UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL AND HUMAN SCIENCES

Department of Social Statistics and Demography

Using Multilevel Models to Investigate Interviewer Effects on Nonresponse Bias and Measurement Error

by

Denize A. Barbosa

Thesis for the degree of Doctor of Philosophy

June 2015

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL AND HUMAN SCIENCES
Department of Social Statistics and Demography

Doctor of Philosophy

USING MULTILEVEL MODELS TO INVESTIGATE INTERVIEWER EFFECTS ON NONRESPONSE BIAS AND MEASUREMENT ERROR

by Denize A. Barbosa

Nonresponse and measurement error are common characteristics in almost all sample surveys. These nonsampling errors can negatively affect the quality of the sample estimates, because of the possibility of selection and measurement biases. One potential factor that can influence these sources of errors is the survey interviewer. This study aims to detect empirically interviewer effects on nonresponse bias and measurement error in sample surveys.

Multilevel modelling techniques are employed to variables of interest from real datasets from four different surveys. For the investigation of interviewer effects on nonresponse bias, the available datasets are quite rich since they contain linked auxiliary variables for respondents and nonrespondents from different sources. For the investigation of interviewer effects on measurement error on variables of interest, data from a survey and administrative records are linked to provide a dataset consisting of observed and potentially more accurate measures for essentially the same variables for all respondents.

The results suggested that there were significant nonresponse bias and evidence of interviewer effects on nonresponse bias for some of the dependent variables of interest. In addition, there was evidence of significant interviewer effects on measurement error.

This research provides intuitive approaches to assess interviewer effects on nonresponse bias and measurement error in variables of interest. The models discussed in Chapters 2 and 3 can also be used to identify and control for explanatory variables that may help to explain the interviewer–level variation and partially explain the nonresponse bias. Whilst the models discussed in Chapter 4 can be useful to monitor interviewers performance as well as to identify where to improve interviewer training to avoid the occurrence of measurement error.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Denize A. Barbosa , declare that the thesis entitled *Using Multilevel Models to Investigate Interviewer Effects on Nonresponse Bias and Measurement Error* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission

Signed:.....

Date:.....

I dedicate this thesis to my husband, Damião, and our son, Daniel, for their love, remarkable patience and support over the course of my PhD.

Acknowledgements

First and foremost, I would like to thank God for being my strength and guide during the writing of this thesis and throughout my life.

I would like to express my sincere gratitude to my supervisors, Prof. Peter W. F. Smith and Dr. Gabriele B. Durrant, for the continuous support to my PhD study and research. I am also thankful to their patience, motivation and all insightful discussions. Without their guidance and encouragement, this thesis would not have materialized.

My deepest gratitude to Prof. Chris J. Skinner who envisaged some ideas for this thesis, contributed with many comments and suggestions, was part of my supervisory team and kindly sponsored my first year of study in this university.

I am particularly grateful to my internal examiner, Dr. Nikos Tzavids, for his valuable advice and feedback on my upgrade exam.

My gratitude to Dr. Barry Schouten and Prof. Li–Chun Zhang for providing datasets that were used in some of the applications of the methods proposed in this thesis and for their endless patience in answering to all my queries about the datasets.

I wish to thank my employer in Brazil, Federal University of Rio Grande do Norte (UFRN), for having granted me this opportunity to conduct PhD studies at the University of Southampton. Special thanks to my colleagues at the Department of Statistics of UFRN, particulary Dr. André Pinho and Dr. Jeanete Moreira, for their support before coming here and during this whole journey.

I would like to thank our friends from Brazil for the emotional support and prayers dedicated to me and my family during these four years that we have been away from home. Many thanks also to our "local" Brazilian friends Henrique and Gabriella for having shared with us many enjoyable moments, helped us on numerous times since we arrived in this country and made us feel as part of their beautiful family.

Many thanks to my friends from the University of Southampton for their friendship, constant encouragement, exchanging of experiences and helping me to stay

sane during these four years. Thanks also to the faculty and members of the staff that contributed in any way for the realization of this thesis.

I am also thankful to the Brazilian research agency CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) for sponsoring me during the last two years of my PhD.

Finally, I would like to thank my family for all their love, encouragement and prayers. Special thanks to: my mother who raised me and my siblings on her own and always encouraged us to study; my mother–in–law Francisca and her sister Domicina for restless support during good and difficult times; my son Daniel who is our companion in all our adventures and is always a laugh with his remarkable sense of humour; and, most of all, to my loving, supportive, encouraging, and patient husband Damião whose faithful support and insightful discussions during the final stages of this PhD are so appreciated.

Thank you all!

Acronyms

BA	Bachelor of Arts
BDIC	Bayesian Deviance Information Criterion
BHPS	British Household Panel Survey
BSc	Bachelor of Science
BTEC/Edexcel	Business and Technology Education Council
CAPI	Computer Assisted Personal Interview
CATI	Computer Assisted Telephone Interviewing
CCS	Consumer Confidence Survey
EPSEM	Equal probability of selection method
ESRC	Economic and Social Research Council
ESS	European Social Survey
ESS–Norway	Norwegian sample of the European Social Survey
EU	European Union
GCSE	General Certificate of Secondary Education
IAS	Interviewer Attitude Survey
i.i.d.	Independent and identically distributed
ISER	Institute for Social and Economic Research
LFS	Labour Force Survey
MA	Master of Arts
MAR	Missing at random
MCAR	Missing completely at random
MCMC	Markov Chain Monte Carlo
`MLwiN`	Statistical software package for fitting multilevels models
MQL	Marginal quasi–likelihood
$N(0, \cdot)$	Normally distributed with zero mean and $\cdot$ variance
NMAR	Not missing at random
NVQ	National Vocational Qualification
ONS	Office for National Statistics
PGCE	Postgraduate Certificate in Education

PhD	Doctor of Philosophy
PQL	Penalized (or predicted) quasi–likelihood
RISQ	Representativity Indicators for Survey Quality
RSA/OCR	Oxford, Cambridge and RSA Examinations
TSE	Total Survey Error
UK	United Kingdom

Chapter 1

Introduction

Sampling surveys can be subject to sampling and nonsampling errors. Sampling error is the deviation of a survey estimate from its population value due to surveying a sample rather than the whole population. Nonsampling error can arise due to nonresponse, coverage error and measurement error. Nonresponse occurs when the required information from the sampled units cannot be completely obtained. Coverage error results from when the sample frame covers more than the target frame (over coverage) or less than the target frame (under coverage). Measurement error happens when there is a discrepancy between an observed measure and its true value.

This research discusses approaches to detect the effects of interviewers on nonresponse bias and on measurement error in sample surveys by applying multilevel modelling techniques. This first chapter gives an outline for the thesis, where the main research questions are specified. The chapter provides an introduction to nonresponse in sample surveys, the definition of bias, a discussion about nonresponse bias, the potential effect of the survey interviewer on nonresponse and an introduction to measurement error. Additionally, an overview of multilevel models, which are commonly used in the literature to analyse interviewer effects, is

presented. This chapter also gives a detailed description of the datasets used in the applications of the methods proposed.

1.1 Outline of the thesis

This study aims to detect empirically interviewer effects on nonresponse bias and on measurement error in sample surveys. For both cases, multilevel modelling techniques are employed to variables of interest from four different datasets. Since, in order to examine the interviewer effects on nonresponse bias on survey estimates, information on both respondents and nonrespondents has to be available, the datasets considered for this purpose contain auxiliary variables from different sources. The dataset used to assess interviewer effects on measurement error on dependent variables of interest also contains auxiliary information for all respondents since, for this analysis, observed and true information for these variables are required. In addition to this introductory chapter, this thesis is divided into four subsequent chapters.

Chapter 2 aims to introduce a novel and intuitive approach to assess interviewer effects on nonresponse bias. This approach consists of applying two–level random coefficient models for variables of interest using the response indicator as an explanatory variable rather than the dependent variable. To support the proposed method, an application considering two dependent variables, employment and academic qualification, from a linked dataset is described. These variables are used to test for nonresponse bias and for assessing interviewer effects on nonresponse bias. Information at the individual–level on both respondents and nonrespondents is acquired by linking each household member in the 2001 UK Labour Force Survey dataset to the 2001 UK individual census records (2001 UK Census Linked Study).

In Chapter 3, the aim is to investigate further the assessment of interviewer effects on nonresponse bias by making use of survey data from other collection modes and alternative sources of auxiliary variables. In contrast to Chapter 2, where the dataset used comes from the linkage of a cross–sectional face–to–face survey and individual census records, the focus in this chapter is on datasets from different data collection modes: a CATI survey, the Dutch Consumer Confidence Survey (CCS), and a longitudinal study, the British Household Panel Survey (BHPS). Another distinction from Chapter 2 is that these datasets are linked to other sources of auxiliary variables, since census records are not available. For the CCS, the source of auxiliary variables linked to the survey data in order to acquire information on respondents and nonrespondents is administrative data. Whilst for the BHPS, time invariant variables from its first wave are linked to respondents and nonrespondents from wave 10. The models proposed in Chapter 2 are applied to three variables, jobs in the household, type of household and size of household, from the CCS linked data and also to two variables, mother and father working indicators, from the BHPS.

In Chapter 4, the aim is to introduce an approach to assess interviewer effects on measurement error. Two–level random intercept models are applied to measurement error indicators for variables of interest, education level and household size, from the 2010 Norwegian sample of the European Social Survey data linked to administrative records. These measurement error indicators are created based on the joint distribution of essentially the same variables of interest from the survey and administrative data.

Finally, in Chapter 5, the aim is to present a general conclusion for all data analyses, summarise the limitations of this research and discuss possible ideas for future work.

The research questions of this study are:

- In Chapter 2, the models consider the response indicator as an explanatory variable rather than a dependent variable. Does this specification of the models help to assess the nonresponse bias driven by interviewers? Is there evidence of interviewer effects on nonresponse bias in the estimated proportion of individuals belonging to the categories of dependent variables of interest from the Labour Force Survey?

- In Chapter 3, how effective is telephone survey and other sources of auxiliary variables (i.e., administrative data and time invariant data from previous waves), as opposed to respectively face–to–face survey and census linked data for assessing interviewer effects on nonresponse bias?

- In Chapter 4, what type of categorical dependent variable (binary or multinomial) would be more appropriate to detect interviewer effects on measurement error by applying multilevel models?

1.2 Sampling and nonresponse framework

Consider a population $U = \{1, 2, \ldots, N\}$ of N units supposed here to be individuals, which could also be partitioned into strata or clusters of individuals. Let y_i and $\mathbf{x}_i$ denote the values of a survey variable y and a vector $\mathbf{x}$ of individual–level characteristics or explanatory variables for the i–th unit of the population.

Suppose a sample s of n units is drawn from the population U by a probability sampling design $p(\cdot)$. That means that s is the outcome of a random subset S selected from U with probability $p(s)$ and the sampling design $p(\cdot)$ is such that $p(s) \geq 0$ for all $s \subset U$ and $\sum_s p(s) = 1$. Based on the sample s, define the vector

of sampling inclusion indicators $\boldsymbol{I} = (I_1, I_2, \ldots, I_N)^\top$, where

$$I_i = \begin{cases} 1, & \text{if unit } i \text{ is included in the sample } s, \\ 0, & \text{otherwise,} \end{cases}$$

for all $i = 1, 2, \ldots, N$.

A common characteristic in almost all sample surveys is the presence of nonresponse. This problem occurs when there is failure in obtaining the measures that should be collected from the sample units selected for the study (Cochran, 1977). Sometimes, sample units do not provide any information requested in a survey (unit nonresponse), other times they provide only part of the requested information (item nonresponse). Common causes of nonresponse are noncontact with the sampling unit, the inability of the sample unit to provide a response and refusals to participate in the survey (Lynn et al., 2002).

The nonresponse literature identifies common characteristics from sampled units that may contribute to the fact they are less likely to respond a survey request, causing particular groups to be underrepresented in a sample. In many cases these characteristics depend on the survey topic. Generally, studies have highlighted that, among others, no children in the household, single household, male and younger age groups (Durrant and Steele, 2009; Durrant et al., 2010; Campanelli et al., 1997) are potentially associated with nonresponse.

When there is nonresponse, either unit or item, in a sample survey, there is always the possibility that the characteristics from respondents and nonrespondents differ. In this case, Bethlehem et al. (2011) suggest that steps can be made to adjust for the difference from respondents and nonrespondents, based on the knowledge about the cause of the missing data mechanism.

Little and Rubin (2002) distinguish the mechanisms that leads to nonresponse between Missing Completely at Random (MCAR), Missing at Random (MAR) and Not Missing at Random (NMAR). To define each of these mechanisms, suppose that only unit nonresponse happens in the survey and that the values of the auxiliary variables in the sample $\{\mathbf{x}_i : i \in s\}$ are fully observed. In this case, consider the vector of response indicators $\boldsymbol{R} = (R_i : i \in s)^\top$ where

$$R_i = \begin{cases} 1, & \text{if sampled unit } i \text{ responds to the survey,} \\ 0, & \text{otherwise,} \end{cases}$$

for all $i \in s$. Hence, y_i is measured only when $R_i = 1$. Let r_i be a possible value for R_i and let $\boldsymbol{r} = (r_i : i \in s)^\top$. An analogous response indicator could be defined for item nonresponse, but this thesis focusses on unit nonresponse only. Hence, with respect to missingness in a given survey variable y, the terminology in Little and Rubin (2002, p. 12) states the nonresponse process is MCAR, MAR and NMAR according whether the distribution of $\boldsymbol{R}$ does not depend on the sample data $\{(y_i, \mathbf{x}_i) : i \in s\}$, depends only on the observed data and depends on the missing data, respectively. Therefore, by using $f(\cdot|\cdot)$ to denote the conditional distribution of the response indicator vector and assuming these distributions could be indexed by a vector $\boldsymbol{\phi}$ of possibly unknown parameters, a MCAR nonresponse process means that

$$f(\boldsymbol{r}|\{(y_i, \mathbf{x}_i) : i \in s\}; \boldsymbol{\phi}) = f(\boldsymbol{r}|\boldsymbol{\phi}),$$

for all values of $(y_i, \mathbf{x}_i)$ $i \in s$ and all $\boldsymbol{\phi}$. Here and what follows (Y_i, R_i) are assumed to be independent across individuals. On the other hand, the MAR nonresponse mechanism implies that

$$f(\boldsymbol{r}|\{(y_i, \mathbf{x}_i) : i \in s\}, \boldsymbol{\phi}) = f(\boldsymbol{r}|\{y_i : i \in s, R_i = 1\}, \{\mathbf{x}_i : i \in s\}; \boldsymbol{\phi}),$$

for all possible values of $\{(y_i, \mathbf{x}_i) : i \in s, R_i = 0\}$ and all $\boldsymbol{\phi}$. Finally, in an NMAR mechanism, $f(\boldsymbol{r}|\{(y_i, \mathbf{x}_i) : i \in s\}, \boldsymbol{\phi})$ is a function not only of $\{(y_i, \mathbf{x}_i) : i \in s, R_i = 1\}$ and $\boldsymbol{\phi}$, but also of $\{(y_i, \mathbf{x}_i) : i \in s, R_i = 0\}$.

