another result which reflects the problems of contacting mobile persons is the extreme gap between the percentage of self-employment in the young cohort 1960: this category is completely absent in the NRS file, whereas in the main study (1996) there is about 9%.

Summary: *The labour market transformation separates two groups of nonrespondents. The oldest nonrespondents (1930) have stable careers and fewer jobs up until 1991; the youngest nonrespondents (1960) were forced into mobility and have more jobs since 1989.*

Politics

The complex of the psychological and political variables was chosen in order to see whether there is a connection to refusing. It turned out that the selected self-esteem items did not distinguish nonrespondents from survey participants. The field report (infas(1995:16)) gives hints about political attitudes playing a role in refusals. A ranking of the most frequently mentioned refusal reasons reveals three outstanding reasons (with equal frequencies). One of the top three[70] is dissatisfaction with the political and social situation, "nothing will be changed". The available political variable for nonrespondents as well as for survey persons is party membership related to the year 1989[71]. The comparisons between the NRS and the main study were significant for the two opposite cohorts 1930 and 1960. The next figure shows the percentage of "SED" members.

FIGURE 22: Party Membership in the SED per Cohort and Gender



[69] For example, during the official status of being unemployed, they might have participated in qualification courses, even in other cities.

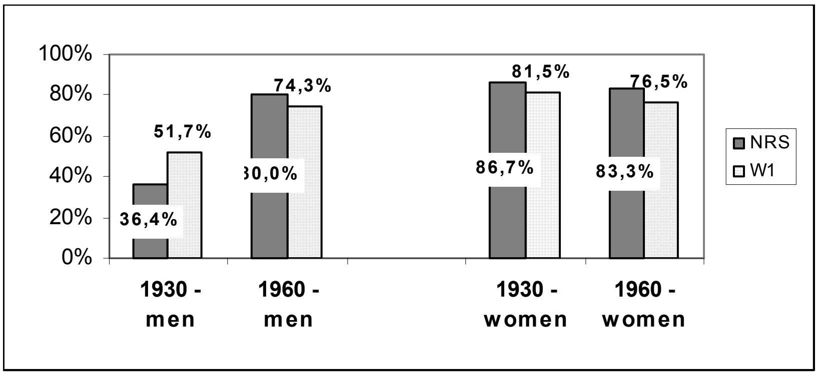
[70] The other two reasons are: "not interested in surveys in general" and "target persons were influenced by husband/wife".

[71] Means: ever being party member until 1989.

The "SED"[72] was the communist party in the GDR. The overall proportion of SED members in the complete sample is about 21%. There were also several other parties[73] under the GDR regime with an overall average of 4,7%. The majority of the interviewed persons had not been members of any party (72,4%). We can recognise from FIGURE 22 that women participated in official political life to a far lesser extent than men did. The differences between nonrespondents and main study are not very high, except for the oldest women, where the NRS curve is below the main sample. Among the older male nonrespondents, however, there are more SED members. This is the exact opposite for the youngest group.

In the next figure, we see the proportion of persons who had never been in any party until 1989.

FIGURE 23: Percentage of Persons Not Members of Any Party 1989 - Cohorts 1930 and 1960



The percentages of women (1930 and 1960) and of the 1960 men are slightly higher, although not by a great deal. In fact more than three quarters of the young men and of the women in general had never participated in any party: in the case of the male nonrespondents only 36% had done so. It has to be mentioned that, on the other hand, the mere party membership cannot be considered as 100%-equivalent to a convinced and active participation in a state's political life. Particularly persons who wanted to obtain better careers often decided to become members of the communist party (SED) just for the sake of convenience.

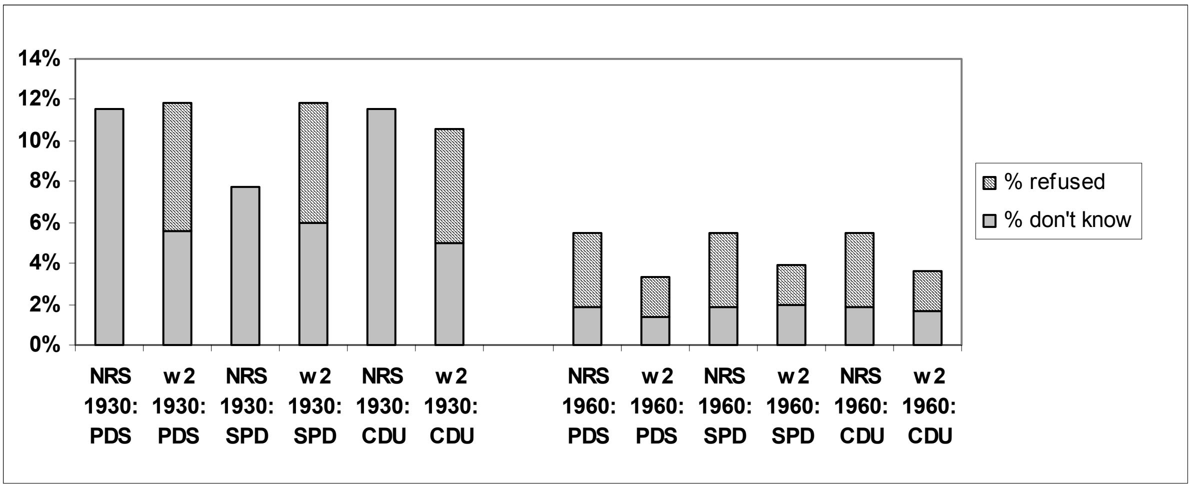
To obtain some more information about the political profile of the nonrespondents, I finally decided to examine some additional variables measuring attitudes towards political parties. The wave 2 and the NRS study had inquired the sympathy for the political parties in the German parliament. The party attitude items used a scale ranging from -5 ("no sympathy at all") to +5 ("highest amount of sympathy"). Comparing the mean values of the valid answers explicitly by cohort and gender, no remarkable differences between the panel and

[72] SED=**S**ozialistische **E**inheitspartei **D**eutschlands

[73] Other parties = CDU (during the GDR era); LDPD; NDPD; DBD; block parties. The codebook, see appendix A, also has some marginal cases for the SPD from the early time of establishing new parties.

NRS participants could be detected. However, the examination of the categories "don't know" and "item refused" turned out to be more interesting. The following figure shows the percentage of these categories for the two cohorts 1930 and 1960.

FIGURE 24: Percentage of "Don't know" and "Refused" in the Party Attitude Scales - Cohort 1930 and 1960 -



The graph shows the percentages for the three selected parties according to a left to right spectrum from the "PDS" (the successor of the former communist party), "SPD" (the social-democrats) to the "CDU" (the christian democrats). In Cohort 1960, the "don't know" category is constant, the item refusal proportion is slightly higher in the NRS study. Contrary to that, we see a different picture for cohort 1930. Instead of item refusal the NRS persons chose the "don't know" alternative. Counting together the "refused" and "don't know" percentages in the panel study, we find almost the same proportion (at least for the parties "PDS" and "CDU").

Although I have to leave out here a specialised analysis of party attitude measurement, the findings for cohort 1930 fit to the impression: the NRS study is missing the hard-core-refusals containing, however, a proportion of soft refusals in cohort 1930 who tend to the more convenient answers. The "convenience" attribute for the soft refusals in cohort 1930 is assisted by Huinink's[74] (1995:41) characterisation of a part of the GDR citizens. He describes them as participating in a common "freerider" strategy, which means they use the state's advantages without accepting a personal duty for a positive engagement in the society.

[74] Huinink calls it in German: "Trittbrettfahrer"-Strategie which I translated to "freerider" strategy. In : Huinink, J. (1995) Individuum und Gesellschaft in der DDR. In: Huinink/Mayer (1995:41). Kollektiv und Eigensinn.

Summary: The party membership during the GDR era distinguishes two groups: younger persons and women in general did not participate; the nonrespondents differ only slightly with a tendency to lower participation. The male nonrespondents of the older cohorts were engaged in political parties to a higher degree. We, therefore, have the result:

- *The mobile nonrespondents do not differ from the survey persons with regard to the political aspect. They generally tend towards lower party participation.*
- *The political participation was much higher for male nonrespondents of the older cohorts, as well as for men of the "refusing" 1930 generation in particular.*

General

All in all the nonresponse file is a biased extract, predominantly produced by filtering out persistently hard-to-contact persons. Some of them might have been former "soft" refusals, but the "hard-core-refusals" are once again excluded. The NRS data is characterised by:

- a noticeable loss of cohort 1930
- a loss of men in the older age groups
- a remarkably high proportion of unmarried persons, mainly never married persons.

Educational variables separate the NRS from the main study for the older cohorts. The older nonrespondents tend to be better educated. The younger nonrespondents differ with regard to labour market mobility.

This analysis of the reasons of loss in the sampling process together with the comparison of the realised nonresponse and survey interviews resulted in the description of

- **cohort 1930 as the "refusing generation"** and
- **cohort 1960 as the "mobile generation".**

This means that the "mixture hypothesis" about the profile of the nonresponse study makes sense. In the following, I will focus on the two polarising cohorts 1930 and 1960. Having discussed the nonrespondents of the East German Life History Study variable by variable, the next chapter will turn to multivariate relationships. It shall be analysed whether the collected explaining variables can predict nonresponse when interacting together in a regression model.

3. Predicting Nonresponse

This chapter discusses how the identified characteristics are able to predict whether a person is in the main study or in the NRS study. We particularly want to detect the structure of the predictors and their relative importance in a higher dimensional context. We cannot conclude this from one-dimensional distributions, so the task will consequently be performed by a multivariate regression model. According to the hypothesis, *"Cohort 1960 stands for hard-to-reach people"*, I expect that non-contact predictors play a significant role for this cohort. If the hypothesis, *"Cohort 1930 stands for refusals",* holds, significant effects of refusal indicator variables should appear.