The probability of response

$$p_i = P(R_i = 1; \boldsymbol{\phi})$$

is commonly referred to as the propensity score for unit i. In more general terms, this probability could be written as

$$p_i = P(R_i = 1|\{(y_i, \mathbf{x}_i) : i \in s\}; \boldsymbol{\phi})$$

and may be termed as the propensity score of unit i, extending the terminology given in Rosenbaum and Rubin (1983) and David et al. (1983). Thus, when the p_i are a constant for all i, then nonresponse is MCAR. An important example of a MAR nonresponse is when the propensity scores $p_i = p(\{\mathbf{x}_i : i \in s\}, \boldsymbol{\phi})$. In the NMAR case, $p_i = p(\{(y_i, \mathbf{x}_i) : i \in s\}, \boldsymbol{\phi})$, that is the propensity scores depend on the missing y_i and cannot be completely explained by the observed data.

1.3 Nonresponse bias

The main consequence of nonresponse is the possibility of bias in the estimates produced. The bias of an estimator $\widehat{\varphi}$ of a parameter φ is defined as the difference between the expected value of the estimator and the true value of the parameter, that is

$$Bias\left[\widehat{\varphi}\right] = E\left[\widehat{\varphi}\right] - \varphi.$$

If $Bias\,[\widehat{\varphi}] = 0$, then the estimator $\widehat{\varphi}$ is unbiased for φ. In many cases, characteristics of interest in the study differ between respondents and nonrespondents and, therefore, analyses of the information collected only in the respondents may lead to distorted inferences about the population and model parameters of interest.

Before discussing how to handle data with a clustered structure, suppose that the pairs (y_i, r_i) are realizations of the random variables (Y_i, R_i) and that sampling is noninformative in the sense that the joint distribution of the (Y_i, R_i) does not depend on the outcome of the sampling. That is, the conditional distribution

$$f(y_i, r_i|\boldsymbol{I}) = f(y_i, r_i),$$

for all possible values of the sampling inclusion indicator $\boldsymbol{I} = (I_1, I_2, \ldots, I_N)^\top$. For simplicity, start by assuming that the pairs (Y_i, R_i) are identically distributed so that the nonresponse bias may be written as $E(Y_i|R_i = 1) - E(Y_i)$ or, alternatively, as $(1-p_i)\Delta$, where p_i is the propensity score of unit i and $\Delta = \mu_1 - \mu_0 \equiv E(Y_i|R_i = 1) - E(Y_i|R_i = 0)$. The regression of Y_i on R_i may then be expressed as

$$E\left(Y_i\,|R_i = r_i\,\right) = \mu_0 + \Delta r_i. \tag{1.1}$$

Hence, the nonresponse bias is closely related to the coefficient of r_i in the linear regression. If nonresponse is missing completely at random (MCAR), that is Y_i and R_i are pairwise independent, the nonresponse bias will be zero. More generally, it may be of interest to include a vector $\mathbf{x}_i$ of auxiliary variables in the regression for at least two reasons. First, if such variables are available at the sample or population level, they may be used for nonresponse adjustment, such as by weighting or imputation. In this case, the adjusted estimator may be approximately unbiased if nonresponse is missing at random (MAR) given $\mathbf{x}_i$, that is if Y_i and R_i are pairwise independent conditional on $\mathbf{x}_i$. Second, as considered below

$\mathbf{x}_i$ has a role in the assessment of interviewer effects. A possible linear regression model including conditioning on $\mathbf{x}_i$ is

$$E\left(Y_i \,|R_i = r_i, \mathbf{x}_i\right) = \mu_0 + \Delta r_i + \boldsymbol{\gamma}^\top \mathbf{x}_i, \tag{1.2}$$

where $\boldsymbol{\gamma}$ is a vector of regression coefficients associated with the components of $\mathbf{x}_i$.

The regressions in (1.1) and (1.2) relate to the conditional distribution of Y_i given R_i or given R_i and x_i and may be viewed as a pattern–mixture model in the terminology of Little (1993). These parametrizations are natural if one is interested in testing MCAR or MAR mechanisms. Suppose first that Y_i is observed for $r_i = 0$. Thus the hypotheses that the mechanism is MCAR or MAR correspond to the hypotheses that $\Delta = 0$ in model (1.1) or (1.2), respectively, and may be tested using methods of linear regression analysis. These hypotheses correspond, respectively, to the absence of nonresponse bias in an unweighted estimator or to the absence of nonresponse bias in a weighted or imputation estimator which adjusts for nonresponse bias under the assumption of MAR nonresponse. The issue of missing Y_i values for $r_i = 0$ in the usual nonresponse setting is overcome by acquiring "proxies" for those missing values that are available in the linked data.

The same type of test from the above discussion involving (1.1) and (1.2) still holds if Y_i is a binary variable. In the case of a binary Y_i, the left–hand side of Equation (1.2) may be written as

$$g\left\{E\left(Y_i \,|R_i = r_i, \mathbf{x}_i\right)\right\} = \mu_0 + \Delta r_i + \boldsymbol{\gamma}^\top \mathbf{x}_i,$$

where $g\{.\}$ is an appropriate link function such as the logit.

1.4 Measurement error

Measurement error mostly arises from four sources: interviewer, respondent, questionnaire and the mode of data collection (Groves, 2004, p. 295). Examples of possible causes for this error can include, but are not limited to: differences that may occur in reactions of respondents to different interviewers (Ruddock, 1998), e.g. to interviewers of their own sex or own ethnic group; answers given by respondents may be influenced by the desire to impress an interviewer or their answers do not always reflect their true beliefs because they may feel under social pressure not to give an unpopular or socially undesirable answer, e.g. the use of illicit drugs may be underreported (Mensch and Kandel, 1988); inadequate interviewer training; the wording of questions may be unclear, ambiguous or difficult to answer, e.g. it may require remembering past dates or facts; and respondents may answer questions differently depending on the manner in which the questionnaire is administered, whether in the presence of an interviewer, via telephone or self–administered (Lepkowski, 2004). For instance, in an internet survey on self–reported sexual behaviour, men reported significantly higher sociosexuality than women. However, there was no difference between men and women's reports when the concern about confidentiality was not compromised (Beaussart and Kaufman, 2013).

Measurement error is an undesirable feature because it may negatively affect the quality of the data and yield inaccurate estimates (Biemer and Lyberg, 2003). For example, consider the following simple situation. Suppose an observed variable X_i containing measurement error can be related to its true value x_i, that is the value that would have been obtained in the absence of measurement error, by the linear model

$$X_i = x_i + e_i, \tag{1.3}$$

where e_i is a random variable describing the measurement error of the i–th observation. The unobserved variable x_i is sometimes called *latent variable* and it may be a covariate in a regression coefficient, such as

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{1.4}$$

where β_0 and β_1 are regression coefficients and ϵ_i is the regression error term associated with the i–th observation. Since x_i is unobserved, the regression coefficients must be estimated under the basis of the observed data $\{(Y_i, X_i)\}$. A key general reference providing theoretical results for the treatment of measurement error in covariates of regression models is Fuller (1987). Under some assumptions for the distribution of (x_i, e_i, ϵ_i), the Ordinary Least Squares estimator of β_1 obtained by replacing x_i by X_i is biased toward zero. This problem, known as *attenuation*, may lead one to declare a non–significant effect of x_i on the regression of Y_i given x_i when this effect may, in fact, be significant. Extensions to the treatment of measurement error in nonlinear regression problems are discussed by Carroll et al. (2006).

In the case where both observed and "true" values for a variable are categorical, the term misclassification is commonly used to denote the existing measurement error (Buonaccorsi, 2010). In this context, let W be an observed (error prone) measure for a variable X. Assuming that W and X are binary, the measurement error model is specified by the misclassification probabilities

$$\theta_{w|x} = P(W = w|X = x),$$

with $\theta_{1|1} = P(W = 1|X = 1)$ denoted as *sensitivity* and $\theta_{0|0} = P(W = 0|X = 0)$ denoted as *specificity*. $\theta_{1|1}$ and $\theta_{0|0}$ are the probabilities of an observation being correctly classified, whereas $\theta_{1|0}$ and $\theta_{0|1}$ are the probabilities of an observation

being misclassified.

To assess measurement error in variables of interest from sample surveys, generally the survey responses are compared with potentially more accurate data from other sources for the same respondents, since the true population values are normally unknown. These sources usually involve auxiliary variables from census records, data from re–interviews or administrative data. As such, to ensure information accuracy, the time points between the data collection from the survey and the information gathered from the auxiliary variables should be as close as possible. In addition, it is important that variables of interest have compatible definitions in the survey and in the other sources.

1.5 Interviewer effects

In sample surveys there are several factors that have a negative impact on the quality of an estimator of a population parameter. Examples of these factors can be the poor quality of the data, small sample sizes and ineffective coverage of the target population, among others. These factors can contribute to errors that belong to a more comprehensive class of errors called Total Survey Error (TSE), which can be defined as consisting of basically two components: the sampling and nonsampling errors (Biemer and Lyberg, 2003). This class provides a framework for maximising data quality in the survey. Biemer (2010) suggests an approach to efficiently allocate the survey budget to minimising the TSE and, as a consequence, maximising the data quality. However, even if it would be possible to eliminate the sampling error component of the TSE, for example, by drawing a census of the population rather than a sample (which may be impracticable since it is usually quite costly), nonsampling errors could still occur as a result of nonresponse, coverage error and measurement error, among other sources.

The survey interviewers may have an influence in all sources of nonsampling errors. For example, some interviewers' characteristics may impact the decision of sampled units to participate in a survey (Hox and De Leeuw, 2002; Durrant et al., 2010). Interviewers may be somehow prevented to visit all sampled units or they may deliberately choose not do so. Also, their persuasion skills may induce responses that do not correspond to the truth from reluctant respondents (Biemer, 2001; Groves, 2006; Fricker and Tourangeau, 2010). Therefore, interviewers may play a significant role in the TSE and, consequently, in the quality of the survey estimates.

In interviewer administered surveys, which is the focus of this research, one factor that can potentially influence nonresponse is the survey interviewer, since interviewers are supposed to use their abilities and strategies to convince sample units to participate in a survey. Brunton-Smith et al. (2012) argue that some interviewers use inadequate strategies to approach the sampling units and, therefore, this may result in one of the common causes of nonresponse. They also emphasise that efforts to reduce nonresponse have been undertaken by survey organisations by intensifying the interviewer training.

Interviewers may have an effect on nonrespondents when the interviewers' characteristics and attitudes can somehow prevent some sampled units to respond to a survey request or to some specific items. In addition, if the respondent pool has characteristics that differ from those in the nonrespondent pool, this can lead to nonresponse bias. Interviewers may also have influence on respondents when: (i) in the interaction with interviewers, respondents provide socially desirable answers for sensitive questions, for example, reports of drug use (Mensch and Kandel, 1988); (ii) interviewer characteristics are related to the survey topic, an example is a study by Hatchett and Schuman (1975) in the USA in which African–American interviewers obtain less negative behaviour reports towards African Americans

from white respondents; and, even though controversial, (iii) experienced interviewers tend to rush reading questions and fail following protocols compared to new interviewers (Bradburn et al., 1979).

Despite of the interviewer training, the interaction between interviewers and sampling units is still an enigma, since it involves persuasion strategies (interviewer), decision–making processes (sampling units), interviewers' attributes and social environment. Additionally, some of these factors are still out of the control of the survey researchers (Groves and Couper, 1998). Thus, many studies link interviewer characteristics to nonresponse. In their paper Hox and De Leeuw (2002) investigate whether interviewer response rates are predicted by their attitude and behaviour. In the analysis, they apply multilevel models to a nonresponse indicator outcome and use interviewers' characteristics as explanatory variables. Although they conclude that (non)response rates vary across interviewers, they do not find strong evidence that interviewers' attitude and behaviour predict (non)response rates.

Although in a different scenario, Pickery and Loosveldt (2002) consider a multilevel multinomial model to analyse interviewer effects on different types of nonresponse. In their paper, three possible outcomes for an interview are considered: completed interview, refusal and noncontact. In addition to control for respondents' characteristics in the model, they also control for interviewers' characteristics. The analysis showed that both chances for refusals and for noncontacts are subject to interviewer effects. They found evidence of a relation between refusal and noncontact: interviewers who report more noncontacts are more likely to report refusals. This conclusion was possible because a multinomial multilevel model was used, i.e., the same conclusion would not be possible if a dichotomous response model had been applied.

Researchers also associate interviewer performances and characteristics to measurement error and consequently to poor data quality (Groves, 2004; Biemer et al., 1991). For instance, Kennickell (2002) investigates the effects of interviewers on data quality in a survey. He addresses indicators of quality such as editing of survey data, resetting incorrect variables to missing values and number of times respondents refused to give an answer or gave a "don't know" answer, focussing on the role of interviewers. He emphasises that the interviewers' performance on gathering quality data may be misleading since usually particularly good interviewers may be assigned to somehow difficult cases. As a main finding, interviewer effects were significant for all fitted models.

Due mainly to the urge to reduce administration costs, surveys organisations have increasingly choose to collect data through multiple modes, leaving interviewer administered mode as a last resource because of its high cost. However, this data collection mode is still the most popular one since it ensures larger response rates. As a disadvantage of this mode, Schouten et al. (2013) argue that in interviewer administered surveys, respondents may be prone to give socially desirable answers that may differ from the truth or they may be inhibited to elaborate their complete answers, leading to measurement error. In their paper, they also mention the speed of the interview that may vary depending on the mode of data collection, e.g., telephone, web or face–to–face survey, as another contributing factor to exasperate measurement error.

The application of multilevel techniques to investigate interviewer effects on sources of nonsampling errors is quite common in the literature. To give a few examples of papers that employ these models to analyse the effects of interviewers on different types of survey nonresponses and on measurement error, one can refer to O'Muircheartaigh and Campanelli (1999), Pickery and Loosveldt (2004), Durrant et al. (2010), Blom et al. (2011) and Sinibaldi et al. (2013). However, research

is still needed to investigate the effects of interviewers on nonresponse bias and measurement error by applying multilevel approaches.

In this study, approaches are proposed to detect interviewer effects on nonresponse bias and on measurement error by applying multilevel models. The literature suggests that the effects of interviewers on these two errors may be linked since significant interviewer effects can arise from nonresponse bias and/or measurement error. This linkage can happen, especially when there is an extra effort from the interviewers to convert refusals into responses (Olson, 2006), as well as in the case of questions about sensitive or attitudinal items (Beaussart and Kaufman, 2013). In some cases, this persuasion results in a response that may differ from the truth. West and Olson (2010) have a goal on analysing simultaneously the effects of interviewers on these two sources of error by proposing a procedure to quantify how much of the interviewer variation is due to nonresponse error and how much is due to measurement error. Their dataset contains the same variables from the survey and from a frame of certificates for both respondents and nonrespondents. Although their approach is quite interesting, the datasets that are used in this research for the investigation of interviewer effects on nonresponse bias (Chapters 2 and 3) do not contain the same variables from the surveys and from other more accurate sources. The variables considered here come from auxiliary sources (e.g., census records and administrative data). On the other hand, the dataset used in the investigation of interviewer effects on measurement error (Chapter 4) do contain the same variables from the survey and from a more reliable source. However, the approach that is used for this investigation takes into account only the respondents. For these reasons, this study treats the effects of interviewers on these two sources of error separately.

1.6 Datasets

In the investigation of interviewer effects on nonresponse bias and measurement error, datasets from four different surveys are used. These datasets come from various types of surveys such as face–to–face and telephone surveys as well as cross–sectional and longitudinal surveys. This can provide a richer source of application for the models that are described in Section 1.7 and which are considered specifically in Chapters 2, 3 and 4, given the particular features involved in each survey.

In order to investigate the effects of interviewers on nonresponse bias, three datasets are considered. In Chapter 2, data from the 2001 UK Labour Force Survey (LFS) linked to the 2001 census records are used in an application of the models proposed in that chapter. Two other applications, which are presented in Chapter 3, use data from a telephone survey linked to administrative data carried out in the Netherlands, namely the Consumer Confidence Survey (CCS), and data from the British Household Panel Survey (BHPS).

Furthermore, for the investigation of interviewer effects on measurement error, presented in Chapter 4, data from the 2010 Norwegian sample of the European Social Survey (ESS) linked to administrative data are used. Additional details on these datasets can be found in the next sections. However, details regarding the analysis samples used in the applications of the models that are proposed for each investigation as well as nonresponse matters are discussed in Chapters 2, 3 and 4.

1.6.1 Labour force survey

This research makes use of data from the UK LFS. This is a quarterly sample survey of households conducted by the Office for National Statistics in England,

Scotland, Wales and, since 1994, Northern Ireland. This survey was first carried out in 1973 and was repeated every two years up to 1983. Then, from 1984 to 1991 it was carried out annually and since 1992 it has been running as a quarterly survey. The LFS provides information on various aspects of the UK labour market, including employment, unemployment and economic activity rates. It also covers a range of related topics, such as income, qualifications, training and disability.

The sample design is an unclustered sample of households who live at private addresses in the UK. The quarterly survey has a panel design whereby individuals stay in the sample for 5 consecutive quarters (or waves), with a fifth of the sample replaced each quarter. Thus, there is an 80% overlap in the samples for each successive survey. At a sampled address, it aims to interview every household member aged 16 and over.

1.6.2 Consumer confidence survey

Another application discussed in this study utilizes data from the Consumer Confidence Survey (CCS), which is a Computer Assisted Telephone Interviewing (CATI) survey from the Netherlands. This survey is used in the European Representativity Indicators for Survey Quality (RISQ) Project (for more details on the EU RISQ Project see: `http://www.risq-project.eu/index.html`). The CCS was conducted monthly throughout 2005 and carried out by the Statistics Netherlands.

The CCS is a two–stage sample, in which municipalities are the clusters in the first stage. In the second stage, addresses are drawn from the clusters, using simple random sample without replacement. The data are collected by interviewing one person from a selected household and asking for his/her opinion on the state of the economy. Interviewers worked in a centralised telephone unit, where their CATI

management system automatically assigned phone numbers to interviewers each day.

Groves and Magilavy (1986) discuss that CATI system provides a way, with relatively low cost, to have approximately interpenetrated designs, which are based on the idea of random assignment of sampled units to interviewers (Mahalanobis, 1946). This setting helps in the evaluation of interviewers and to avoid the potential confounding effects between interviewers and geographic areas where sampled units live since interviewers are usually assigned to specific areas. Therefore, a significant interviewer effect, using an interpenetrated design, is more likely to be a "pure" interviewer effect than in the case of other designs. On the other hand, they emphasise that interviewer effects in telephone surveys tend to be attenuated compared to face–to–face surveys, maybe because there is less room for interviewer–respondent interaction through telephone than face–to–face (West et al., 2013). Also, they find that the age of the respondents is related to interviewer effects, with older respondents more likely to be influenced by interviewers than younger ones.

1.6.3 British household panel survey

This study also makes use of data from the British Household Panel Survey (BHPS). The BHPS had its first wave in 1991 and has been conducted annually since then. Its first sample consisted of approximately 10,000 interviewed individuals in around 5,500 households. The data are collected using a stratified clustered sample design, selected from the UK Postcode Address File. In the sampled addresses, all household members aged 16 or over are interviewed. In addition, from 1994 onwards children aged 11 through 15 from these households are

also asked to complete a short interview. This survey covers topics such as household composition, housing conditions, residential mobility, education and training, health and the usage of health services, labour market behaviour, socio–economic values, income, benefits and pensions.

Since the BHPS is a panel survey, the same individuals are followed in successive waves. If some household member from the original sample forms a new household, all adults (16+) from this new household are also interviewed. Moreover, there were some extension samples over the years in the BHPS. To enable individual country analysis and comparison between countries within the UK, in 1999 1,500 households in each of Scotland and Wales were added to the main BHPS sample. Also, a sample of 2,000 households was added in Northern Ireland in 2001.

As is argued in Groves and Peytcheva (2008), to examine nonresponse bias on survey estimates of interest, such as estimates of means and percentages, auxiliary information for respondents and nonrespondents should be available. Therefore, for the investigation of interviewer effects on nonresponse bias (Chapters 2 and 3), information on the dependent and explanatory variables at individual–level for both respondents and nonrespondents was acquired by linking each household member in the LFS dataset to the 2001 UK individual census records. In the case of the CCS and BHPS, linked census records are not available. However, other types of auxiliary variables that also provide information on respondents and nonrespondents are considered. These variables are administrative data and time invariant variables from previous waves. The linkage of survey variables with auxiliary variables containing individual or household–level information on respondents and nonrespondents will be discussed further in Chapters 2 and 3.

1.6.4 European social survey

Yet another dataset used in this study comes from the 2010 Norwegian sample of the European Social Survey (ESS) (for more details on the ESS see: `http://www.europeansocialsurvey.org`). The ESS has been carried out every two years since 2002. The most recent round of this survey was in 2012 and approximately 30 countries took part in each round. The ESS was mainly initiated to provide a reliable source of cross–national data collected with highly methodological standards to enable to draw inferences about changes over time in Europe. Thus, this survey aims to explain changes in attitude and behaviour patterns in Europe and improve methods of survey measurements across countries.

The ESS collects information on media use, social and public trust, political interest and participation, socio–political orientations, moral, political and social values, national, ethnic and religious allegiances, well–being, health and security, demographics and socio–economics.

The next section introduces multilevel models since multilevel techniques are often used to analyse interviewer effects.

1.7 Multilevel models

Regression analyses based on unclustered data is often handled in the survey practice by appropriate standard (single–level) regression models. For example, suppose the sample observations are denoted by $(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), ..., (y_n, \mathbf{x}_n)$, where $y_i \in \{0, 1\}$. In this case, a possible model that can be applied to analyse the data is a single–level logistic regression model. This model can be defined by specifying two components. The first is a stochastic component that assumes that $y_1, y_2, ..., y_n$

are realizations of independent random variables $Y_1, Y_2, ..., Y_n$ by which

$$Y_i \mid \mathbf{x}_1, ..., \mathbf{x}_n \ \sim \ \text{Bernoulli}(\pi_i), \quad i = 1, ..., n, \tag{1.5}$$

where $\pi_i = \pi(\mathbf{x}_i) = E(Y_i \mid \mathbf{x}_1, ..., \mathbf{x}_n)$. The second component, the systematic part of the model, assumes that there is a vector of unknown regression coefficients $\boldsymbol{\beta}$, of the same dimension as the vectors of covariates $\mathbf{x}_i$, such that

$$\text{logit}\,(\pi_i) = \mathbf{x}_i^\top \boldsymbol{\beta}, \quad i = 1, ..., n, \tag{1.6}$$

where $\text{logit}(p) = \log\{p/(1-p)\}$. The model defined in Equations (1.5) and (1.6) is a generalized linear model (Nelder and Wedderburn, 1972) with binomial response and *logit* link. Other binary regression models in this family can be specified by changing the link function in (1.6). Some other possible choices in practice are the probit, the complimentary log–log and the log–log links, which are defined respectively by the functions $\Phi^{-1}(p)$, $\log\{-\log(1-p)\}$ and $-\log\{-\log(p)\}$ (McCullagh and Nelder, 1989, p. 108), where Φ is the cumulative distribution of a Normal probability function.

One first difficulty to apply single–level regression models in sample surveys comes from the facts that survey data are usually collected by subjecting units to grouping or clustering and observations within a given cluster are usually more similar than observations between clusters. Clustering in the survey data can happen due to either the sample design or the hierarchical nature of the survey variable of interest. For instance, in a study to analyse a characteristic of interest (e.g., an attitudinal or sensitive item or a neighbourhood related issue) from individuals where the data are collected through interviewers, if interviewers' characteristics have influence on the responses from individuals, individuals interviewed by the same interviewer tend to give more similar responses than individuals interviewed

by different interviewers. In addition, surveys interviewers are usually assigned to specific geographic areas, so depending on the characteristic of interest, individuals living in the same area may have points of views alike. Thus, individuals' responses may be nested within interviewers and interviewers may or may not be nested within areas. This means that in addition to individuals' characteristics, their responses may also depend on the interviewer and on the area that the individuals live. Therefore, the analysis of this characteristic of interest must take into account the clustering structure since the dependence of the observations within clusters violates the independence assumption of the single–level regression models (Maas and Hox, 2005).

A second problem to consider with the application of a single–level regression model to some surveys is that it may be of interest, for instance, to control for difference among the survey interviewers and areas in the statistical model. The single–level model could indeed be specified by creating dummy variables for interviewers and areas. However, there will be a large number of parameters to estimate and interpret if the number of interviewers and/or areas is large. A potential alternative to overcome this problem would be by including these interviewers and areas in the model as random effects at different levels, provided that these random effects are uncorrelated with possible explanatory variables.

A class of statistical models that can be applied to circumvent the two problems aforementioned is the multilevel models. These type of models provide a set of flexible tools for modelling clustered discrete and continuous survey data, linear and nonlinear relationships and fixed and/or random effects. These models can also be applied in settings such as the analysis of longitudinal data, in which the repeated observations for each individual are considered to be the level–one variables whereas individuals are the level–two variables. A detailed account of

multilevel models, their historical development and applications is given in Goldstein (2011) and Hox (2010), among others. In general, a multilevel model allows one to take into account the existing dependence within each level of clustering, which if ignored may harm inferences as a consequence of obtaining statistically inefficient estimates of regression coefficients, incorrect standard errors, less conservative confidence intervals and significance tests (Goldstein, 2011).

Since the datasets used in this study may potentially have a clustered structure, where individuals are considered to be nested within interviewers, the multilevel modelling technique is a natural choice to analyse these datasets. In a broader context, consider level–one units nested within level–two units. In the example above, individuals are the level–one units and interviewers are the units at level–two. Consider also dependent variables defined as binary outcomes such that

$$y_{ij} = \begin{cases} 1, & \text{if level–one unit } i \text{ within level–two unit } j \text{ has a characteristic of interest} \\ 0, & \text{if level–one unit } i \text{ within level–two unit } j \text{ does not have the characteristic,} \end{cases}$$

where $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$, and let x_{ij} be a level–one unit characteristic (explanatory variable). The dependent variables may correspond to binary indicators of education and employment, for instance. One of the multilevel models considered for these variables is the two–level random intercept logistic model. This model can be specified by introducing a vector of random effects $\boldsymbol{u}_0 = (u_{01}, ..., u_{0J})^\top$ and assuming that

$$\begin{aligned} y_{ij} \mid x_{ij}, \boldsymbol{u}_0 &\sim \text{ indep Bernoulli}(\pi_{ij}), \quad i = 1, \ldots, n_j \text{ and } j = 1, \ldots, J, \\ \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) &= \beta_{0j} + \beta_1 x_{ij}, \\ \beta_{0j} &= \beta_0 + u_{0j}, \quad u_{0j} \sim \text{ i.i.d. } N(0, \sigma^2_{u_0}) \quad j = 1, \ldots, J, \end{aligned} \tag{1.7}$$

where the notation "$\sim$ indep" and "$\sim$ i.i.d." denote "has independent distributions" and "has independent identically distributed distributions", $\pi_{ij} = P(y_{ij} = 1 \mid x_{ij}, u_{0j})$ is the probability of the dependent variable y_{ij} having the characteristic, the β_{0j} are level–two unit dependent intercept, β_1 is the coefficient of the explanatory variable, the u_{0j} are random effects representing unexplained level–two unit effects and the variance parameter $\sigma^2_{u_0}$ is the residual between level–two unit variance in the log–odds of the dependent variable.