Since the nonresponse study does not contain the individual information about who was a refusal and who a non-contacted person in the initial sampling, I have to adopt an indirect approach. It is necessary to reason about which kind of conclusions we can draw from the model by variables which explain participation in the nonresponse study. Are the explanations for persons who are in the NRS file also suitable for a generalisation about the non-contacts and the refusals?

1) Generalisation about non-contacts:

It is reasonable to take the non-contact explanations and to draw conclusions from the people in the NRS study about the hard-to-contact people in general. The argument is that the biased nature of the study is a filtered extract of the hard-to-reach persons. The study undertook additional efforts and consequently succeeded in getting interviews with exactly these very persons.

2) Generalisation about refusals:

It is not reasonable to proceed in the same way for conclusions about refusals because most of them are in fact not in the study. Explanatory variables for refusals might be found for cohort 1930 in the sense that only meek refusal tendencies are explained. This means that it is not those who are in the study that tell us most about refusals, but rather those who are extremely underrepresented.

The models can be done from two points of view which correspond to the design of the study. Attention will first be paid to the time of the first wave in 1991. This was the original sampling situation in which the loss of target persons occurred. Attention will next be paid to the time when wave 2 was performed in 1996, as additional data is available for panel persons as well as for nonresponse persons. The data will be censored at the comparable point in time: in 1996. The panel participants are the best persons from the empirical researcher's point of view, because they could easily be contacted and were willing to give interviews twice. According to the concept of random sampling, the nonrespondents also had to originally be in the pool of the virtual ideal sample. Can a model predict who is a "dark chapter person" or who is the "ideal panel participant"?

I do not expect extremely strong models of prediction as this would mean that the main study had systematically excluded a number of individuals. Empirical projects in the social sciences undertake efforts to avoid this result by careful sampling designs and control of the fieldwork[75]. The special problem of nonresponse lies in the fact that effects of nonresponse are not easy to assess. I will, therefore, argue that if nonrespondents are not a 100% random subsample, the prediction models will produce a potentially weak explanation, which is better, however, than pure guesswork. Four logit[76] models with a dichotomous dependent variable (1=being in the nonresponse sample/ 0=being in the main study) will be presented for both cohorts 1930 and 1960 and then related to the two points in time of the study.

3.1 Predicting Nonresponse for Cohort 1930

I will begin with cohort 1930 which was entitled "the refusing generation". The variables for the prediction models were chosen for two principle reasons. Theoretically, the selection is in accordance with the relevant findings of the exploration chapter. Practically, the number of variables had to be small since more parsimonious models are easier to interpret and increasing the number of parameters can raise the general goodness of fit without being meaningful. On account of the general characteristics of the nonrespondents and of the special findings for this cohort, the following variables were kept in the model:

variables for cohort 1930	***expected effect on the likelihood of being a nonrespondent***
gender: men	negative
married	negative
medium level of education: grade 10	positive
number of jobs up until 1991	negative
party membership	positive
capi interview (1996)	clearly positive

75 This means: clearing out-of-date/incorrect addresses, finding persons who have moved, several attempts at contacts, interviewer control etc.

76 Regression models for a binary dependent variable can be performed by logit or probit regression assuming slightly different distribution functions. In practice both models are reported to be very similar so that results are not expected to change substantially. See: Aldrich/Nelson (1984:30-35) Linear Probability, Logit, and Probit Models. I took "logit" from STATA 6.0.

TABLE 8 shows the estimates of two logit models: one predicting NRS participants against the initial sample 1991, the other one predicting NRS participants against the panel interviews 1996.

TABLE 8: Logit Models Predicting Being in the Nonresponse Study for Cohort 1930
- Wave 1 and Wave 2 -

		Logit Estimates for Cohort 1930			
		Wave 1 / 1991		***Wave 2 / 1996***	
dependent variable	**independent variables**	**coefficient** ***(std. error)***	**P>|z|**	**coefficient** ***(std. error)***	**P>|z|**
NRS (in wave 1)					
NRSPAN (in wave 2)					
	men	-0,148	0,751	-0,272	0,587
		0,466		*0,501*	
	married91	-0,783	0,085	---	---
		0,454			
	married96	---	---	**-1,253**	0,012
				0,496	
	smiddle	**1,098**	0,043	0,997	0,088
		0,543		*0,585*	
	njob91	**-0,244**	0,009	**-0,205**	0,025
		0,093		*0,091*	
	anyparty	0,280	0,565	0,544	0,309
		0,486		*0,534*	
	capi	---	---	**1,770**	0,000
				0,470	
	constant	**-1,635**	0,001	**-1,701**	0,005
		0,505		*0,610*	
	valid observations N=		605		338
	nonrespondents N=		25		25
	survey persons N=		580		313
	Likelihood ratio chi2 =		14,89		32,94
	Prob > chi2 =		0,011		0,000
	% correct predictions by model =		95,8%		92%
	% correct predictions by guess =		91,9%		86%

Significant coefficients at 5%-level are given in bold.

Variables are:

NRS = indicator of study: 1=person is in nonresponse study
0=person is in main study/ wave 1
NRSPAN = indicator of study: 1=person is in nonresponse study
0=person is in main study/ wave 1
men = indicator of gender 1=men/ 0=women
married91 = indicator of being married 1991 (1=yes;0=no)
married96 = indicator of being married 199 (1=yes;0=no)
smiddle= indicator of middle qualification in school (1=grade 10/ 0 otherwise)
njob91= number of reported job-spells up until 1991
anyparty = indicator of party membership up until 1989
(1=SED or block party/ 0=no party)

The Model for 1991

The likelihood ratio test is significant (LR chi2=14,89 with p=0,011) which means that the model is better than the simple estimation of the constant term. There are doubts whether it is useful to interpret the value of the pseudo-R^2 which the programme calculates. Aldrich/Nelson (1984:58) emphasise that there is no generally accepted measure for the goodness of fit in logit models. As an additional description for the goodness of the model, I have therefore calculated the percentage of correct predictions of being a nonrespondent using the model probabilities[77]. As a point of comparison, I also calculated a best guess[78] prediction by taking just the empirical probability of being in the NRS sample. The model percentage is slightly better than the best guess, although not by a great deal (95,8% versus 91,9%).

If we examine the coefficient estimates, we find that they are showing the direction as expected although the variables "men" and "any party" are not significant. We cannot conclude from these multivariate interrelations, therefore, that there are substantial gender differences. We are additionally not on the secure side of error if we interpret the political party variable as increasing the risk of nonresponse.

> As a background information for the interpretation of the party variable I checked how many of the party members had active positions in the party. Although there are more party members in the NRS study for cohort 1930 (especially men, see FIGURE 22 in chapter 2), there are only few single cases who had a position in the party. Contrary to this, about 25% of the 1930 party members in the main study worked actively as a functionary. This result again confirms that cohort 1930 in the NRS study missed the hard-core refusals, it contains, however, those persons who refused by convenience.

We might state, however, that being married (family variable "married91") decreases the risk (10% significance). The most important variables, both of which are 5%-significant, are the school variable "smiddle" and the job variable "njob91". The educational variable represents the grade 10 degree, which is the well established middle position of education. We have the result that better education (here: in the sense of a grade 10 degree) increases the risk to be in the NRS file, whereas a high number of jobs up until 1991 decreases the risk. This confirms that in the 1930 cohort, persons with stable careers during the GDR era tended to be in the nonresponse study. In addition to which it must be mentioned that the educational indicator "high school degree (Abitur)" and the job category "self employed" could not be included in the model. Since these categories are virtually absent in the NRS study, the model cannot be estimated properly due to numerical reasons of absent variance.

There are no explicitly high risks of nonresponse[79] for the 1930 cohort, although we can look at the relative importance of the coefficients. Due to the fact that they cannot be

[77] Predicted probabilities p<0,5 were treated as "failure: not in NRS", whereas p>=0,5 were treated as "success: being in NRS".
[78] According to the empirical binomial distribution, p(NRS=1, i.e. "being in nonresponse file") was calculated. Persons from the main sample and the nonresponse study were assumed to be nonrespondents by the same probability.
[79] Which I interpret here as the risk of being a participant in the NRS study.

interpreted intuitively as in the case of linear models, TABLE 9 (see below) lists examples of predicted probabilities within selected groups.

Women generally have slightly higher values than men; party members higher than non party members. The highest risk values are about p=0,2. If one takes the case of e.g. the influence of the marriage variable, one can recognise that for e.g. "men with grade 10 degree, no party membership, 4 jobs up until 1991", the risk for married men is 0,080 but that this nearly doubles to 0,156 for unmarried men with the same attributes. The relative influence of the school variable is even higher: e.g. "unmarried women, party members with 4 jobs" have a risk of only 0,089 with regards to being in the nonresponse file, given that they do not have a grade 10 degree. This increases to 0,226 for the same women when the school variable is changed to grade 10.

TABLE 9: Predicted Probabilities for Nonresponse in Cohort 1930 / Initial Sampling 1991

Cohort 1930 / Initial Sampling 1991						
			predicted probability for			
likelihood of being			no party membership		in SED or block party	
a nonrespondent is	in group	with	**MEN**	**WOMEN**	**MEN**	**WOMEN**
lowest	married 1991 and	number of jobs up until 1991= 5	0,022	0,027	0,029	0,034
	not grade 10 (school)	4	0,028	0,032	0,037	0,043
	married 1991 and	number of jobs up until 1991= 5	0,064	0,073	0,083	0,094
	grade 10 (school)	4	0,080	0,091	**0,103**	**0,117**
	unmarried 1991 and	number of jobs up until 1991= 5	0,047	0,054	0,062	0,071
	not grade 10 (school)	4	0,060	0,068	0,077	0,089
	unmarried 1991 and	number of jobs up until 1991= 5	**0,130**	**0,147**	**0,164**	**0,186**
highest	**grade 10 (school)**	4	**0,156**	**0,180**	**0,201**	**0,226**
		mean value for #jobs in complete group is 5,3	*p > 0,10 in bold*			

(Predictions were also calculated if combination of values was not observed.)

The Model for 1996

This model is generally better than a simple guess. We have a significant likelihood-ratio test (LR chi2=32,94 with p=0,000).

The marriage variable is actualised[80] for the year 1996. The additional variable is the interview method "capi". The general tendency is that the wave 2 model shows a similar significance structure. Also when modelling the NRS participation against the panel sample, gender and party membership have no significant coefficients. The lower risk for married persons is confirmed. The school variable (indicating grade 10) is only 10% significant. The

[80] Apart from more persons being widowed, this does not in fact mean that there are relevant changes for the cohort under consideration at the age of approximately 66 years.

essential meaning of the job variable is retained: this states that persons with fewer jobs are more likely to be nonrespondents.

The interview method "capi" is a highly significant predictor for the nonresponse study. This seems to assist the "refusing generation" theory, as the general availability of the telephone can be assumed to be equal within the cohort. A special selection mechanism might instead be the explanation: the capi method is the reason for the successful conversion of soft refusals.

The calculated predicted probabilities for this model are shown in the next table.

TABLE 10: Predicted Probabilities for Nonresponse in Cohort 1930 / Wave 2 1996

Cohort 1930 / Wave 2 1996							
				predicted probability for			
				no party membership		in SED or block party	
likelihood of being a nonrespondent is	in group	***keeping number of jobs up until 1991 at njob91=5***		**MEN**	**WOMEN**	**MEN**	**WOMEN**
lowest	married 1996 and not grade 10 (school)	interview method is:	cati	0,014	0,018	0,024	0,031
			capi	0,077	0,098	**0,126**	**0,159**
	married 1996 and grade 10 (school)	interview method is:	cati	0,037	0,048	0,062	0,080
			capi	**0,185**	**0,229**	**0,281**	**0,339**
	unmarried 1996 and not grade 10 (school)	interview method is:	cati	0,048	0,061	0,079	**0,101**
			capi	**0,227**	**0,278**	**0,335**	**0,398**
	unmarried 1996 and grade 10 (school)	interview method is:	cati	**0,119**	**0,151**	**0,189**	**0,234**
highest			**capi**	**0,442**	**0,510**	**0,577**	**0,642**
		mean value for #jobs in complete group is 5,3		*p > 0,10 in bold*			

(Predictions were also calculated if combination of values was not observed.)

To keep the table clear, the number of jobs up until 1991 is kept at 5, which is approximately the average number (5,3). The table reveals the great influence of the capi variable. Given a personal interview, the probability of being in the NRS file is almost four times as high as for telephone interviews (e.g. "men, unmarried, no party, grade 10" p=0,119 for cati, but p=0,442 for capi).

Summary for Cohort 1930:

Being unmarried and having stable careers explain the higher risk of being in the nonresponse study. Participation in political parties is not significant, whereas the most predictive power lies in the personal interview method. This combination of attributes can be interpreted as a weak confirmation of the refusing theory, since particularly the capi-variable seems to be related to the refusal rate. As the general predictive power of the model is rather weak, however, we have no strong forecast for nonresponse.

3.2 Predicting Nonresponse for Cohort 1960

As before, the selection of the variables for the model is motivated by the exploratory results. For this cohort, "the mobile generation", the family and labour market variables are expected to be important. For the family status, I chose the dichotomous variable "never married". Additionally, the process of founding a family will gain importance in the sense that people without or with fewer children are more likely to be nonrespondents. To reflect the labour market influence, the variables "being unemployed 1991", "number of jobs up until 1991" and "number of jobs between 1989 and 1996" were chosen. Contrary to the oldest cohort, educational variables were not different in the exploration chapter. Also the capi variable is non discriminating. Though the party membership variable showed that the NRS study has a higher proportion of persons who are not members of any party, I did not select the political variable for the model. Following the analysis for this cohort, there were no arguments which could justify the assumption that a special political profile was linked to the hard-to-reach characterisation. (Indeed tests with the party variable in several models never gave significant estimates.) Though no gender differences are expected, I kept the variable "men" in the model as the control variable in the multiple interaction.

The following variables are included in the presented models:

variables for cohort 1960	***expected effect on the likelihood of being a nonrespondent***
gender: men	negative, but ought not to be significant
never married 1991 and 1996	positive
number of children 1996	negative
unemployment 1991	positive
number of jobs up until 1991	negative
number of jobs 1989-1996	positive (opposite direction of njob91)

The corresponding logit estimates of the models for cohort 1960 are given in the next table.

TABLE 11: Logit Models Predicting Being in the Nonresponse Study for Cohort 1960 - Wave 1 and Wave 2 -

		Logit Estimates for Cohort 1960 *Wave 1 / 1991*		*Wave 2 / 1996*	
dependent variable	**independent variables**	**coefficient** *(std. error)*	**P>|z|**	**coefficient** *(std. error)*	**P>|z|**
NRS (in wave 1) NRSPAN (in wave 2)					
	men	-0,236	0,437	-0,434	0,181
		0,303		*0,325*	
	nevmarri91	**0,832**	0,015	---	---
		0,341			
	nevmar96	---	---	0,481	0,269
				0,436	
	njob91	**-0,247**	0,015	**-0,332**	0,003
		0,101		*0,111*	
	alo91	0,718	0,068	---	---
		0,393			
	nj89_96	---	---	**0,374**	0,008
				0,142	
	nchi96	---	---	**-0,562**	0,008
				0,211	
	constant	-1,729	0,000	-0,705	0,189
		0,390		*0,537*	
valid observations N=			629		418
nonrespondents N=			55		55
survey persons N=			574		363
Likelihood ratio chi2 =			18,26		30,15
Prob > chi2 =			0,001		0,000
% correct predictions by model =			91,3%		87,1%
% correct predictions by guess =			84,1%		77%

Significant coefficients at 5%-level are given in bold.

Variables are:

NRS =	indicator of study:	1=person is in nonresponse study 0=person is in main study/ wave 1
NRSPAN =	indicator of study:	1=person is in nonresponse study 0=person is in main study/ wave 1
men =	indicator of gender	1=men/ 0=women
nevmarri91 =	indicator of never being married up until 1991 (1=yes;0=no)	
nevmar9696 =	indicator of never being married up until 1996 (1=yes;0=no)	
njob91=	number of reported job-spells up until 1991	
alo91 =	indicator of being unemployed 1991 (1=yes;0=no)	
nj89_96 =	number of job-spells between 1989 and 1996[81]	
nchi96 =	number of children[82] 1996	

The Model for 1991

The general fit of the model is better than it is for cohort 1930. The likelihood ratio-test is significant (LR chi2=18,26, p=0,001) and the percentage of correct predictions by the model

[81] I kept this variable overlapping the time of njob91, because the wave 2 questionnaire asked explicitly for jobs between 1989 (the end of the GDR) and the actual interview time. This enabled an anchoring of the respondents' memory in the historical moment of the beginning of the transformation years.

[82] Counted number of child-spells: interviewed persons reported dates about their (natural or step-) children. They may or may not live together in the same household.

is about 7 percentage points above the best guess (91,3% versus 84,1%). Gender as a controlling variable indicates lower probabilities for men but is not significant (as it was expected). Unemployment during the field time increases NRS participation (alo91, p=0,068). However, the important variables in this model are the never married status and the number of jobs up until 1991. The same trend as the one for cohort 1930 can be seen: persons with fewer jobs are more likely to be in the NRS file - as well as never married persons of course. TABLE 12 shows the predicted probabilities as estimated by the model.

TABLE 12: Predicted Probabilities for Nonresponse in Cohort 1960 / Wave 1 1991

Cohort 1960 / Initial Sampling 1991				
likelihood of being a nonrespondent is	in group	with	predicted probability for **MEN**	**WOMEN**
lowest	married or divorced 1991	number of jobs up until 1991= 4	0,050	0,062
	no unemployment	3	0,062	0,078
	married or divorced 1991	number of jobs up until 1991= 4	0,097	0,119
	unemployed	3	0,121	0,148
	never married 1991	number of jobs up until 1991= 4	0,107	0,132
	no unemployment	3	0,133	**0,163**
	never married 1991	number of jobs up until 1991= 4	**0,198**	**0,238**
highest	**unemployed**	3	**0,240**	**0,285**
		mean value for #jobs in complete group is 3,4	*p > 0,15 in bold*	

(Predictions were also calculated if combination of values was not observed.)

We see the greatest influence exerted by the unemployment variable: e.g. the risk doubles when "married or divorced women with 4 jobs" being not unemployed (p=0,078) is compared to the same group which is unemployed (p=0,148). There is also a doubling effect for the "never married" variable: e.g. the risk for "unemployed men with 3 jobs", married or divorced (p=0,121) doubles for never married men (p=0,240).

The Model for 1996

The general fit is better than for the previous model (LR chi2=30,15), as 10% more correct predictions are obtained than by the best guess (87,1% versus 77%).The logit estimates for 1996 (see last column in TABLE 11) reflect the labour market influence in the transformation. The family status variable is no longer significant, although the two "number of jobs" variables are. Whereas the results for the njob91 variable once again means that persons with more stable careers are to be found in the NRS study, the new job mobility variable (number of jobs between 1989 and 1996) indicates the opposite trend. Persons who had more jobs during the transformation years are more likely to be nonrespondents. Together with the unemployment result of the 1991 model, this means that the main study lost those persons who were forced into greater job mobility. Both job variables are 1% significant. The coefficients have a similar magnitude, although in the opposite direction. The family variable "number of children" is also 1% significant and favours being in the nonresponse study when having fewer (or no) children. If we recall in particular that the 1996 model estimates nonresponse study versus panel participantion[83], we see that it is easier to find those persons, who have a family, at home. The relative influence of the variables in the 1996 model can be examined in the table of the predicted probabilities.