Model (1.7) allows the intercept to vary across level–two units, whereas the coefficient of x_{ij} is the same between level–two units. This model may be extended to allow the coefficient to vary across level–two units. For this extension, consider the set of random effects $\boldsymbol{u} = \{\boldsymbol{u}_1, ..., \boldsymbol{u}_J\}$, where $\boldsymbol{u}_j = (u_{0j}, u_{1j})^\top$ for all $j = 1, ..., J$. The resulting two–level random coefficient logistic model could be written as

$$\begin{aligned} y_{ij} \mid x_{ij}, \boldsymbol{u} \ &\sim \text{ indep Bernoulli}(\pi_{ij}), \quad i = 1, \ldots, n_j \ \text{ and } \ j = 1, \ldots, J, \\ \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) &= \beta_{0j} + \beta_{1j} x_{ij}, \\ \beta_{0j} &= \beta_0 + u_{0j}, \\ \beta_{1j} &= \beta_1 + u_{1j}, \\ \boldsymbol{u}_j = (u_{0j}, u_{1j})^\top &\sim \text{ i.i.d. } N(\mathbf{0}, \boldsymbol{\Omega}_u) \ \ j = 1, \ldots, J, \end{aligned} \tag{1.8}$$

where β_{0j} is the same as in (1.7), β_{1j} is the level–two unit dependent coefficient of the explanatory variable, the $\boldsymbol{u}_j = (u_{0j}, u_{1j})^\top$ are vectors of random effects representing unexplained level–two unit effects assumed to follow the bivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Omega}_u)$ with mean vector zero and covariance matrix

$$\boldsymbol{\Omega}_u = \begin{bmatrix} \sigma^2_{u_0} & \sigma_{u_0 u_1} \\ \sigma_{u_0 u_1} & \sigma^2_{u_1} \end{bmatrix}.$$

In this covariance matrix, the variance parameters $\sigma^2_{u_0}$ and $\sigma^2_{u_1}$ are respectively

the variation in the intercepts across level–two units and the variation in the coefficients across level–two units, whereas $\sigma_{u_0u_1} = cov(u_{0j}, u_{1j})$ is the covariance between random intercept and random coefficient effects.

If the clustering structure has three levels, Model (1.7) could easily be extended to a three–level random intercept logistic model. In this case, let y_{ijk} denote the binary indicator that the level–one unit i within level–two unit j within level–three unit k has the characteristic of interest and let x_{ijk} be a level–one unit explanatory variable, $i = 1, ..., n_{jk}$, $j = 1, ..., J$ and $k = 1, ..., K$. Considering vectors of random effects $\boldsymbol{u}_0 = (u_{011}, ..., u_{0JK})^\top$ and $\boldsymbol{v}_0 = (v_{01}, ..., v_{0K})^\top$, a possible three–level random intercept logistic model would consider

$$
\begin{aligned}
y_{ijk} \mid x_{ijk}, \boldsymbol{u}_0, \boldsymbol{v}_0 \;&\sim\; \text{indep Bernoulli}(\pi_{ijk}), \quad i = 1, \ldots, n_{jk}, j = 1, ..., J \;\text{ and} \\
&\qquad\qquad\qquad\qquad\qquad\qquad k = 1, \ldots, K, \\
\log\left(\frac{\pi_{ijk}}{1-\pi_{ijk}}\right) &= \beta_{0jk} + \beta_1 x_{ijk}, \\
\beta_{0jk} &= \beta_{00k} + u_{0jk}, \\
\beta_{00k} &= \beta_{000} + v_{00k}, \\
\boldsymbol{u}_0 &= (u_{011}, ..., u_{0JK})^\top \sim N(\mathbf{0}, \boldsymbol{\Omega}_{u_0}), \\
\boldsymbol{v}_0 &= (v_{01}, ..., u_{0K})^\top \sim N(\mathbf{0}, \boldsymbol{\Omega}_{v_0}),
\end{aligned}
\tag{1.9}
$$

where β_{0jk} is the intercept in level–two unit j within level–three unit k, β_{00k} is the average intercept in level–three unit k, the u_{0jk} and v_{00k} are the level–two and level–three unit random effects with covariance matrices $\boldsymbol{\Omega}_{u_0}$ and $\boldsymbol{\Omega}_{v_0}$, respectively.

Models (1.7) and (1.8) may also allow for the cross–classification of two higher level units. For example, interviewers may not be strictly nested within geographic areas since interviewers may work in more than one area and some areas may be covered by more than one interviewer. In this case these two higher level effects may be confounded with each other. Hence, level–one unit outcomes are nested within a

cross–classification of level–two unit effects. The multilevel cross–classified logistic model (Goldstein, 2011) may be written with systematic component as

$$\log\left(\frac{\pi_{i(jk)}}{1-\pi_{i(jk)}}\right) = \beta_0 + \beta_1 x_{i(jk)} + u_{1j}x_{i(jk)} + u_{0j} + v_{0k}, \tag{1.10}$$

where the notation $i(jk)$ means that level–one unit i is nested within a cross–classification of level–two unit j and level–two unit k. The other terms of the model have already been explained for the models in Equations (1.7) and (1.8) except for v_{0k}, denoting a random effect representing unexplained level–two unit k effects.

Model (1.7) can also be extended to a two–level random intercept multinomial model to fit categorical dependent variables with three or more unordered categories. If the dependent variable has C categories, the systematic component of the model could be specified as

$$\begin{aligned} \log\left(\frac{\pi_{ij}^{(c)}}{\pi_{ij}^{(1)}}\right) &= \beta_{0j}^{(c)} + \boldsymbol{\beta}_1^{(c)\top}\mathbf{x}_{ij}, \\ \beta_{0j}^{(c)} &= \beta_0^{(c)} + u_{0j}^{(c)}, \end{aligned} \tag{1.11}$$

where $c = 2, \ldots, C$. In Model (1.11) there are $C-1$ equations referring to the number of categories of the dependent variable minus one, where the $C-1$ response probabilities are each compared to the response probability in the reference category. The $(C-1)\times 1$ vector $\boldsymbol{u}_{0j} = (u_{0j}^{(2)}, ..., u_{0j}^{(C)})^\top$ is normally distributed with mean vector zero and covariance matrix $\boldsymbol{\Omega}_{u_0}$ given by

$$\boldsymbol{\Omega}_{u_0} = \begin{bmatrix} \sigma^2_{u_0^{(2)}} & & & \\ \sigma_{u_0^{(3)}u_0^{(2)}} & \sigma^2_{u_0^{(3)}} & & \\ \vdots & \vdots & \ddots & \\ \sigma_{u_0^{(C)}u_0^{(2)}} & \sigma_{u_0^{(C)}u_0^{(3)}} & \cdots & \sigma^2_{u_0^{(C)}} \end{bmatrix}.$$

1.7.1 Estimation methods

Parameters of interest in multilevel models are in many cases estimated by using maximum likelihood methods. For instance, Schall (1991) proposes a general algorithm for estimating fixed effects, random effects and variance components for generalized linear models with random effects. His algorithm yields approximate maximum likelihood estimates of the fixed effects and variance components and approximate empirical Bayes estimates of the random effects. In his paper, he describes two algorithms to obtain maximum likelihood and restricted maximum likelihood estimates for normal linear model with random effects and identity link function. He adapts these two algorithms so that maximum likelihood and restricted maximum likelihood estimates can also be obtained in non–normal and non–identity link generalized linear models with random effects.

In some of the analyses in this study, quasi–likelihood approaches are used to approximate the nonlinear link (e.g., logit) function considered here. In these approaches, the nonlinear function is linearised using a Taylor series expansion, which approximates this function by an infinite series of terms. The Taylor approximation is referred to as either first or second order depending on how many terms of the series are used. Since the Taylor series linearisation of a nonlinear function depends on the values of its parameters, the Taylor series expansion can use the estimated values of the fixed part only, which is referred to as marginal quasi–likelihood (MQL), which was proposed by Goldstein (1991). Alternatively, it can be improved by using both the estimated values of the fixed part and the residuals, which is referred to as penalized (or predicted) quasi–likelihood (PQL), which was examined by Green (1987) for general semi–parametric regression analysis and by Breslow and Clayton (1993) in the context of generalized linear mixed models.

In general, the second order PQL method provides more accurate estimates than the first order MQL, specifically, in the case where the level–two samples are small and the random effects are large. For multilevel models with binary responses, Goldstein and Rasbash (1996) demonstrate that the second order PQL leads to much better estimates than the first order MQL method. However, in this context, PQL can still be biased since it tends to underestimate the variance components and fixed effects (Breslow and Clayton, 1993).

Alternatively to maximum likelihood approaches for estimating parameters from multilevel models, there are, among others, the Bayesian methods. In Bayesian statistics, a probability distribution of possible values is assigned to express the uncertainty about the population value of a model parameter. This probability distribution is called the prior distribution, because it is specified independently from the data. After the collection of the data, this prior distribution is combined with the likelihood of the data to produce a posterior distribution, which describes the uncertainty of the population values after observing the data. The variance of the posterior distribution is usually smaller than the variance of the prior distribution.

For the prior distribution there are two options: an informative prior or an uninformative prior. The prior chosen can either strongly influence the posterior distribution and consequently the conclusions (informative prior) or have very little influence on the conclusions, serving only to produce the posterior (uninformative prior). When the posterior distribution is difficult to describe mathematically, it is approximated using Markov Chain Monte Carlo (MCMC) simulation techniques that generate random samples from a complex posterior distribution of the unknown parameters.

In this study, the models parameters, in most cases, are estimated using MCMC with a diffuse prior and with starting values obtained by using 2nd order PQL.

MCMC was chosen because it yields more accurate estimates compared to PQL (Goldstein, 2011). Browne and Draper (2006) compare various estimation methods, including quasi–likelihood methods (such as 1st order MQL and 2nd order PQL) and Bayesian methods (MCMC with several diffuse prior distribution), for different datasets and verify that, especially for multilevel models with binary responses, the estimates obtained from MCMC are closer to the true parameter values than the ones obtained from quasi–likelihood approaches.

For information on other estimation methods and a fuller description for the ones presented in this thesis one can refer to, for instance, Hox (2010), Gelman and Hill (2007) and Goldstein (2011).

1.7.2 Residual plots

As in multiple regression analyses, it is also strongly recommended to check whether or not the model assumptions are met in multilevel regression. In the latter model, the verification of the posited assumptions usually requires more work than in the traditional fixed regression coefficient setting because of the additional random error terms that are specific to the model levels. For example, the random effects u_{0j} in the logistic multilevel model (1.7) are the level–two error terms as they account for residual deviations from the overall intercept β_0. Similarly, the random effects u_{0jk} and v_{00k} in the three–level random intercept logistic model (1.9) are error terms representing residual deviations of the overall intercepts at levels two and three, respectively. Predictions for these error terms are seen then as estimated residuals or, simply, residuals.

Based on the sets of residuals obtained at a given level of a fitted multilevel model, there are a number of plots that can be constructed to check violations in the model assumptions and detect the presence of outliers. For instance, the

assumption that the random effects of models such as (1.7), (1.8) and (1.9) are a random sample from normal distributions can be verified by a quantile–quantile plot of the corresponding standardised residuals. This procedure plots the sorted values of the standardised residuals in increasing order against standard normal scores. If the normality assumption is correct, the expected pattern in the plot is that the points lie close to a straight diagonal line.

A quantile–quantile plot may also indicate if there are outlying or influential residuals, which would appear in the plot to diverge from a straight diagonal line at the extremities. If that is the case, one should look for explanations for the outliers and whether they resulted from obvious mistakes, such as errors of data coding, and whether they could be fixed before analysis. Additionally, a simple analysis that can be carried out is by refitting the model without the outliers to check for the effect of omitting them on the model parameter estimates. Outlying observations could also be identified by examining boxplots, if there are enough data, and by dotplots, otherwise.

A plot of the residuals of a given level against the corresponding predicted values of the fixed part of the model is useful to check the assumptions of linearity of the fixed part and homoscedastic variance. The expected pattern in this plot when there is no evidence of possible violations of those assumptions is a pattern by which the residuals fluctuate around the zero line in an unstructured way. Indication of violations in the assumptions may suggest one to consider fitting more complex models, modelling nonlinearity directly or taking a heteroscedastic variance structure into account.

Another important graphical procedure for multilevel models is the so called *caterpillar plot*. This plot is constructed as follows: suppose $\hat{u}_j$ and $\hat{\sigma}(\hat{u}_j)$ denote respectively the residuals of a given level and its corresponding standard errors. The caterpillar plot is a side–by–side plot of the separate confidence intervals $\hat{u}_j \pm c\hat{\sigma}(\hat{u}_j)$

versus the values $\hat{u}_j$, where the order of the intervals in the plot is that of the $\hat{u}_j$ in ascending order and endpoints of each interval are connected by a vertical line. The value c is a multiplier factor for the intervals, such as $c = 1.96$. The plot can be used to visualise informally which random effect differs from the overall mean, in the case that the corresponding interval do not include the value zero. Also, an examination whether two random effects appear to be significantly different can be made by checking if two confidence intervals do not overlap.

However, in order for the comparisons in a caterpillar plot to be made as formal hypotheses tests, some search procedure should be used to determine the value of c to control the overall type I error level among all comparisons of interest. For example, Goldstein and Healy (1995) proposed a procedure for pairwise comparisons, where the choice $c = 1.39$ $(= 1.96/\sqrt{2})$, in the case the standard errors where the residuals are about the same, is appropriate to yield an overall significance level of approximately 5%. When the standard errors are not equal, but their ratios of standard errors are at most 2:1, the value of c is impacted slightly and should be increased from 1.39 to 1.4 (Goldstein, 2011, p. 44).

One important assumption in the Goldstein and Healy (1995) procedure, that may not be reasonable in some applications, is that of uncorrelated residuals. If this assumption is violated, it can lead to misleading analyses, as discussed by Afshartous and Wolf (2007). An alternative approach that can be employed in those situations to control the overall significance level of the caterpillar plot is a multiple testing procedure. The well known Bonferroni procedure attains an overall level α by taking $c = z_{\alpha/2K}$, where $z_{\alpha/2K}$ is the standard normal $100(1 - \alpha/2K)$ quantile and K is the number of possible comparisons to be made. One modification of this procedure is the sequential Bonferroni method (Holm, 1979), where one performs separate tests and ranks their p–values from lowest to highest (say,

$p_1 \leq p_2 \leq \cdots \leq p_K$). Then, the null hypothesis associated with the k–th comparison ($k = 1, 2, ..., K$) is rejected if $p_j \leq \alpha/(K - k + 1)$. The Holm method can be very conservative, but it is more powerful than the Bonferroni procedure. However, as pointed out by Afshartous and Wolf (2007, p. 1044), Bonferroni and the Holm methods have the disadvantage of being based on individual p–values and, therefore, their power can be increased by taking dependence in the test statistics into account. Afshartous and Wolf (2007) discuss and extend the procedure of Romano and Wolf (2005), which allows for the incorporation of such type of dependence. However, the implementation of the method is more complex and requires specific computing code or software.

Chapter 2

Investigating the Effects of Interviewers on Nonresponse Bias

2.1 Introduction

Nonresponse has been a key issue in most sample surveys. The major consequence of nonresponse is the lack of reliability of the survey estimates. This happens because, in many cases, characteristics from respondents differ from those of non-respondents, specially if the recruitment process of sample members fails to select units from all distinct groups existing in the target population. Departures of the characteristics from respondents and nonrespondents in a sample may consequently lead to nonresponse bias. Other consequences of nonresponse are smaller sample sizes, which yield incomplete data structure and more complex analyses, as well as the need of specialised software.

Not long ago, the researchers' concern was mostly on the identification of factors that could influence (non)response rates. Merkle and Edelman (2002) investigate the influence of interviewer and voter's characteristics on response rates. In the

survey, the interviewers stand outside of the pooling place waiting for the voters. The authors report that older and middle–aged voters are more likely to respond to older interviewers, whereas younger voters do not seem to mind to respond interviewers from a different age group. Another finding is that interviewer's race does not seem to negatively influence response rates. In addition, the factor that had the bigger influence on response rate was the position of interviewers, those interviewers that were closer to the polling place had higher response rates because the voters apparently understood that the questionnaire was part of the voting process. Although researchers are keen to find ways of reducing nonresponse rates, Groves (2006) and Keeter et al. (2000) stress that (non)response rate alone does not predict nonresponse bias. Nonresponse bias can occur when the survey's main topic is related with the probability of participating in the survey (Bethlehem, 2002). Although much attention is still being given to nonresponse rate issues, Loosveldt and Beullens (2014) report that in recent years the general research focus has shifted towards nonresponse bias in nonresponse research. For example, Groves and Peytcheva (2008) analyse a large number of studies to examine characteristics of the survey design and survey estimates that are related to nonresponse bias. They also examine the relationship between properties of the target population and nonresponse bias.

The assessment of nonresponse bias in sample surveys can be made in a number of ways. Groves (2006) describes five specific approaches, pointing out their advantages and disadvantages. The first approach is based on the comparison of estimates of response rates across sociodemographic variable subgroups. Different response rates across these subgroups yield evidence of nonresponse bias. Although this approach can be easily implemented, it does not estimate the existing nonresponse biases for statistics of interest. A second approach relies on the matching of sampled units with their records from external data sources (e.g.,

rich sampling frames or supplemental matched data). This matching provides variables for both respondents and nonrespondents, and then the respondent and nonrespondent values are compared. Similar values mean no nonresponse bias. This approach allows the nonresponse bias for these variables to be estimated. However, the external data sources may contain measurement error, which can compromise these estimates. Another technique is to compare distributions of sociodemographic variables from respondents with those from the latest census. If these distributions are similar, there is no evidence of nonresponse bias and, as a result, more credibility is given to the survey. Since the key survey variables are usually not the same ones found in the census, nonsampling errors can compromise the comparison. An additional technique consists of comparing subgroups of respondents interviewed in different phases of the data collection (i.e., it compares data from the early survey respondents with data obtained from respondents that needed following up) since these phases may exhibit different nonresponse bias characteristics. This approach is easy to implement, given that the process data are available, and it can be applied to data from many types of surveys. However, it does not take into account the nonrespondents. Lastly, the fifth technique described by Groves is based on contrasting unadjusted respondent–based estimates with post–survey adjusted estimates of the same parameter. The adjustments can be obtained by a combination of, for instance, reweighting of the survey weights, postratification and weighted imputed estimators. This approach has the advantage of allowing the investigation of the impacts of different assumptions in the estimation process, giving the analyst the possibility of drawing safer inferences when the alternative estimates are homogeneous. The problem with this approach is when there are differences among the alternative estimates and no gold standard for comparison is available. In this case, the decision on which estimate to choose from will be unclear since some of the assumptions involved in the estimators may be untestable.

It is widely recognised that the survey interviewer may be a potential factor affecting the survey outcomes and the quality of the data obtained. Interviewers are in charge of contacting and convincing sampled units to take part in the survey and, therefore, play a crucial role in contacting and gaining cooperation from sample survey members (Blom et al., 2011; Durrant et al., 2010; Pickery and Loosveldt, 2002; O'Muircheartaigh and Campanelli, 1999). One line of research on interviewer influences aims to identify interviewer characteristics such as experience, social skills, personality traits and attitudes which could affect survey participation. For example, Jäckle et al. (2013) examine the role of various interviewers' characteristics on achieving cooperation rates, and which of these characteristics are associated with higher cooperation rates, and also investigate the personality traits and inter–personal skills differences between more experienced interviewers and their counterparts. Mainly, they found evidence that interviewers' experience, personality traits and inter–personal skills are associated with cooperation rates. Hox and De Leeuw (2002) attempt to find which interviewer characteristics have influence on (non)response rates. Although they conclude that (non)response rates vary among interviewers, they do not find strong evidence that interviewers attitude and behaviour predict (non)response rates. Snijkers et al. (1999) describe a technique known as concept mapping to extract from experienced interviewers their successful strategies of persuading potential respondents on the doorstep interaction to avoid survey nonresponse. Amongst their strategies, the interviewers found that the most effective are professional competence, tailoring of introduction and maintaining the interaction. Survey interviewers can also affect the response versus nonresponse outcomes not merely in terms of the corresponding rates, but also in terms of a selective composition of the sample of survey respondents that could introduce nonresponse bias (Loosveldt and Beullens, 2014). For example, consider the situation where the sampled units' decision to participate in a survey may be influenced by their interaction with interviewers, so that the interviewers'

request to take part in a survey may not be appealing for some sampled units with specific profiles. Thus, nonresponse bias may result from the influence interviewers might have on the recruitment of respondents and nonrespondents with quite distinct characteristics. If interviewers tend to interview more sampled units from specific groups (such as more people from political party A than from political party B, more older people than younger ones, more employed than unemployed people, among others) rather than from all possible groups, this may lead to nonresponse bias, which may affect negatively the quality of the data (Loosveldt and Beullens, 2014). Another example by which interviewers might have an influence on the recruitment of respondents is when interviewers mention the survey sponsor at the beginning of their interaction with sampled units (Groves et al., 2012). For sampled units that have a positive attitude towards the sponsor, which can be for instance government, scientific organizations or commercial groups, their probability of participation can increase (Groves et al., 2000) and this may lead to an overrepresentation of sample members that support the sponsor (for gratitude, solidarity or civic duty) rather than all types of representations. Although these issues have been discussed more recently in the literature, the effects of interviewers on nonresponse bias have not yet been fully explored.

This study aims to assess interviewer influence on nonresponse bias by applying multilevel modelling techniques. Multilevel models are extensions to standard regression models to deal with clustered structured data. Many methodological developments of these techniques started in the early 1990's (Goldstein, 2011) motivated by the works of Aitkin et al. (1981) and Aitkin and Longford (1986). However, multilevel model analyses incorporated other developments previously established. For example, Snijders and Bosker (2012) point out the issue of aggregate versus individual effects (Robinson, 1950), the distinction between within–group and between–group regression (Davis et al., 1961) and the treatment of regression

intercepts and slopes as outcomes of higher levels (Burstein et al., 1978). Multilevel models are also known as hierarchical linear models (Raudenbush and Bryk, 2002) in reference to the term introduced in the general framework for hierarchical data of Lindley and Smith (1972) and Smith (1973). Comprehensive accounts of multilevel models have been given by Goldstein (1987, 1995, 2003, 2011).

In recent years, multilevel models have been extensively used in the literature to take into account interviewer random effects to investigate the interviewer influence on nonresponse rates. For example, O'Muircheartaigh and Campanelli (1999) apply multilevel models to a binary dependent variable (noncontacts and refusals) and to a multinomial dependent variable (refusals, noncontacts and responses) to investigate the effects of interviewers on survey nonresponse. They also consider a cross–classified multilevel model to disentangle interviewer and area effects using data from an interpenetrated design. Pickery and Loosveldt (2002) also apply multilevel models to a multinomial dependent variable (completed interview, refusal and noncontact) to examine interviewer effects on different types of nonresponse. Additionally, Durrant et al. (2010) consider a cross–classified multilevel model using a binary dependent variable (cooperation and refusal) to identify interviewer characteristics that influence survey cooperation. In the context of the papers aforementioned, interviewers can be seen as clusters since respondents interviewed by the same interviewer tend to give more similar responses than respondents interviewed by different interviewers (Biemer et al., 1991).

In the present study, multilevel models are considered to provide a practical way of assessing the nonresponse bias and the interviewer effects on this bias, when census linked data or auxiliary variables are available. This linkage of survey data with data from other sources containing information on respondents and nonrespondents as a way to assess nonresponse bias is related to one of the approaches described by Groves (2006). Here, the data analysis is carried out by applying

logistic multilevel models in order to predict binary survey variables of interest. One differential feature of these models is that a binary response indicator (response and nonrespose) is used as one of the explanatory variables rather than a dependent variable as in the case of most of the studies in the literature. In addition to a random intercept, a random coefficient is introduced for the response indicator to investigate the effects of interviewers on nonresponse bias. Loosveldt and Beullens (2014) also consider the application of multilevel model with random intercept and random slope to assess nonresponse bias. However, they use the response indicator as the dependent variable and associate a random slope with each explanatory variable. This research makes use of variables from the 2001 UK Labour Force Survey (LFS) dataset, such as interviewer identifying codes, household response outcomes and individual response outcomes, and the 2001 UK census records. The advantage of this dataset is that each household member in the UK LFS has been linked to his/her corresponding 2001 UK individual census records, thus information on both respondents and nonrespondents is available.

In the models considered in this study, in addition to the response indicator, individual–level characteristics are also included to take into account any variation due to assigning interviewers to different types of sample members. The inclusion of these individual–level characteristics is also considered by Loosveldt and Beullens (2014). Even though the effects of interviewer on nonresponse bias are the primary investigation, interviewer effects are potentially confounded with the effects of the areas where the sampled units live, which can make the estimation of interviewer effects more difficult. One example of these confounding effects is when interviewers are assigned to sampled units having similar characteristics or behaviours due to living in the same neighbourhood. Some studies have tried to disentangle interviewer and area effects (Durrant et al., 2010; O'Muircheartaigh and Campanelli, 1999).

The other sections of this chapter are structured as follows. Section 2.2 provides a detailed explanation about the dataset and the variables considered in this study. Section 2.3 defines the statistical models considered for the data analysis. In Section 2.4, the main results are discussed and lastly in Section 2.5 the conclusions from this study are presented.

2.2 Description of the data

The data used in this study comes partly from the 2001 UK Labour Force Survey (LFS) and the other part comes from the 2001 UK census records. The interviewer identifying codes, household response outcomes and individual response outcomes are LFS variables. The information for respondents and nonrespondents regarding the dependent and explanatory variables of the models that will be discussed in Section 2.3 are taken from the Census records.

The 2001 UK LFS is one of the surveys involved in the 2001 UK Census Linked Study. This linked study was designed and performed by the Office for National Statistics (ONS), which also conducts the UK LFS. Including the LFS, response outcomes of six face–to–face major UK Government household surveys, that took place around the time of the 2001 UK Census, have been linked with census records at both the household and the individual–level. Thus, in addition to household and individual characteristics, observations made by the interviewer at the time of the interview about various household and neighbourhood characteristics, interviewer attitudes, area characteristics and survey design features have also been linked (Durrant and Steele, 2009). The Census Linked Study provides a unique dataset containing auxiliary variables for both respondents and nonrespondents. The quality of the linked data was verified by comparing the distributions of key variables before and after the linkage. The survey and census data linkage was

mainly based on households' addresses and the survey records were 95% successfully linked to census records (White et al., 2001).

The LFS collects information about the labour market from a quarterly sample of households throughout the UK. The data collected in this survey include information on household and households' member characteristics, residential details, economic activities, employment status, interviewers and household areas. The interviews that resulted in the data used in this research took place shortly after the 2001 census, between May and June, so that this survey data could be linked to the 2001 UK census variables. Since only one wave of the LFS is considered here, the data are treated as cross–sectional rather than panel. Additionally, this study considers unit nonresponse only.

The analysis sample consists of characteristics from eligible contacted households members who took part in the LFS. Therefore, cases that were deleted from the dataset referred to houses that were vacant second homes, households of size zero, household outcomes which were ineligible (e.g., unable to respond due to language problems), if contact could not be established with anyone in a household (non-contact), individuals without individual–level census data, individuals potentially out of the labour market (i.e., under sixteen or over sixty–four years old) and cases where Interviewer Attitude Survey (IAS) data were not linked (for more details on the IAS, see Freeth et al. (2002)). As a result, the dataset is left with 4,748 cases and 201 interviewers, each with an average workload of approximately 24 interviews.

Amongst the variables in the LFS linked dataset, one that plays an important role in the data analysis is the individual response outcome, which is classified into four categories: 1 = personal response, 2 = proxy response, 3 = nonresponse and 4 = missing. Category 1 means that the response is obtained from the reference person (personally). In category 2, the response is obtained from another person (proxy)

in the household on behalf of the reference person, whereas category 3 implies that neither the reference person responds to the survey nor a proxy responds on his/her behalf. Category 4 is discussed further a bit later in this section. The frequency distribution for this variable is given in Table 2.1, where it can be noticed that there is quite a large number of missing cases (25.7%), which indicates that this variable is poorly recorded.