TABLE 13: Predicted Probabilities for Nonresponse in Cohort 1960 / Wave 2 1996

Cohort 1960 / Wave 2 1996					
likelihood of being a nonrespondent is	in group	***keeping number of jobs up until 1991 at njob91=3***		predicted probability for **MEN**	**WOMEN**
lowest	married or divorced 1996	number of jobs between	1	0,053	0,079
	2 children	1989 and 1996	2	0,075	0,111
	married or divorced 1996	number of jobs between	1	0,089	0,131
	1 child	1989 and 1996	2	0,125	**0,180**
	never married 1996	number of jobs between	1	0,082	0,122
	2 children	1989 and 1996	2	0,116	**0,168**
	never married 1996	number of jobs between	1	0,137	**0,197**
highest	**1 child**	1989 and 1996	**2**	**0,187**	**0,263**
mean value for #jobs up until 1991 in complete group is 3,4				*p > 0,15 in bold*	
mean value for #jobs between 1989 and 1996 in complete group is 2,0					
mean value for #children 1996 in complete group is 1,6					

(Predictions were also calculated if combination of values was not observed.)

The likelihood of being in the NRS sample is about 50% greater for each additional job between 1989 and 1996 (e.g. "men, married or divorced, 2 children" with 1 job: p=0,053, but for 2 jobs: p=0,075). The effect of the children variable means a decrease in risk of roughly 60% for each child (e.g. compare "women, never married, 2 jobs 1989-96" for 1 child: p=0,263, but for 2 children: only p=0,168).

Summary for Cohort 1960:

During the initial sampling period, a significant increase in the risk of the NRS study is given by never having been married, being unemployed and having fewer jobs during the GDR era. Five years later, the family influence is no longer provided by the marital status, but rather by the number of children. Apparently, persons with more children could be reached better by

[83] Regarding panel persons as the "best" survey participants, as they could be contacted and interviewed twice.

the panel, which means that target persons with no or fewer children tend to be hard to reach. This is also confirmed for persons with greater job mobility in the transformation years.

General Summary

The aim of the prediction models was to check whether the profile of the "refusing" and the "mobile" cohort can be confirmed in a multivariate relationship.

The first general result is that ***no significant gender influences*** were provided in the multidimensional view even though the one-dimensional perspective reveals an underrepresentation of men.

The family situation is an important overall explanation for nonresponse. This result means that, from the survey's point of view, ***single persons are harder to contact.*** The single status in the oldest cohort means not being married and includes mainly widowed and divorced persons; whereas the younger single individuals are the never married persons. It was also discovered that, for the younger cohort 1960, the number of children influences the possibility of contacting and interviewing target persons. ***Persons with no or fewer children tend to be harder to contact***. This is also given if controlled by gender.

The labour market situation in the transformation influences the job mobility of the younger cohort. For the year 1991 of the initial sampling, unemployment increases the risk of nonresponse. Approximately five years later in 1996/97, those persons with more jobs during the last years are the nonresponse candidates. Obviously, the economic instability forced younger persons into ***greater job mobility which made it harder to reach them***. This is reinforced by the result that self-employed persons are almost completely excluded from the NRS study. Obviously, it was most problematic to contact this group.

The historical situation of the transforming society had another importance for the oldest cohort 1930. The labour market conditions were not relevant because all of them retired during the 90s.

Since the general level of education was lower during the historical period in which the 1930 cohort attended school, a higher degree in school turned out to be an explanatory variable: the group with a ***secondary qualification in school (grade 10) is the risk group for nonresponse in cohort 1930***. It is not related to the general aspect of how much effort is required to contact this cohort. It can, therefore, be concluded that people with a well established position in the GDR society tended to refuse to participate in surveys under the circumstances of a new system.

This is consistent with the other findings. Firstly, there is an extreme underrepresentation of high school degrees in the NRS study. Secondly, the number of reported job spells up until 1991 (which is the complete working life of this generation) is much lower for the nonrespondents. This leads to the conclusion that ***persons with stable established careers[84] in the GDR system are more likely to refuse***. Pure membership in a political party up until 1989 (i.e. mainly in the communist party) turned out not to be a significant explanation for the meek refusals of the NRS, as I had originally expected.

The ***personal interview method*** is the best predictor of the 1930 nonrespondents. This means that individuals giving gentle refusals can be convinced to participate better by interviewers who are personally present than by a voice on the telephone. Obviously, it is worthwhile for surveys to take these differences into account and to accept higher costs for the fieldwork.

The prediction models in general are not very powerful. The following principle findings of the multivariate model appear to confirm the one-dimensional results, however:

People living alone are a risk group for nonresponse. For younger persons, unclear labour market conditions led to greater mobility and it was, therefore, more difficult to contact the target persons. Tendencies to refuse can be seen in the older people. Persons with better careers benefited from the old system. It was harder to convince them to give interviews.

[84] To interpret these findings, I had a look at the target persons' concrete wording of the reported jobs up until 1989. I checked the "system-stable" jobs: military, police and formulations for political jobs. The percentage of these kinds of jobs in the main study is about 3,8% for cohort 1930. This is a conservative estimate, as sometimes only general formulations like e.g. "employee" were mentioned. I only included clear job information like "profession in the army" etc in the system stable category. In terms of percentages the value is only slightly lower for cohort 1930 in the NRS study. Given only a total of N=55 nonrespondents, however, I detected only 3 jobs in this category. I will assume that many of the adamant refusals were members of these job categories and were suspicious about being interviewed.

4. Estimating Bias in Multivariate Relationships

Having established a profile of the nonrespondents and a model to predict the participation in the NRS sample, the final question is: what can be done if the dependent variable in a multivariate relationship is suspected of being biased by nonresponse? In an OLS[85] regression model, the estimates of the coefficients might no longer be unbiased and consistent, as usually is expected given correct assumptions. The effects of the sample selection bias on OLS regression estimates are explained in more technical detail by Winship/Mare[86] (1992) and Brehm (1993:100-107). Hedayat/Sinha[87] (1991) adopt the idea of two population strata, the respondents and the nonrespondents. They show for the example of means[88] that nearly unbiased estimates for the population mean are obtained under certain assumptions: simple random sampling has to be assumed for both strata with a fixed proportion of sampled nonrespondents. For practical surveys these are rather strong requirements. (Hedayat/Sinha (1991:354-363) offer a perspective for bootstrapping and jack-knifing methods for situations when parameter estimation becomes more complicated, especially concerning the estimation of standard errors. These techniques are a separate field of research and shall only be mentioned briefly here.)

There are several approaches with regard to the correction of nonresponse bias. I will briefly discuss weighting adjustments. I will then focus on the Heckman sample selection model in more detail and present an application for a regression model using cohort 1960 of the nonresponse study and the main study.

4.1 Weighting Adjustments

There are a great number of strategies for weighting adjustments. One possible correction procedure is weighting[89] based on demographic variables, e.g. age, gender, socio-economic status or size of city. This can repair deviating marginal distributions and is an approach appropriate for panel studies when the panel attrition has to be corrected. In panel studies we are in the comfortable situation of knowing the demographic attributes of the persons, who

[85] ordinary least squares
[86] Winship/Mare (1992) Models for Sample Selection Bias. In: Annual Review of Sociology, Vol 18:327-350.
[87] Hedayat,A.S./Sinha,B.K. (1991) Design and Inference in Finite Population Sampling.
[88] The estimation of means can be considered a special case of linear regression, namely estimating only one linear parameter.
[89] I will not discuss weighting techniques based on the sampling design where we know the necessary design information to calculate weighting factors.

had dropped out of the former wave. This information is not given in the initial sampling, however. Census data is, therefore, often used to validate a sample. Weighting based on census information has to cope with several other problems such as e.g. incompatible definition of target groups[90], differing variable categories, incomparable interview situation, etc. Another disadvantage is that such a weighting technique may be able to correct marginal distributions, but that we cannot assess what happens in multivariate interdependencies[91].

Other approaches refer more strongly to the aspect of nonresponse. The underlying idea is to find a weighting adjustment that reflects the different probability of response. One strategy is to take classes which represent the ease with which a person is contacted in order to construct weights as e.g. is done in the Politz-Simmons[92] procedure or weighting by the number of necessary telephone calls.

Burton (1999) suggests weights based on attitudes according to his hypothesis that nonrespondents differ from respondents in their attitudes and opinions.

The last approach I will mention here is weighting by the propensity to respond. Firstly, a regression of a dichotomous variable (being a respondent versus not being one) on explanatory variables is performed, then the inverse of the predicted probabilities will be taken as weights. This is discussed by Little/Schenker[93] (1995) and Brehm (1993:118). Persons with a low prognosis for response will get more weight by this method. A good model is, therefore, necessary in order to predict who will be a respondent and who not. In other words, one needs information about the excluded "dark chapter persons" as well. Schnell (1997:249-250) discusses the method as an interesting alternative under the title "propensity weights". Brehm[94] (1993) is more sceptical about the positive effect, particularly in multivariate analyses. Little/Schenker (1995:47) mention high variances for weighted estimates and general problems of variance estimation in such a case.