Table 2.1: Frequency distribution of individual–level response outcome (LFS linked dataset)

Response outcome	Frequency	Percent
1 Personal response	2286	48.1
2 Proxy response	1177	24.8
3 Nonresponse	66	1.4
4 Missing	1219	25.7
Total	4748	100.0

One can obtain a deeper understanding of these missing cases by comparing the response outcomes at individual–level with the response outcomes at household–level to investigate the possibility of inconsistencies in the responses or nonresponses between both levels. The variable that gives the final response outcome at household–level is assigned as follows. Its value is 1 if all members of a household have responded to the survey (full cooperation), it is 2 if some members of a household have responded to the survey, but others have not (partial cooperation) and it is 3 if nobody in the household has responded to the survey (refusal). A cross–tabulation for the individual–level response outcome and the household–level response outcome is presented in Table 2.2.

In the refusal column, in Table 2.2, all 668 cases are classified as missing at individual–level. Since this corresponds to refusal at household–level, it makes more sense to consider this as nonresponse at the individual–level. In the case of full cooperation, it implies that all members of the household have responded

to the survey, then this 471 (in the full cooperation column) missing should be considered as response at individual–level. Since partial cooperation means that not every member of a household has responded to the survey, there are several possibilities for the 80 missing cases at individual–level in the partial cooperation column: (i) all 80 could be response; (ii) the whole amount could be nonresponse; or (iii) only part of them could be response. Since this amount is not much, it is, as a working assumption, considered as response for the analysis. Therefore, all cases from category 4 (missing) of the individual–level response outcome variable, in Table 2.2, are distributed among its first three categories and the new individual–level response outcome variable contains only the categories: 1 = personal response, 2 = proxy response and 3 = nonresponse.

Table 2.2: Frequency distribution of individual–level response outcome and household–level response outcome (LFS linked dataset)

	Household–level response outcome			
Individual–level response outcome	1 Full cooperation	2 Partial cooperation	3 Refusal	Total
1 Person response	2209	77	0	2286
2 Proxy response	1165	12	0	1177
3 Nonresponse	0	66	0	66
4 Missing	471	80	668	1219
Total	3845	235	668	4748

Based on the response outcome at individual–level, a response indicator is defined to be used as one of the explanatory variables in the models. The categories 1 and 2 from the new individual–level response outcome variable is assigned as response (response indicator = 1) and category 3 as nonresponse (response indicator = 0). A cross–tabulation of the response indicator and the household response outcome is shown in Table 2.3.

To assess potential nonresponse bias in the variables from the LFS, it is crucial that variables for respondents and nonrespondents are available. Therefore, instead

of using the LFS target variables, fully observed census variables are used as dependent variables (and as explanatory variables) in the models.

Table 2.3: Frequency distribution of the response indicator and the household response outcome (LFS linked dataset)

	Household response outcome			
Response indicator	1 Full cooperation	2 Partial cooperation	3 Refusal	Total
0 Nonresponse	0	66	668	734
1 Response	3845	169	0	4014
Total	3845	235	668	4748

Census binary variables such as employment and academic qualification are chosen to be dependent variables for the models, since they are observed for all 4748 cases, i.e. for both respondents and nonrespondents and they are directly related to the LFS main investigation. The variable employment is coded 0 if the person is unemployed and 1 if the person is employed, whereas the variable academic qualification is coded as 0 if the person has no academic qualification, O levels GCSEs or other qualification (e.g. City and Guilds, RSA/OCR, BTEC/Edexcel) and as 1 if the person has A levels, first degree (e.g. BA, BSc), higher degree (e.g. MA, PhD, PGCE, post–graduate certificate diplomas) or NVQ levels. Table 2.4 presents the percentages for each dependent variable for nonrespondents and respondents.

Table 2.4: Distributions (percentages) of the dependent variables (LFS linked dataset)

Dependent variable	Nonrespondents	Respondents	Total
Employment	68.94	70.98	70.66
Academic qualification	23.98	30.07	29.13

According to the percentages in Table 2.4, the bias for employment and academic qualification are respectively 0.315 and 0.934 percentage points. As in Groves

(2006), different statistics (in this case, percentages) from the same survey may be subject to different nonresponse biases. Here, the biases are computed based on the definition given by Groves and Couper (1998, p. 3), that is, $Bias(\bar{y}_r) = (m/n)(\bar{y}_r - \bar{y}_m)$, where m/n is the nonresponse rate, $\bar{y}_r$ is the respondent mean and $\bar{y}_m$ is the nonrespondent mean.

Table 2.5 presents the nonrespondent/respondent percentages for each category of the original variable academic qualification. The binary coding used in the models is chosen such that the 1st, 2nd and 5th categories (namely, No academic qualification, O levels GCSEs and Other qualifications, respectively) are assigned to low academic qualification and the other two (A levels, 1st degree, higher degree and NVQ levels) are assigned to high academic qualification. This choice respects the natural ordering of the various qualifications. In any case, if the variable is free from nonresponse bias, regardless of the chosen assignment of its categories, this bias should not be significant. In some cases, the binary coding can actually mask some hidden effects. For instance, Pickery and Loosveldt (2002) argue that they only find a significant relationship between two categories of a three–category variable in their study, because they consider a multinomial dependent variable in their models. This conclusion would not be possible if a dichotomous response model had been used.

Table 2.5: Distributions (percentages) of the orginal codings for academic qualification (LFS linked dataset)

Academic qualification	Nonrespondents	Respondents	Total
No academic qualification	28.88	25.16	25.74
O levels GCSEs	38.69	38.61	38.63
A levels	7.36	9.07	8.80
1st degree, higher degree, NVQ levels	16.62	21.00	20.32
Other qualifications	8.45	6.15	6.51

In addition to the response indicator, further potential explanatory variables that could be included in the models to predict the dependent variables of interest, i.e. variables related to economic status, are individual–level and area characteristics, such as: gender, marital status, student indicator, health, carer, pensioner indicator, dependent child indicator, ethnic group, age and urban/rural indicator. These explanatory variables are used in the models to account for any variation due to different types of sample members assigned to interviewers. Frequency distributions for the explanatory variables are given in Appendix A.

Regarding the structure of the data, it is assumed that individuals interviewed by the same interviewer tend to give more similar responses than individuals interviewed by different interviewers (Biemer et al., 1991). Hence, individual outcomes are considered nested within interviewers. Furthermore, interviewer effects are potentially confounded with area effects. For the 2001 UK LFS linked dataset, interviewers do not seem to be purely nested within local authority districts (areas) where the sampled units live and to which interviewers are assigned. As it is shown in Tables 2.6 and 2.7, there are interviewers assigned to more than one area and some areas are covered by more than one interviewer. Therefore, in the case where area effects are taken into account in the multilevel analyses, individual outcomes are considered to be nested within the cross–classification of both interviewers and areas.

Table 2.6: Frequency distribution of number of interviewers per area (LFS linked dataset)

No. of Interviewers per Area	No. of Areas
1	157
2	108
3	44
4	14
5	9
6	1
Total	333

Table 2.7: Frequency distribution of number of areas per interviewer (LFS linked dataset)

No. of Areas per Interviewer	No. of Interviewers
1	21
2	61
3	56
4	32
5	18
6	10
7	2
9	1
Total	201

2.3 Methodology

Multilevel models (Goldstein, 2011) are used to analyse nonresponse bias and to evaluate the influence of the interviewers on nonresponse bias. The aim is to investigate nonresponse bias in the dependent variables such as employment and academic qualification. Since both dependent variables of interest are binary variables, logistic multilevel models with random interviewer effects are used. Multilevel models are widely used in the literature to analyse the effects of interviewers (Durrant et al., 2010; Scott and Davis, 2001; Hox and De Leeuw, 2002 and O'Muircheartaigh and Campanelli, 1999).

In the nonresponse literature, in order to investigate interviewer effects through multilevel models, the response indicator is generally used as a dependent variable (Durrant et al., 2010). However, in this study, the response indicator is used as an explanatory variable. In this context, the dependent variables are defined as

binary outcomes such that

$$y_{ij} = \begin{cases} 1, & \text{if person } i \text{ interviewed by interviewer } j \text{ has a characteristic of interest} \\ 0, & \text{if person } i \text{ interviewed by interviewer } j \text{ does not have the characteristic,} \end{cases}$$

where $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$. Let r_{ij} be the response indicator of person i, interviewed by interviewer j, which means that $r_{ij} = 1$ if the person i interviewed by interviewer j participates in the survey and $r_{ij} = 0$ otherwise. Let also $\boldsymbol{x}_{ij}$ be a vector of individual–level characteristics (explanatory variables).

Since, in this study, both r_{ij} and y_{ij} are binary variables, the odds ratio of the nonresponse model with y_{ij} as a dependent variable is equivalent to the odds ratio of the nonresponse model with r_{ij} as a dependent variable. The development below shows how the model with y_{ij} as a dependent variable relates to the model with r_{ij} as a dependent variable. For simplicity, it is considered $r_{ij} = r$, $y_{ij} = y$ and $\boldsymbol{x}_{ij} = x$. The model for nonresponse may be written as

$$P(r = 1 \,|\, y, x, j).$$

From Bayes' rule this probability may be expressed as

$$P(r = 1 \,|\, y, x, j) = \frac{P(r = 1 \,|\, x, j) P(y \,|\, r = 1, x, j)}{P(y \,|\, x, j)}. \tag{2.1}$$

Similarly

$$P(r = 0 \,|\, y, x, j) = \frac{P(r = 0 \,|\, x, j) P(y \,|\, r = 0, x, j)}{P(y \,|\, x, j)}. \tag{2.2}$$

Dividing (2.1) by (2.2) gives

$$\frac{P(r=1\,|\,y,x,j)}{P(r=0\,|\,y,x,j)} = \frac{P(r=1\,|\,x,j)P(y\,|\,r=1,x,j)}{P(r=0\,|\,x,j)P(y\,|\,r=0,x,j)},$$

which may be written as

$$\text{odds}(r|y,x,j) = \text{odds}(r|x,j)\frac{P(y\,|\,r=1,x,j)}{P(y\,|\,r=0,x,j)}. \tag{2.3}$$

Therefore, if y is binary

$$\begin{aligned}\frac{\text{odds}(r|y=1,x,j)}{\text{odds}(r|y=0,x,j)} &= \frac{P(y=1\,|\,r=1,x,j)P(y=0\,|\,r=0,x,j)}{P(y=1\,|\,r=0,x,j)P(y=0\,|\,r=1,x,j)} \\ &= \frac{\text{odds}(y\,|\,r=1,x,j)}{\text{odds}(y\,|\,r=0,x,j)}.\end{aligned} \tag{2.4}$$

Nonresponse is missing at random (MAR) if $\text{odds}(r\,|\,y,x,j) = \text{odds}(r\,|\,x,j)$. Hence, nonresponse is MAR if $\text{odds}(y\,|\,r=1,x,j) = \text{odds}(y\,|\,r=0,x,j)$. The extent of not missing at random (NMAR) may be measured by the odds ratio

$$\frac{\text{odds}(r\,|\,y=1,x,j)}{\text{odds}(r\,|\,y=0,x,j)},$$

which from (2.4) is the same as the odds ratio

$$\frac{\text{odds}(y\,|\,r=1,x,j)}{\text{odds}(y\,|\,r=0,x,j)}.$$

Considering the full notation for the expressions, the first multilevel model of interest is the two–level random intercept logistic model, which may be written as

$$
\begin{aligned}
y_{ij} \mid r_{ij}, \boldsymbol{x}_{ij}, \boldsymbol{u}_0 \ &\sim \text{ indep Bernoulli}(\pi_{ij}), \quad i = 1, \ldots, n_j, j = 1, \ldots, J \ \text{ and} \\
&\qquad\qquad\qquad\qquad\qquad\qquad \boldsymbol{u}_0 = (u_{01}, ..., u_{0J})^\top, \\
\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) &= \beta_{0j} + \beta_1 r_{ij} + \boldsymbol{\beta}^\top \boldsymbol{x}_{ij}, \\
\beta_{0j} &= \beta_0 + u_{0j},
\end{aligned}
\tag{2.5}
$$

where $\pi_{ij} = P(y_{ij} = 1 \,|\, r_{ij}, \boldsymbol{x}_{ij}, j)$ is the probability of the dependent variable having the characteristic of interest, β_{0j} is the interviewer–dependent intercept, β_1 is the coefficient of the response indicator, $\boldsymbol{\beta}$ is a vector of coefficients and u_{0j} are random effects representing unexplained interviewer effects. The random effects are assumed to be independent and identically distributed random variables from a normal distribution, i.e., $u_{0j} \sim$ i.i.d. $N\left(0, \sigma^2_{u_0}\right)$. The variance parameter $\sigma^2_{u_0}$ is the residual between–interviewer variance in the log–odds of the dependent variable. Substituting the expression of β_{0j} into the model for the log–odds of π_{ij} in (2.5), this model becomes

$$
\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_1 r_{ij} + \boldsymbol{\beta}^\top \boldsymbol{x}_{ij} + u_{0j}.
$$

This model assumes that the bias, if present, is the same for all interviewers, since β_1 does not depend on interviewers. In the context of the models investigated in this study, β_{0j} rather than be the interviewer–dependent intercept, it is an interviewer assignment effect since the dependent variables used are census variables, which no interviewer is involved in their collection since census questionnaires are self–administered. This is a novel interpretation of an interviewer–level effect since typically the interviewers have collected the data. The coefficient of the response indicator, β_1, has a particular meaning with regards to nonresponse bias. Using the same reasoning as in Agresti (2013, p. 183), one can note that, controlling for the other explanatory variables and for the interviewer random effects, β_1 is the difference of logits of the dependent variable between respondents ($r_{ij} = 1$) and

nonrespondents ($r_{ij} = 0$), since

$$[\beta_{0j} + \beta_1(1) + \boldsymbol{\beta}^\top \boldsymbol{x}_{ij}] - [\beta_{0j} + \beta_1(0) + \boldsymbol{\beta}^\top \boldsymbol{x}_{ij}] = \beta_1. \tag{2.6}$$

In this case, it follows that the $\text{odds}(y \mid r = 1, x, j) = \exp(\beta_1) \times \text{odds}(y \mid r = 0, x, j)$. This means that if β_1 is significantly different from zero, the odds of response differs from the odds of nonresponse, which indicates that there is nonresponse bias for all interviewers.