4.2 The Heckman Sample Selection Model

The fundamental idea of the Heckman model is to correct a possibly biased regression model in the error term. The error adjustment is calculated on the basis of a previous model which

[90] As already mentioned, the EGLHS sampled East German persons using the concept of their origin as the basis, whereas the German microcensus samples according to the place of residence.
[91] This is also mentioned by Schnell (1997:248).
[92] It uses informationon how often people could have been contacted. See: Brehm (1993:118).
[93] Little/Schenker (1995) Missing Data. In: Arminger/Clogg/Sobel (1995) Handbook of Statistical Modeling for the Behavioral Sciences. pp. 46-47.
[94] See critical discussion in Brehm (1993:118-121).

predicts being in the sample or not. The general formal procedure is the following (the development of the formula see in appendix C):

Let $\mathbf{Y_i = X_i' \beta + \varepsilon_i}$ *regression equation* ***(*)***

be the linear equation of interest for i = 1, ..., n observations. Y is the dependent variable[95]and there are k independent X-variables. If all the standard regression assumptions[96] hold, the expected value of the error term is zero and best linear unbiased estimates of the β's are obtained. If the dependent variable is not observed for all sampled units (as in the case of nonresponse) we might formulate a model of sample selection saying:

Y_i is observed if : $\mathbf{Z_i' \gamma + u_i > 0}$ *selection equation* ***(**)***

having m independent Z-variables in the selection equation with at least one Z-variable which is not included in the initial regression equation. The important idea of the Heckman model is that it takes into account a correlation between the error terms ε_i and u_i and hence corrects for the resulting bias. So the original regression (*) involves a probit model (**) in order to predict sample participation. The statistical package STATA has implemented this procedure automatically in a "heckman"[97] command. Winship/Mare (1992:340-341) point out that the Heckman selection model is based on normality and linearity assumptions for the error terms and also that it is sensitive as to how effectively the selection equation predicts the sample participation. Brehm's (1993:123) argument which is a contrast to weighting, however, is: *"correct the model for nonresponse, don't correct the data".*

4.3 Application of Correction: An Example With an Intentionally Biased Sample

To understand how the Heckman model works, it will be applied to an example. We want to estimate the household income 1996 for women of birth cohort 1960 by an OLS regression model. For the year 1996 there is data available for the panel persons and for persons in the nonresponse study. We assume that taking the nonrespondents (as a part of the initially sampled persons) together with the main sample yields better information about the "true"

[95] assumed to be continuous

[96] For an explanation of the regression assumptions, see: Pindyck/Rubinfeld (1991:73-100) Econometric Models and Economic Forecasts.

[97] Two possible estimations are implemented in STATA: the Heckman two-step procedure and the full maximum likelihood version. The presented calculations in this chapter were done with the two-step version. Parallel computations revealed that the maximum likelihood version produced smaller confidence intervals. Also Brehm (1993:123) reports better asymptotical efficiency for maximum likelihood. I took the two-step procedure, however, as some tests with the data, taking the ML module, produced programme warnings related to convexity. This could mean that the numerical algorithm is unstable with the data. The STATA manual states that the Heckman two-step procedure is more stable with problematic data.

relationship than would provided were the nonrespondents to be left out.[98] We are interested in the amount of bias for the estimated regression coefficients which influence the dependent variable and we want to correct it.

An extreme example shall demonstrate what happens when the data is severely biased. For this purpose, I have systematically excluded cases with higher education which also means leaving out cases with higher income.

The idea of this experiment is as follows: a decision as to whether a correction method should be applied can never be 100% secure. This is also argued by Winship/Mare (1992). So many aspects effect the quality of a sample. But we expect to see matters clearly under controlled extreme conditions. The same reasoning is given in mathematical proofs. To prove that a theorem on e.g. natural numbers is true for all numbers, is often complicated. Given it is true, however, we can verify it for any arbitrary number including both - the extreme and the trivial cases.

Now the example:

"OS" stands for the original data, "BS" for the biased data. Both files are described as follows:

OS = original sample as reference

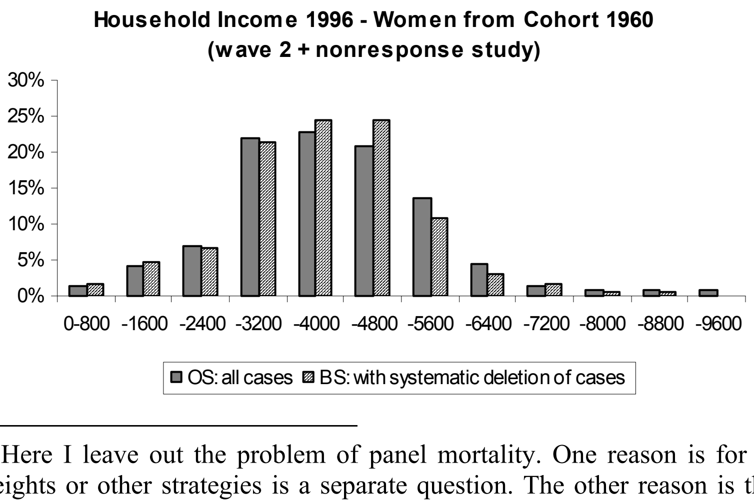
Women of birth cohort 1960; original data from main study/wave 2 and from nonresponse study (N=234 persons).

BS = intentionally biased sample

More than 80% of the women with the highest qualification in school ("Abitur") were eliminated and nearly all persons with completed university degrees. Therefore, lower household income appears more often in this sample (remaining N=179 persons).

The following figure shows the histogram of the household income 1996 for both samples.

FIGURE 25: Histogram of Household Income in the Original and in the Biased Sample



mean values for household income 1996	in DM
original data wave 2 + nonresponse study valid N=220	3928,25
intentionally biased data only selected cases valid N=168	3780,74

EGLHS, cohort 1960, women

---

[98] Here I leave out the problem of panel mortality. One reason is for simplicity as assessing possible panel weights or other strategies is a separate question. The other reason is the simple fact that data is available for panel persons and nonrespondents, but not really for panel drop-outs. Thus there is no possibility of comparing results.

A logit model predicts whether a person is a participant of the biased sample or not. The dependent dichotomous variable is "inbs" (1=person is in BS; 0=person was excluded). Due to the construction of the biased sample, the variables SCHOOL (level of education in school on a scale of 4 degrees) and VHIGH (0/1 indicator of completed university education) strongly predict the BS sample participation as we can see in the list of results.

TABLE 14: Logit-Model to Predict Selection for the Biased Sample

dependent variable	independent variables	coefficient *(standard error)*		
inbs	**SCHOOL**	**-2,9293**		
(being in		*0,6617*	number of observations	234
the biased	**VHIGH**	**-3,7130**	LR test chi2(2)	130,06
sample)		*1,1167*	Prob > chi2	0,0000
	constant	**8,1753**	% correct predictions by model	93,6%
		1,4336	% correct predictions by guess	64%

All estimated coefficients and constant are significant with p<1%.
Calculations done by STATA 6.0.

All coefficients in the sample selection equation are significant. The model produces good predictions for participation in the biased sample BS. The likelihood-ratio-test shows that the model differs significantly from the mean model which estimates only the constant. I added the percentage of correct predictions[99] as calculated by the model in order to provide a contrast to a best-guess-prediction. We see that the model is clearly better. It will be the sample selection model in the Heckman procedure.

The same linear regression model for household income 1996 will now be run in three different manners:

1) with the artificially biased sample (BS) alone
2) with a heckman sample selection model, taking the above prediction model
3) with the original sample (OS).

The results are given in the following synopsis.

[99] Predictions p>=0,5 were counted as "being in the biased sample", p<0,5 as not. The "best-guess-prediction" was calculated by taking the binomial empirical relative frequency of being in the sample. Both groups, "in" and "not in" the sample, were distributed to a yes/no prediction assuming the empirical probability.

TABLE 15: Linear Regression Model for Household Income 1996 / Three Different Calculations

		Biased Sample BS		***Heckman Correction***		***Original Sample OS***	
dependent variable	**independent variables**	**coefficient** *(std. error)*	**P>|t|**	**coefficient** *(std. error)*	**P>|z|**	**coefficient** *(std. error)*	**P>|t|**
hhinc96	nphh96	**209,32**	0,049	218,26	0,113	**292,04**	0,003
		105,63		*137,88*		*98,67*	
	marrie96	**1215,78**	0,000	**1194,62**	0,000	**1267,27**	0,000
		251,13		*339,87*		*252,02*	
	alo96	**-687,96**	0,006	**-688,51**	0,013	**-845,30**	0,001
		246,12		*277,69*		*252,74*	
	capi	**-354,21**	0,048	-352,46	0,093	**-392,25**	0,028
		177,35		*209,86*		*177,16*	
	humkap	**368,48**	0,000	**332,89**	0,050	**255,87**	0,000
		67,34		*169,54*		*45,91*	
	constant	**-2212,65**	0,012	-1897,99	0,183	-1194,14	0,080
		874,61		*1426,08*		*678,56*	
valid observations			161		216		211
panel observations					161		
nonrespondents					55		
	F		21,91				24,67
	Prob > F		0,000				0,000
	r**2		0,414				0,376
	Wald chi2				32,48		
	Prob > chi2				0,000		
	ρ				0,449		
	ρσ				468,471		

Bold coefficients mark 5%-significance level.

Variables are:

	hhinc96 =	household income 1996
	nphh96 =	number of persons in household (nphh96),
	marrie96 =	being married 1996 (1 if married; 0 otherwise),
	alo96 =	being unemployed 1996 (1 if yes; 0 otherwise),
	capi =	CAPI[100] interview 1996 (1 if yes; 0 otherwise),
	humkap =	cumulated education in school and vocational training in years[101]

Regression Estimates for the Model in the Original Sample

The regression model for household income 1996 in the original sample, taking panel participants and nonrespondents together, will be interpreted as the "validation model" in this context (see last column in TABLE 15). The significant F-test indicates that we can assume that

[100] The CAPI variable is treated as a proxy variable for socio-economic status in this cohort. Households without a telephone tend to have a lower income and a lower level of education (see chapter 2, FIGURE 14).

[101] The intention of the scale is to reflect the usual number of years that are necessary for obtaining a certain qualification level. A scale in years is also used by Blossfeld/Rohwer (1995:80-93) for education as the independent variable in transition rate models. For the data here, it was not unambiguous to define a clear projection of vocational degrees onto years. The reason is that in the GDR there were several possible dynamic ways to obtain degrees, like e.g. vocational training including a final high school degree ("Berufsausbildung mit Abitur"); part time studying; correspondence and evening courses. I have, therefore, defined a rough scale as follows:
humkap=SCHOOL + VOCAT in years having
SCHOOL: <grade 8=7 years; grade 8=8 years; grade 10=10 years; Abitur=12 years
VOCAT: no training at all=0 years; semiskilled=1 year; skilled worker=2 years; technicians/special schools=3 years; university=4 years.

not all coefficients are simultaneously zero. The model explains about 38% of the variance (r**2=0,376). All coefficients are significant at a 5%-significance level. They appear in the direction that is theoretically expected, according to the analysis in the exploration chapter, and can be interpreted in terms of their financial contribution to the household income 1996 for women of cohort 1960.