To relax the assumption of the same bias for all interviewers, another possible model of interest is the two–level random coefficient logistic model

$$\begin{aligned} y_{ij} \mid r_{ij}, \boldsymbol{x}_{ij}, \boldsymbol{u}_1, ..., \boldsymbol{u}_J \;\sim\; & \text{indep Bernoulli}(\pi_{ij}), \quad i = 1, \ldots, n_j \;\text{ and} \\ & \qquad\qquad\qquad\qquad\qquad\qquad j = 1, \ldots, J, \\ \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) &= \beta_{0j} + \beta_{1j} r_{ij} + \boldsymbol{\beta}^\top \boldsymbol{x}_{ij}, \\ \beta_{0j} &= \beta_0 + u_{0j}, \\ \beta_{1j} &= \beta_1 + u_{1j}, \end{aligned} \tag{2.7}$$

where β_{0j} is the same as in (2.5), β_{1j} is the interviewer–dependent coefficient of the response indicator and $\boldsymbol{u}_j = (u_{0j}, u_{1j})^\top$ is a vector of random effects representing unexplained interviewer effects. The random effects are assumed to follow a bivariate normal distribution, i.e., $\boldsymbol{u}_j \sim N(\boldsymbol{0}, \boldsymbol{\Omega}_u)$ and

$$\boldsymbol{\Omega}_u = \begin{bmatrix} \sigma^2_{u_0} & \sigma_{u_0 u_1} \\ \sigma_{u_0 u_1} & \sigma^2_{u_1} \end{bmatrix},$$

where the parameters $\sigma^2_{u_0}$ and $\sigma^2_{u_1}$ are respectively the random intercept variance and the random coefficient variance. The term $\sigma_{u_0 u_1}$ is the covariance between random intercept and random coefficient effects. Substituting the expressions of

β_{0j} and β_{1j} into the logit model of (2.7), it becomes

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_1 r_{ij} + u_{1j} r_{ij} + \boldsymbol{\beta}^\top \boldsymbol{x}_{ij} + u_{0j}.$$

This model relaxes the assumption of constant nonresponse bias for interviewers in Equation (2.5), allowing the bias, if present, to vary across interviewers. Similar to above, the random coefficient has a meaning with regards to nonresponse bias. Whereas β_{0j} could be only attributed to interviewer assignment effects, it is possible to attribute β_{1j} to the effect of interviewers (as well as possibly in combination with the effect of interviewer assignment) since r_{ij} does relate to the LFS interviewers. The distinction between interviewer effects and interviewer assignment effects could be better explained as follows. Groves and Magilavy (1986) define interviewer effects as the variation among the responses obtained from sampled units if different interviewers were assigned to respondents. In the LFS, the response indicator differentiates the responses from nonresponses (refusals), thus this response indicator may be related to the interviewers since they are in charge of convincing sampled units to take part in the survey. Therefore, a significant variation at interviewer–level on the coefficient of this response indicator represents the difference of logits of the dependent variable between respondents and nonrespondents for a specific interviewer, that is, an interviewer effect. On the other hand, interviewer assignment effects are defined as the significant variation at interviewer–level on variables for which interviewers are not related to the collection of these variables. In this chapter, the dependent variables considered in the models, employment and academic qualification, for both respondents and nonrespondents are self–administered Census variables. This interviewer assignment effect is a novel interpretation of an interviewer–level effect since typically the interviewers collect the data.

The random coefficient represents the difference of logits of the dependent variable between respondents and nonrespondents for interviewer j, since

$$\begin{aligned}&[\beta_0 + \beta_1(1) + u_{1j}(1) + \boldsymbol{\beta}^\top \boldsymbol{x}_{ij} + u_{0j}] - \\ &[\beta_0 + \beta_1(0) + u_{1j}(0) + \boldsymbol{\beta}^\top \boldsymbol{x}_{ij} + u_{0j}] = \\ &\beta_1 + u_{1j} = \beta_{1j}.\end{aligned} \tag{2.8}$$

Thus, the odds$(y \,|\, r = 1, x, j) = \exp(\beta_{1j}) \times$ odds$(y \,|\, r = 0, x, j)$, implying that if β_{1j} is significantly different from zero there is evidence of nonresponse bias for interviewer j.

The above models may be extended to multilevel cross–classified logistic models (Goldstein, 2011), since interviewer effects are potentially confounded with area effects. In this case, individual outcomes are nested within a cross–classification of interviewers and areas. The multilevel cross–classified logistic model may be written similarly to model (2.7) using the systematic component

$$\log\left(\frac{\pi_{i(jk)}}{1-\pi_{i(jk)}}\right) = \beta_0 + \beta_1 r_{i(jk)} + u_{1j} r_{i(jk)} + \boldsymbol{\beta}^\top \boldsymbol{x}_{i(jk)} + u_{0j} + v_{0k}, \tag{2.9}$$

where the notation $i(jk)$ means that individual i is nested within a cross–classification of interviewer j and area k. The other terms of the model have already been explained for the models in Equations (2.5) and (2.7) except for v_{0k}, which are i.i.d. $N(0, \sigma^2_{v_0})$, denoting a random effect representing unexplained area effects.

Again, the random coefficient in (2.9) represents the difference of logits of the dependent variable between respondents and nonrespondents for each interviewer

j in the same area k, since

$$\begin{aligned}
&[\beta_0 + \beta_1(1) + u_{1j}(1) + \boldsymbol{\beta}^\top \boldsymbol{x}_{i(jk)} + u_{0j} + v_{0k}] - \\
&[\beta_0 + \beta_1(0) + u_{1j}(0) + \boldsymbol{\beta}^\top \boldsymbol{x}_{i(jk)} + u_{0j} + v_{0k}] = \\
&\beta_1 + u_{1j} = \beta_{1j}.
\end{aligned} \tag{2.10}$$

This has the same interpretation as in (2.8).

A novel reparametrization of the model in Equation (2.9) is considered to analyse simultaneously the residual between–interviewer variance components for the respondents and nonrespondents. This also provides a more intuitive model for assessing the bias. Equation (2.11) presents this reparametrization.

$$\begin{aligned}
\log\left(\frac{\pi_{i(jk)}}{1-\pi_{i(jk)}}\right) &= \beta_{0k} + \beta_{1j} r_{i(jk)} + \beta_{2j}(1 - r_{i(jk)}) + \boldsymbol{\beta}^\top \boldsymbol{x}_{i(jk)} \\
\beta_{0k} &= \beta_0 + v_{0k} \\
\beta_{1j} &= \beta_1 + u'_{1j} \\
\beta_{2j} &= u'_{0j},
\end{aligned} \tag{2.11}$$

where $\left(1 - r_{i(jk)}\right)$ is the nonresponse indicator, β_{2j} is the interviewer–dependent coefficient of the nonresponse indicator. If $r_{i(jk)} = 1$, u'_{1j} is the sum of the random effects u_{0j} and u_{1j} from Equation (2.9), whereas if $r_{i(jk)} = 0$, u'_{0j} is the same as u_{0j} from Equation (2.9). Also, $\boldsymbol{u}'_j = (u'_{0j}, u'_{1j})^\top$ is a vector of random effects representing unexplained interviewer effects following a bivariate normal distribution, i.e.

$$\boldsymbol{u}'_j = \begin{bmatrix} u'_{0j} \\ u'_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} ; \begin{bmatrix} \sigma^2_{u'_0} & \sigma_{u'_0 u'_1} \\ \sigma_{u'_0 u'_1} & \sigma^2_{u'_1} \end{bmatrix} \right),$$

where $\sigma^2_{u'_0} = \sigma^2_{u_0}$ and $\sigma^2_{u'_1} = \sigma^2_{u_0} + \sigma^2_{u_1} + 2\sigma_{u_0u_1}$. By substituting the expressions for β_{0k}, β_{1j} and β_{2j}, (2.11) may be rewritten as

$$\log\left(\frac{\pi_{i(jk)}}{1-\pi_{i(jk)}}\right) = \beta_0 + \beta_1 r_{i(jk)} + u'_{1j} r_{i(jk)} + u'_{0j}(1 - r_{i(jk)}) + \boldsymbol{\beta}^\top \boldsymbol{x}_{i(jk)} + v_{0k}. \quad (2.12)$$

Using the same reasoning as before, the difference between the random coefficient and the interviewer random effect represents the difference of logits of the dependent variable between respondents and nonrespondents for interviewer j in the same area k, since

$$\begin{aligned}
&\left[\beta_0 + \beta_1(1) + u'_{1j}(1) + u'_{0j}(1-1) + \boldsymbol{\beta}^\top \boldsymbol{x}_{i(jk)} + v_{0k}\right] - \\
&\left[\beta_0 + \beta_1(0) + u'_{1j}(0) + u'_{0j}(1-0) + \boldsymbol{\beta}^\top \boldsymbol{x}_{i(jk)} + v_{0k}\right] = \\
&\beta_1 + u'_{1j} - u'_{0j} = \beta_{1j} - u'_{0j}.
\end{aligned}$$

This expression implies that the $\text{odds}(y \mid r = 1, x, j) = \exp(\beta_{1j} - u'_{0j}) \times \text{odds}(y \mid r = 0, x, j)$. Thus, if the difference between β_{1j} and u'_{0j} is significantly different from zero, there is nonresponse bias for interviewer j.

2.3.1 Modelling strategy

The modelling strategy is as follows: first, single–level logistic models with only the response indicator as the explanatory variable are fitted to dependent binary variables of interest from the 2001 UK LFS linked dataset. Thereafter, since the nonresponse literature has identified a number of characteristics that contribute to nonresponse, such as no children in the household, single household, male, younger age groups, among others (Durrant and Steele, 2009; Durrant et al., 2010 and Campanelli et al., 1997), individual–level characteristics are included into the models as additional explanatory variables to control for differences in sample members assigned to interviewers. Then, the intercept is allowed to vary across interviewers

(two–level random intercept logistic model) to explore if the dependent variables of interest differ across interviewers, reflecting interviewer assignment effects since the dependent variables are self–completed census variables. At this point, if the response indicator is significant, it means that there is nonresponse bias in the dependent variable. However, the bias is assumed to be the same across all interviewers. After that, the difference between response and nonresponse within an interviewer is allowed to vary across interviewers (two–level random coefficient logistic model). If this random coefficient is significant, it means that there is difference between the characteristics from respondents and nonrespondents for an interviewer, i.e. the nonresponse bias depends on interviewers. Lastly, the models include a cross–classification of interviewers and areas where the sampled units live aiming to disentangle these two effects that are potentially confounded.

2.3.2 Estimation methods

To estimate random intercept and random coefficient logistic multilevel models, 2nd order predictive quasi–likelihood (PQL) (Green, 1987; Breslow and Clayton, 1993) is used, since for multilevel models with binary responses Goldstein and Rasbash (1996) demonstrate that the second order PQL leads to much better estimates than the 1st order marginal quasi–likelihood (MQL) (Goldstein, 1991). PQL, however, can still be biased since it tends to underestimate the variance components and fixed effects (Breslow and Clayton, 1993).

The multilevel logistic cross–classified models are estimated using Markov Chain Monte Carlo (MCMC) with a diffuse prior and with starting values obtained by using 2nd order PQL. The MCMC method also considers 150,000 iterations after a burn–in of 10,000. The models are fitted using `MLwiN` (Rasbash et al., 2012). MCMC was chosen because in addition to yield more accurate estimates compared

to PQL (Goldstein, 2011), the software used here only permits the estimation of cross–classified models by MCMC methods.

The trajectory plots suggest that the length of 10,000 was enough for the burn–in chains. For the majority of the models, 150,000 iterations are more than enough to estimate the variance terms since the Monte Carlo Standard Error, which is a measure of inaccuracy of Monte Carlo samples, never went more than 0.005 (and only approximately 28% of models had values for the Monte Carlo Standard Error between 0.004 and 0.005) and the chains mix well. The estimates for the covariance terms were less accurate though since the chains are less well behaved.

2.4 Results

Firstly, in this section, the response rate for the 2001 UK LFS is explored across interviewers using descriptive statistics. Then, the multilevel analysis for the dependent variables of interest is carried out.

Figure 2.1 shows the distribution of response rates achieved by the interviewers. These response rates were computed as the percentage of successful interviews out of the total number of interviews of an interviewer. According to the histogram, the response rates vary considerably across interviewers. This is also found in prior research in this area (see for example West et al. (2013)). Additionally, this response rates are skewed to the left, meaning that most interviewers have high response rates. The average workload for all interviewers is approximately 24 interviews with a standard deviation of 13.48, whereas the average workload for interviewers who have 100% response rates is 13 interviews with a standard deviation of 9.33.

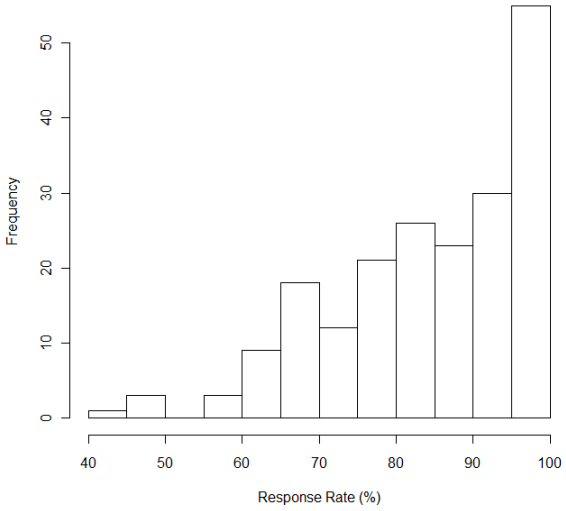


Figure 2.1: Histogram of the interviewers' response rates (LFS linked dataset)

The multilevel models are fitted considering two binary dependent variables: employment and academic qualification. As mentioned earlier, these dependent variables from the census are chosen because they collect similar information as the LFS target variables.