Having income as the dependent variable, it is easy to interpret the coefficients. We see that the income is rising by the number of persons (one more person contributes 292,04 DM) and the most positive effect on income is given for married women in comparison to unmarried (1267,27 DM). Being unemployed decreases the household income. We also see that target persons from the capi field have lower household income compared to persons from the telephone field. The effect is nearly half of the unemployment effect. We also see that longer education contributes positively to the household income. Following the constructed definition of this variable, we may interpret one more year of education in school or later vocational training as being worth 255,87 DM in income, keeping all other coefficients of the model comparably constant.

Regression Estimates for the Model in the Biased Sample

The column with the regression estimates for the biased sample shows that the significance structure and the overall fit is comparable. Also we do not find severe changes in the sense that positive coefficients turned to negative (or the reverse). We see that some estimates differ greatly, however, e.g. the coefficient for humkap (368,48, although only 255,87 in the original sample). Of course this is the bias impact which was intentionally implanted in this sample. (The high education groups were excluded.) As a consequence, the effect of years of education on the household income is overestimated in the remaining cases.

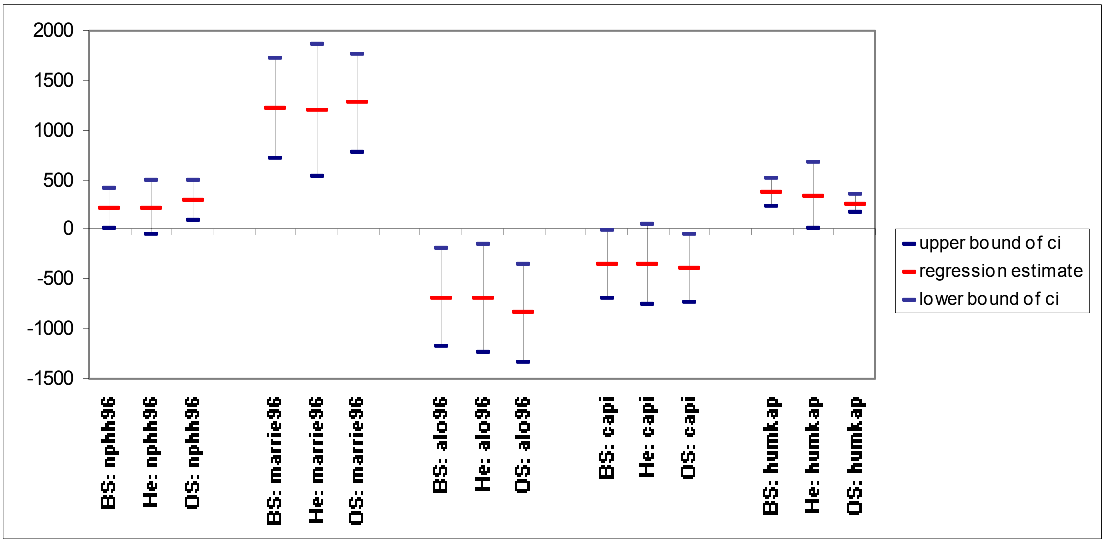
Regression Estimates for the Heckman Sample Selection Model

We see in the general structure that the variables for number of persons and capi interview are no longer significant. There is an overall increase in the standard errors. The significant Wald-test also indicates that the model differs from the null hypothesis assuming all coefficients to be zero. The value for rho (ρ=0,449) shows the relatively high correlation between the error terms of the selection and the regression equation. Rho sigma ($\rho\sigma$=468,471) is the value of the coefficient of the additional correction regressor (the Inverse Mill's Ratio) and is an overall estimate of the magnitude of the selection bias (see: Brehm (1993:122); STATA 6.0 manual, pp.18-20). $\rho\sigma$ is rather high in our case.

How did the Heckman correction work in the artificial extreme example? The expectation is that it should correct in the direction of the original sample (which we use as a validation). Except for the marriage variable (marrie96), where the correction estimate is outside the gap between BS and OS, all other Heckman estimates are either almost identical (alo96, capi), or have moved clearly into the direction of the original sample (nphh96, humkap). The constant is also closer to the original sample. This means that there is in fact an improvement.

It has to be emphasised, however, that parameter estimation also means getting the position of the parameter <u>within a confidence interval</u>[102]. The following figure, therefore, contrasts visually the calculated regression coefficients within their confidence intervals.

FIGURE 26: Position of Estimated Regression Coefficients within Confidence Intervals
- Biased Sample, Heckman Correction, Original Sample -

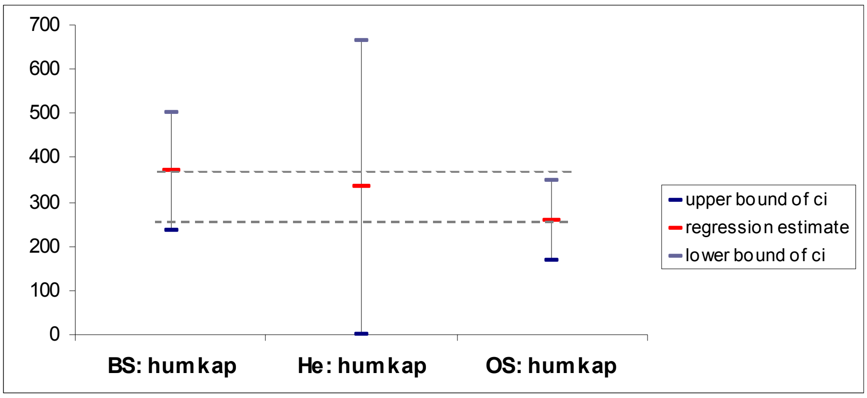


BS = OLS regression with intentionally biased sample (high education and income eliminated)
He = Heckman sample selection model
OS = OLS regression in original sample (women of cohort 1960; wave 2 and nonresponse study)

We see that the structure of the relative importance and magnitude of the coefficients is rather stable. The Heckman confidence intervals are the widest but they properly include the validation intervals (OS). The general tendency is that the bias correction moves in the expected direction. We see this result in more detail in the next figure which focuses on the coefficient estimate for the variable "humkap".

[102] Here: the frequently used 95%-interval.

FIGURE 27: Bias Correction of Coefficient Estimate for Educational Variable



----
marks the "bias region" between estimates for original and biased sample.

We have, therefore, learnt how the Heckman model for bias correction works from the extreme example with the intentionally biased sample. All in all it turned out that improvements were made by remaining on the conservative side with regard to confidence intervals and significance levels.

4.4 The Heckman Correction in a Regression Model for Women of Cohort 1960

I will now leave the demonstration example and return to the real data. The regression for household income will be run once again and compared to three different models.

Firstly, the regression is estimated for the 1960 women of the panel study. This is the usual survey situation modelling a multivariate relationship upon the basis of the respondents' data.

Secondly, the same model is run taking data for the panel participants together with the initial nonrespondents. One could of course raise the objection that this enlarged sample is not the original "true" one, because it does not include all or at the very least a random subset of all the initial nonrespondents. I am arguing, however, that a realised survey data set always reflects decisions about the termination of the field work (due to time and financial budget). As the nonrespondents of the NRS study belong to the originally sampled persons, they might have been successfully interviewed, even as panel cases, if the initial fieldwork had made a little more effort. As a consequence, a model which also includes these persons ought to come closer to the complete "true" sample than it would do were they to be left out.

The third model is the Heckman sample selection concept as introduced in the previous section. It will give an estimate for the total bias ($\rho\sigma$) and correct the regression coefficients. The necessary selection equation is given in the following table, which lists the results given by a separate logit model.

TABLE 16: Logit-Model to Predict Panel Participation against Nonresponse Study
- Women of Cohort 1960 -

dependent variable	independent variables	coefficient *(standard error)*		
sele	nevmar96	-0,5849		
		0,6941		
	nchi96	0,5572		
		0,3072		
	njob91	**0,5852**		
		0,1717	number of observations	234
	nj89_96	**-0,5988**	LR test chi2(4)	27,30
		0,2105	Prob > chi2	0,000
	constant	0,4371	% correct predictions by model	87,2%
		0,7167	% correct predictions by guess	77,8%

Bold coefficients indicate 5%-significance-level.

Variables are:	sele[103] =	dichotomous variable for sample selection: 1=panel person; 0=nonrespondent
	nevmar96 =	never married 1996 (1 if never married; 0 otherwise,i.e. married,divorced,widowed)
	nchi96=	number of children 1996
	njob91=	number of jobs up until 1991
	nj89_96=	number of jobs between 1989 and 1996

The coefficients for the labour market variables "number of jobs up until 1991" and "number of jobs between 1989 and 1996" are 5%-significant . The coefficients for "number of children" and "never married" are not significant, but I keep them in the selection equation since they characterise the detected deviations of the 1960 nonrespondents from the panel participants and are used as control variables. The likelihood ratio test is significant and the percentage of correct predictions by the model is 10% above the best-guess percentage (87,2% versus 77,8%). We naturally do not have such a strong predictability as is given in the artificial example, due to the fact that the main study should not have left persons out so systematically. The following table gives the synopsis of the models.

[103] Contrary to chapter 3, which deals with the predictions for being a nonrespondent, the dependent variable here means participation in the main study which is the binary complement of the same thing. The dependent variable has now changed the 0/1 code. As a consequence, the coefficients (in comparison to the previous chapter) change the sign. For example, the positive coefficient for njob91 has to be interpreted as: the higher the number of jobs up until 1991 there are, the more likely it is that the persons will be in the main study.

TABLE 17: Heckman Sample Selection Model Using Panel and Nonresponse Study - Women of Cohort 1960 -

		Only Panel Persons		*Panel with Heckman Correction*		*Panel Persons + Nonrespondents*	
dependent variable	**independent variables**	**coefficient** *(std. error)*	**P>|t|**	**coefficient** *(std. error)*	**P>|z|**	**coefficient** *(std. error)*	**P>|t|**
hhinc96	nphh96	**274,68**	0,015	**277,43**	0,035	**292,04**	0,003
		112,03		*131,38*		*98,67*	
	marrie96	**1302,08**	0,000	**1300,91**	0,000	**1267,27**	0,000
		276,12		*292,70*		*252,02*	
	alo96	**-917,24**	0,001	**-915,77**	0,001	**-845,30**	0,001
		265,68		*269,47*		*252,74*	
	capi	**-400,41**	0,040	-400,75	0,051	**-392,25**	0,028
		193,63		*204,97*		*177,16*	
	humkap	**254,18**	0,000	**254,39**	0,000	**255,87**	0,000
		51,52		*56,46*		*45,91*	
	constant	-1058,97	0,154	-1077,99	0,212	-1194,14	0,080
		739,97		*864,40*		*678,56*	
valid observations			183		213		211
panel observations					183		
nonrespondents					30		
	F		21,78				24,67
	Prob > F		0,000				0,000
	r**2		0,3809				0,376
	Wald chi2				96,63		
	Prob > chi2				0,000		
	ρ				0,027		
	$\rho\sigma$				32,183		

Bold coefficients mark 5%-significance-level.

Variables are:
- hhinc96 = household income 1996
- nphh96 = number of persons in household (nphh96),
- marrie96 = being married 1996 (1 if married; 0 otherwise),
- alo96 = being unemployed 1996 (1 if yes; 0 otherwise),
- capi = CAPI interview 1996 (1 if yes; 0 otherwise),
- humkap = cumulated education in school and vocational training in years

In the left column we find the estimates for the survey data (only panel persons), in the middle the Heckman correction and in the right column the coefficients for the complete sample containing panel persons and nonrespondents. The last one was also the validation sample OS in the bias example. The same results are listed again.

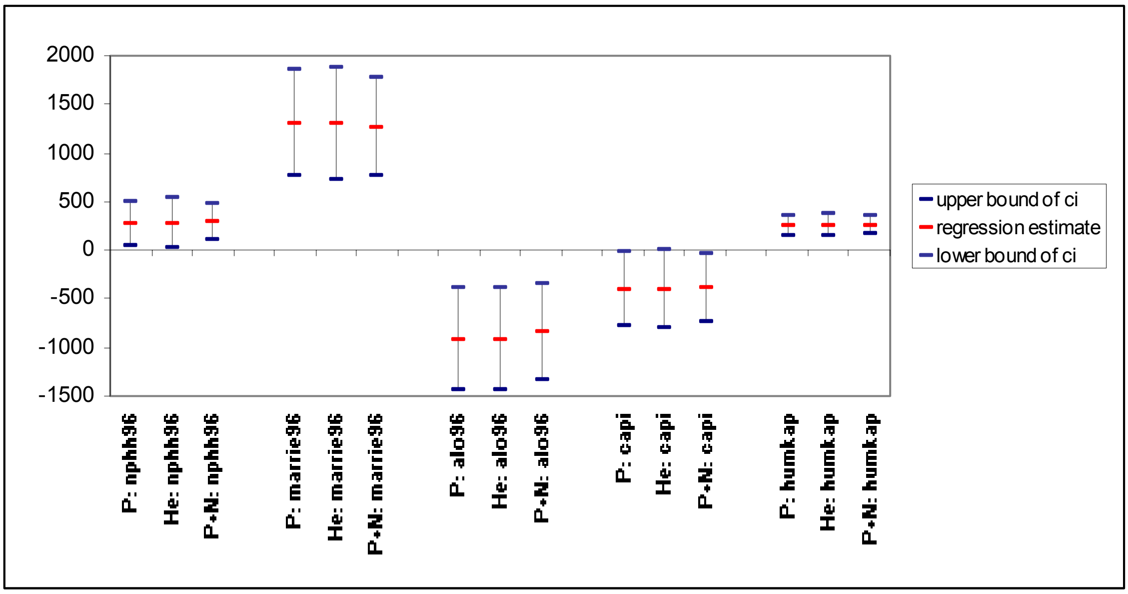
In general, there are no great differences between panel sample and panel+nonresponse. The Heckman corrections are not strong, although each[104] of the coefficients, including the constant, goes in the right direction of the validation sample. Also the level of significance of the estimates does not change substantially (capi being near to 5%). We recognise with ρ=0,027 that the selection equation is not very closely related to the

[104] Considering the estimate for capi as equal.

regression model and also that the total bias $\rho\sigma$=32,183 is much lower than in the extreme example. The confidence intervals are only slightly wider for the Heckman estimates.

In the same type of graphic as those above, FIGURE 28 shows the coefficients within their confidence intervals.

FIGURE 28: **Pattern of Regression Results in Panel Study, Heckman Model, and Panel + Nonresponse - Position of Coefficients Within Confidence Intervals -**



P = regression model with panel sample
He = Heckman sample selection model
P+N = regression model with panel + nonresponse sample (all models: only women of cohort 1960)

Summary

In this chapter I presented applications of the Heckman sample selection model to correct nonresponse bias.

Firstly, an extreme example that was deliberately biased was demonstrated. Under controlled conditions a selection equation was given that could predict the sample participation extremely well. I took the complete unbiased sample as a validation. It turned out that the Heckman estimates, given that there were differences in the first place, were corrected in the right direction. On the other hand, along with the improvement one had to then accept lower precision with wider confidence intervals and more conservative significance levels.

In the second step, I applied the Heckman correction method to the regression for household income 1996 with real life data. I compared the bias correction with the results for the same model taking the 1960 women of the panel study on the one hand and the panel persons together with the corresponding nonrespondents on the other hand. In general the structure of the regression results did not change very much. The relative magnitude of the coefficients and their level of significance was stable. The Heckman results were situated between the survey and the validation sample and a low total bias was estimated. Though the bias correction generally tended to be only slight, the Heckman method appeared to work on the secure side. No correction was outside the expected range.

Due to the fact that the general frame of this thesis is limited, I cannot describe other applications of corrections in detail. I will, therefore, only briefly mention tests with propensity weights. Taking the same data and the presented selection equation, I checked the same regression with weights calculated by the inverse of the predicted probability for sample participation. Though some estimates were corrected, it turned out that others were far out of range between the panel and the total sample, even in the extreme bias example. So Brehm's (1993:119-121) statement that weighting is no solution to the problem of nonresponse in multivariate analyses was verified.

The decision as to whether a correction method is required and which one is useful depends upon the purpose of the regression model. If a precise prediction is needed, then the exact coefficient estimate in numbers is important and the danger of bias is relevant. If we are more interested in explaining the structure of interrelated variables - and this is often the case in sociological models - we focus on the interpretation of the relative importance of the coefficients and not on the absolute number. The efficiency of the Heckman correction depends upon a well specified selection equation (also the propensity weights method) which is more difficult to specify in the nonresponse problem. The income regression example showed that the interpretation of the significant effects was essentially unaltered.

.

5. Final Discussion

This thesis examined patterns of nonresponse and discussed strategies of bias corrections using the nonresponse study of the East German Life History Study. At the end of the analysis of the nonresponse study, I will repeat and discuss the general findings.

In chapter 1, I pointed out the conflict between the ideal theoretical random sample and the practical world of surveys. Several inevitable sources of error always influence the quality of survey data. One of these influences is the nonresponse problem, the guiding subject for this thesis. Mainly the sample drop-out in the initial phase of sampling is a problem, as usually no individual information about the lost persons is available. The concentration on this topic emphasises its importance. One should be aware, however, that the nonresponse problem must be seen in a sequence of research steps. The careful control of other conditions (e.g. the sampling design or the formulation of the questions contributes to a high quality of the survey data.

In chapter 2, I introduced the nonresponse study of the East German Life History Study. The accent of the chapter was the data exploration. I presented comparative descriptions of the nonresponse study, and I will briefly summarise the main results.

First of all, it has to be underscored that the nonresponse study in principle is biased, as most of the refusals are missing again. This constraints generalisations about refusing behaviour (the overall refusal rate in the study is about 70%). Predominantly, one can assume to have information about hard-to-contact persons and some of the "soft" refusals. However, the big advantage of the nonresponse file is to have collected individual data using almost the same interview instrument as the main study did.

The cohort design of the German Life History Study suggests a separate data inspection per birth cohort. It turned out that in fact the characterisations of the NRS persons differ by cohort. On the one hand, this can be interpreted as age-specific: persons of birth cohorts 1930, 1940, 1950, 1960 were about 60, 50, 40, 30 years old in the year of the initial sampling. On the other hand, the cohort differences reflect a different impact of historical situations during the life course.

In a one-dimensional comparison, the NRS data is described by an extreme loss of birth cohort 1930, a loss of (older) men, and a higher proportion of single persons. In addition to these results, variables indicating a stable social position (better education, stable careers, a higher proportion of party membership for men) describe differences for the older cohorts.

Besides the attributes "single and fewer children" , the younger cohorts are characterised by indicators for higher job mobility in the transformation years. Nearly no differences between NRS and main study were found for cohort 1950. Together with the findings of the methodological reports, the nonresponse data explorations resulted in the description of **cohort 1930 as the "refusing generation"** and **cohort 1960 as the "mobile generation".**

In chapter 3, I concentrated on the polarising cohorts 1930 and 1960 and presented logit models to predict the participation in the nonresponse study. The aim was to check whether the detected one-dimensional differences could be confirmed in a multivariate view.

The status of being single, stable careers, and the personal interview method are predictors for being in the nonresponse study for cohort 1930, whereas the political variable for party membership is not significant in the multidimensional context. The models for cohort 1960 showed that the family situation "single, fewer/no children", and a greater job mobility in the transformation are relevant predictors.

The general predictive power of the models is weak and only slightly better than a best-guess prediction. Therefore the question is: about what kind of persons does the NRS study tell us anything? My conclusion is that the interviews can be considered as the result of special fieldwork efforts. This means: the observed nonrespondents represent such a segment of the initial sample which might have been collected initially, if one had decided to spend more costs and time. Focusing on hard-to-contact persons, the effort appears as profitable and one could consider the NRS study as additional cases. Considering hard refusals, there seems to be a general limit to convince them, and investing more effort would not improve the refusal rate.

In chapter 4, I took the nonresponse data as an expansion of the main study and discussed strategies for bias correction. The augmented sample was treated as a validation in order to check the Heckman sample selection procedure. In the household income regression model for women of cohort 1960, the Heckman bias corrections worked in the right direction. Concerning the precision, the model resulted in wider confidence intervals which means staying on the secure side. The conclusion is that the Heckman bias correction offers a solid chance to improve regression estimates. However, the magnitude of the correction was only small, as the Heckman procedure depends on a good prediction model for the sample drop-outs. But even the example with an artificially distorted sample showed an amazing stability in the structure of the regression estimates. I argue that - given a survey had undertaken careful attempts to minimise all the other error influences - structural explanations of social phenomena appeared to be relative robust against nonresponse. (At least this can be concluded by the validation for nonresponse in the sense of converted soft refusals and finally reached non-contacts.)

Finally, there are two general results which the analysis of the nonresponse study made evident.

Firstly, the analysis revealed that the profile of the nonrespondents is cohort-specific. This result cannot be explained as a mere influence of age. It emphasises instead the importance of the individual's specific situation in the life course. The historical and political situation in which a survey occurs is experienced differently in different positions of life. As a consequence of that, the decision to participate in a survey or to refuse depends on conditions related to typical patterns in the life course. It is important to keep this in mind when one plans a survey and works on how to contact and to convince people.

Secondly, the hard-core refusals remain the "dark-chapter" persons. It could be suspected, however, that they tend to be those people who experienced the change in the society as a disadvantage concerning their own life. Given the political background of the transformation in East Germany, this means that not the underqualified persons, but the former elites prefer to refuse.

The additional effort on the fieldwork was successful for the hard-to-contact cases. Finally, several more cases could be realised. Naturally a survey, which is based on random sampling and which was carefully performed, cannot be remarkably improved by adding a few more cases. The practical data analyst, however, sometimes wants to have available some more cases. This is the case, particularly, when one inspects detailed subgroups with almost empty cells. From this practical point of view, the nonresponse study can be used to enrich the data pool. The analysis of the refusing 1930 cohort, however, showed that the study succeeded in getting converted soft refusals who refused because of convenience and not because of their convictions. As multidimensional structural explanations are not changed essentially, it can be doubted whether a survey should run after each of these target persons. There are always limitations concerning time and financial budget. The substantial task is to convince persons. This implies for empirical research under democratic conditions, that the respondent must be taken seriously and must get the opportunity to refuse. The hope is that it is just the existence of this opportunity which might convert and convince sampled target persons.

References

Aldrich, John H./Nelson, Forrest D. (1984) Linear Probability, Logit, and Probit Models. Sage Publications, Newbury Park, pp. 30-35,58.

Blossfeld, Hans-Peter/Rohwer, Götz (1995) Techniques of Event History Modeling. Lawrence Erlbaum, Mahwah, New Jersey, pp. 80-93.

Brehm, John (1993) The Phantom Respondents: Opinion Surveys and Political Representation. University of Michigan Press, Ann Arbor, pp. 16-20,23-38,93-107,118-123.

Burton, Jonathan (1999) Public Attitudes and Responses to Survey Research. University of Essex. PhD Thesis.

Converse, Jean M. (1987) Survey Research in the United States. Roots and Emergence 1890-1960. University of California Press, Berkeley and Los Angeles, chapt. 1-4, and pp. 54-104,137-144,189-211,360-361.

Erickson, B.H. and Nosanchuk, T.A. (1992) Understanding Data (Second Edition). Open University Press, Buckingham, pp. 3-11.

Feller, William (1950) An Introduction to Probability Theory and Its Applications. Vol. I. John Wiley & Sons. New York, London, pp. 155-156,191-196.

Groves, Robert M. (1989) Survey Errors and Survey Costs. John Wiley & Sons. New York, pp.6-80,265-267,271-279,360-484.

Groves, Robert M./Couper, Mick P. (1998) Nonresponse in Household Interview Surveys. John Wiley & Sons, New York, pp.4-5,49-50,157-173.

German Life History Study

Data Documentation and Codebook for The East German Life History Study/Wave 1:
Datendokumentation (1995) Lebensverläufe und historischer Wandel in der ehemaligen DDR/ Codebook für die 1. Welle, bearbeitet von Ralf Künster, Max-Planck-Institut für Bildungsforschung, Berlin.

Data Documentation and Codebook for The East German Life History Study/Wave 2:
infas Sozialforschung (1996). Projekt: "Lebensverläufe in Ostdeutschland - 2. Welle 1996/97", Dokumentation des Erhebungsinstrumentes für die Kohorten 1929-31, 1939-41, 1951-53 und 1959-61 (Codebook). Bonn.

general information available on the Internet:
http://www.mpib-berlin.mpg.de/BAG/Studien/ZA-Info38/Wagner.htm (for the West German data)
http://www.mpib-berlin.mpg.de/BAG/Studien/ZA-Info38/solga.htm (for the East German data)

Hedayat,A.S./Sinha,B.K. (1991) Design and Inference in Finite Population Sampling. John Wiley & Sons. New York, pp.354-363.

Huinink, Johannes (1995) Individuum und Gesellschaft in der DDR - Theoretische Ausgangspunkte einer Rekonstruktion der DDR-Gesellschaft in den Lebensläufen ihrer Bürger. In: Huinink, Johannes/Mayer, Karl Ulrich (1995) Kollektiv und Eigensinn. Akademie Verlag, Berlin. pp. 25-44.

Huinink, Johannes/Mayer, Karl Ulrich/Trappe, Heike (1995) Bildungs- und Berufsverläufe. In: Huinink, Johannes/Mayer, Karl Ulrich (1995) Kollektiv und Eigensinn. Akademie Verlag, Berlin. pp. 106-111.

infas Sozialforschung (1995) Lebensverläufe und historischer Wandel in der ehemaligen DDR, Methodenbericht der Hauptstudie. Bonn, pp. 2,16-23.

infas Sozialforschung (1998) Ostdeutsche Lebensverläufe im Transformationsprozeß, Methodenbericht zur Hauptstudie. Bonn, pp. 9,18-20,27-28.

infas Sozialforschung (1998) Lebensverläufe in Ostdeutschland - Non-Response-Studie. Dokumentation des Erhebungsinstrumentes (Fragebogen und Codebook). Bonn, pp. 26-30.

Kalton, Graham (1983). Introduction to Survey Sampling. SAGE Publications. Newbury Park, pp. 8-57, 69-81.

Kish, Leslie (1965) Survey Sampling. John Wiley & Sons. New York, pp.6-8,53-59,77-82, 148-166, 531-535.

Kolmogorov, A.N. (1956) Foundations of the Theory of Probability. Second English Edition. Chelsea Publishing Company. New York, pp.2,8-12,14-16,61-68.

Little, Roderick J. A./Rubin, Donald B. (1989) The Analysis of Social Science Data with Missing Values. In: Sociological Methods and Research, Vol. 18, Nos. 2 & 3, November 1989/February 1990. Sage Publications, pp. 292-326.

Little, Roderick J.A./Schenker, Nathaniel (1995) Missing Data. In: Arminger/Clogg/Sobel (1995) Handbook of Statistical Modeling for the Behavioral Sciences. Plenum Press. New York, pp. 46-47.

Mikrozensus 1991: information on the Internet
http://www.zuma-mannheim.de/data/microdata/brd/microdata

Pindyck, Robert S./Rubinfeld, Daniel J. (1991) Econometric Models and Economic Forecasts. McGraw-Hill, New York, pp. 73-100.

Schnell, Rainer (1997) Nonresponse in Bevölkerungsumfragen. Ausmaß, Entwicklung und Ursachen. Leske + Budrich. Opladen, pp. 23-30,72-76,87,135-139,190-209,219-223,248-250.

Schnell, Rainer/Hill, Paul B./Esser, Elke (1995) Methoden der empirischen Sozialforschung. R. Oldenbourg Verlag, München, pp. 40,297-351.

Stata 6.0 (1999) Reference Manual. Stata Press, College Station, Texas.

Statistical Yearbook 1997 for the Federal Republic of Germany. Federal Statistical Office, Wiesbaden, p. 62.

Tufte, Edward Rolf (1990) Envisioning Information. Graphics Press. Cheshire, Connecticut. pp. 53-66,81-96.

Tukey, John W. (1977) Exploratory Data Analysis. Addison-Wesley Publishing Company. Reading, Mass.

Verma, Vijay (1998) Sampling Methods. Training Handbook #2, Statistical Institute for Asia and The Pacific Tokyo. Essex Summer School 1998, Course Notes: Part 1 and 2, pp.1.1-1.40,3.1-3.20,4.1-4.18,5.2-5.9.

Wehner, Sigrid (1999) Exploring and Visualizing Event History Data. Materialien aus der Bildungsforschung Nr.65. Max-Planck-Institut für Bildungsforschung. Berlin, p.14.

Wehner, Sigrid (2000) Untersuchung der Stichprobengüte der Panelbefragung LVOST / 2. Welle. Arbeitspapier, Max Planck Institute for Human Development, Berlin.

Winship, Christopher /Mare, Robert D. (1992) Models for Sample Selection Bias. In: Annual Review of Sociology, Vol 18:327-350.