2.4.1 Modelling employment

The models presented in Table 2.8 for the log–odds of being employed are the standard (single–level) logistic model with only the response indicator as the explanatory variable (Model E1), the single–level logistic model with additional explanatory variables (Model E2), the two–level random intercept logistic model (Model E3) and the two–level random coefficient logistic model (Model E4). The intercept in Model E1 is the log–odds of being employed given that the person did not respond to the survey. This intercept can be computed from the percentages in Table 2.4, by taking the logarithm of the ratio between the nonresponse proportion for those employed (0.6894) and the nonresponse proportion for those unemployed

(1 – 0.6894), that is 0.797. The sum of the coefficients, in Model E1, can also be worked out from the percentages in Table 2.4, by taking the logarithm of the ratio between the response proportion for those employed (0.7098) and the response proportion for those unemployed (1 – 0.7098), that is 0.894 (= 0.797+0.097, from Model E1). In Model E1, the response indicator is not significant at the 5% level to explain the odds (probability) of being employed.

Table 2.8: Parameter estimates (with standard errors in brackets) for the models for employment

Fixed effect	Model E1	Model E2	Model E3	Model E4
Constant	0.797 (0.080) **	-0.242 (0.145) *	-0.255 (0.152) *	-0.246 (0.154)
Response indicator				
Response	0.097 (0.087)	0.148 (0.091)	0.161 (0.093) *	0.148 (0.098)
Sex				
Female		-0.676 (0.068) **	-0.685 (0.069) **	-0.686 (0.069) **
Marital status				
First marriage		-0.139 (0.084) *	-0.156 (0.086) *	-0.156 (0.086) *
Re–married		-0.233 (0.130) *	-0.255 (0.133) *	-0.257 (0.134) *
Sep & legally married		-0.506 (0.204) **	-0.534 (0.208) **	-0.533 (0.209) **
Divorced		-0.482 (0.127) **	-0.511 (0.130) **	-0.512 (0.130) **
Widowed		-0.825 (0.233) **	-0.869 (0.236) **	-0.871 (0.237) **
Student indicator				
Not full–time		1.701 (0.133) **	1.746 (0.136) **	1.752 (0.137) **
Ethnic group				
Mixed group		0.370 (0.453)	0.334 (0.469)	0.334 (0.470)
Asian group		-0.646 (0.191) **	-0.735 (0.200) **	-0.748 (0.201) **
Black group		-0.673 (0.262) **	-0.750 (0.275) **	-0.754 (0.276) **
Other group		-0.621 (0.271) **	-0.622 (0.284) **	-0.615 (0.284) **
Random effect				
Random intercept:				
Interviewer variance $\sigma^2_{u_0}$			0.143 (0.037)	0.206 (0.135)
Random coefficient:				
Interviewer coef. variance $\sigma^2_{u_1}$				0.103 (0.151)
Interv. inter.-coef. covariance $\sigma_{u_0u_1}$				-0.081 (0.131)
Interv. inter.-coef. correlation $\rho_{u_0u_1}$				-0.553

The base categories for the explanatory variables are Nonresponse, Male, Single, Full-time student and White group. Model E1 is the standard (single-level) logistic model with only the response indicator as the explanatory variable, Model E2 is the single-level logistic model with additional individual-level characteristics (explanatory variables), Model E3 is the two-level logistic random intercept model and Model E4 is the two-level logistic random coefficient model.
** Significant at the 5% level
* Significant at the 10% level

Controlling for other explanatory variables (individual–level characteristics), in Model E2, the response indicator is still not significant. However, after allowing

the intercept to vary across interviewers (Model E3), the coefficient of the response indicator is significant at the 10% level. According to the interpretation of the response indicator coefficient from Equation (2.6), this means that, at the 10% level, the nonresponse bias is significant for all interviewers. Also, the Wald test for the inclusion of the interviewer variance provides a value of 14.893 on 1 degree of freedom, which is significant (p–value < 0.001), meaning that there is evidence that the probability of being employed varies across interviewers, controlling for the explanatory variables.

Allowing the difference between respondents and nonrespondents within an interviewer to vary across interviewers (Model E4), the coefficient of the response indicator is not significant, which indicates that, on average, there is no nonresponse bias. Moreover, in Model E4, the interviewer random effects variance for the response indicator coefficient is not significant.

2.4.2 Modelling academic qualification

Table 2.9 presents the parameter estimates for the models using academic qualification as the binary dependent variable. The models considered are the same ones as in Table 2.8. The intercept, in Model A1, is the log–odds of being highly academically qualified given that the person did not respond to the survey. As mentioned before, this intercept can be computed from the percentages in Table 2.4, by taking the logarithm of the ratio between the nonresponse proportion for those with high academic qualification (0.2398) and the nonresponse proportion for those with low academic qualification $(1 - 0.2398)$, that is -1.154. In the same way as before, the sum of the coefficients, in Model A1, can be also worked out from the percentages in Table 2.4, by taking the logarithm of the ratio between the response proportion for those with high academic qualification (0.3007) and

the response proportion for those with low academic qualification $(1 - 0.3007)$, that is -0.844 $(= -1.154 + 0.310)$. In Model A1, the response indicator is significant at the 5% level to explain the odds (probability) of being highly academically qualified.

Table 2.9: Parameter estimates (with standard errors in brackets) for the models for academic qualification

Fixed effect	Model A1	Model A2	Model A3	Model A4
Constant	-1.154 (0.086) **	-0.547 (0.102) **	-0.563 (0.116) **	-0.638 (0.129) **
Response indicator				
Response	0.310 (0.093) **	0.303 (0.096) **	0.284 (0.101) **	0.358 (0.123) **
Marital status				
First marriage		-0.188 (0.087) **	-0.141 (0.092)	-0.139 (0.093)
Re–married		-0.780 (0.155) **	-0.736 (0.162) **	-0.749 (0.164) **
Sep & legally married		-0.522 (0.226) **	-0.476 (0.234) **	-0.486 (0.237) **
Divorced		-0.272 (0.140) *	-0.318 (0.145) **	-0.313 (0.147) **
Widowed		-0.522 (0.309) *	-0.448 (0.321)	-0.439 (0.323)
Health				
Fairly good		-0.455 (0.088) **	-0.480 (0.092) **	-0.493 (0.093) **
Not good		-0.640 (0.143) **	-0.695 (0.150) **	-0.713 (0.152) **
Dependent child				
Yes		-2.586 (0.331) **	-2.641 (0.342) **	-2.676 (0.345) **
Age				
35 – 49		-0.200 (0.088) **	-0.195 (0.092) **	-0.199 (0.093) **
50 – 64		-0.726 (0.104) **	-0.775 (0.109) **	-0.783 (0.110) **
Urban/rural indicator				
Rural		0.356 (0.104) **	0.296 (0.113) **	0.319 (0.114) **
Random effect				
Random intercept:				
Interviewer variance $\sigma^2_{u_0}$			0.328 (0.058)	0.598 (0.218)
Random coefficient:				
Interviewer coef. variance $\sigma^2_{u_1}$				0.587 (0.247)
Interv. inter.-coef. covariance $\sigma_{u_0u_1}$				-0.407 (0.213)
Interv. inter.-coef. correlation $\rho_{u_0u_1}$				-0.687

The base categories for the explanatory variables are Nonresponse, Single, Good, Not a dependent child, 16 – 34 and Urban.
Model A1 is the standard (single-level) logistic model with only the response indicator as the explanatory variable, Model A2 is the single-level logistic model with additional individual-level characteristics (explanatory variables), Model A3 is the two-level logistic random intercept model and Model A4 is the two-level logistic random coefficient model.
** Significant at the 5% level
* Significant at the 10% level

Including other explanatory variables in the model (Model A2), the coefficient of the response indicator is significant, meaning that people who took part in the survey are more likely to be highly academically qualified, controlling for the

variables marital status, health, dependent child indicator, age and urban/rural indicator.

Model A3, in Table 2.9, extended Model A2 to allow the probability of being highly academically qualified to vary randomly across interviewers. In this case, the interviewer effects are actually interviewer assignment effects, since academic qualification is a variable from the census, for which there is no interviewers involved in its collection since census questionnaires are self–administered. The coefficient of the response indicator is significant at the 5% level. As discussed in Section 2.3, this indicates that there is nonresponse bias for all interviewers, even though the bias is the same for all of them. The addition of the variance of the interviewer (assignment) random effects is highly significant (p–value < 0.001) based on the Wald test (test statistic=31.621 on 1 degree of freedom), which means that there are significant differences in the probability of being highly academically qualified across interviewers, controlling for the explanatory variables. This interviewer (assignment) variance may be significant due to failure of equally allocating highly academically qualified people to interviewers as shown in Figure 2.2.

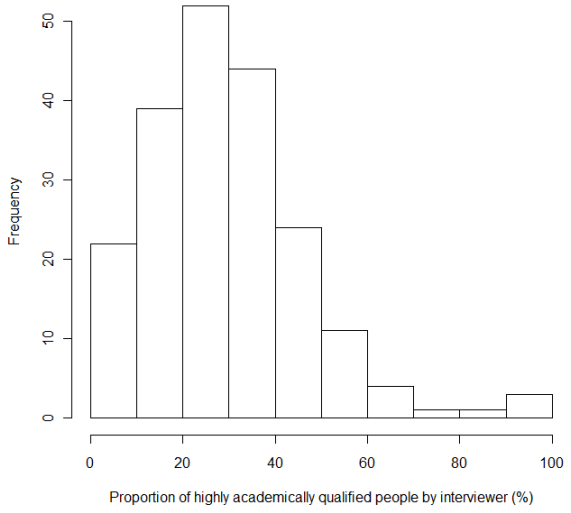


Figure 2.2: Histogram of the proportion of highly academically qualified people by interviewer (LFS linked dataset)

In order to investigate whether or not the nonresponse bias, if present, depends on interviewers, the coefficient of the response indicator in Model A3 is allowed to vary across interviewers. Model A4, in Table 2.9, is the two–level random coefficient logistic model.

In Model A4, the coefficient of the response indicator is significant at the 5% level. From the interpretation that follows Equation (2.8), this means that there is nonresponse bias and it depends on interviewers. Also, this means that the odds of being highly academically qualified for respondents differ from the odds of being highly academically qualified for nonrespondents by interviewer. Furthermore, the Wald test statistic for the inclusion of the new parameters (variance of u_{1j} and covariance between u_{0j} and u_{1j}) is 6.038 (on 2 degrees of freedom), which means that they are significant (p–value=0.049). Therefore, there is evidence that also the effects of response/nonresponse vary by interviewer for the probability of being highly academically qualified, controlling for the other explanatory variables.

On average, after adjusting for the effects of the other explanatory variables, the log–odds of being highly academically qualified is 0.358 higher for respondents than for nonrespondents. The interviewer–level variance for the nonrespondents is 0.598 whereas for the respondents is 0.317 (= 0.598 + 2(-0.407) + 0.587). Hence, the interviewer–level variation is larger between the nonrespondents.

In order to have a better understanding of the random effects pattern in Model A4, a scatterplot for the random effects of the intercept (u_{0j}) versus the random effects of the response indicator coefficient (u_{1j}) is presented in Figure 2.3. The plot shows a negative association between the predicted random effects for the intercept and for the coefficient. The correlation between the two predicted effects is -0.687.

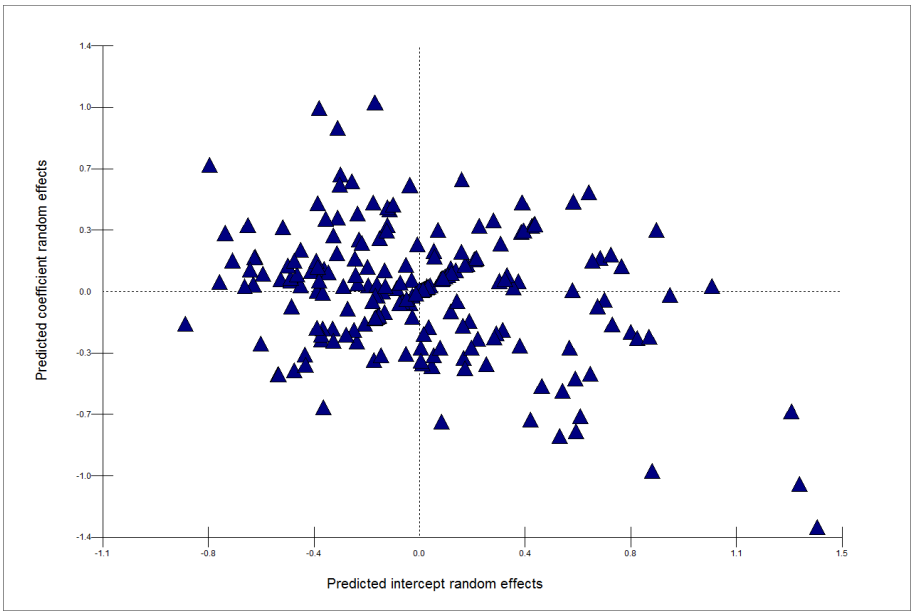


Figure 2.3: Scatterplot of the predicted u_{0j} versus u_{1j} predicted for academic qualification

Here, the random intercept reflects differences in interviewer assignment effects, whereas the random coefficient reflects interviewer effects on nonresponse bias. For interviewer assignments with positive random effects, the random effects of the nonresponse bias tend to be negative. This means that interviewers assigned to highly academically qualified people tend to have a lower than average bias, whereas interviewers assigned to lower academically qualified people tend to have a higher than average bias. The role of these random effects must be investigated further in order to be well understood.

2.4.3 Cross–classified models

Despite the analysis for the two dependent variables of interest provides an assessment of the nonresponse bias, it is not clear yet why the interviewer random intercept effects variation is significant, since the dependent variables in the models are census variables, for which there are no interviewers involved in their collection. One possible explanation could be that interviewer effects may be confounded with

the effects of the areas where the sampled units live. Even though the LFS design is unclustered, interviewers are still assigned to specific geographic areas, usually neighbouring areas. Thus, the effects of interviewers and areas can still be confounded. In order to check this possibility, the models are extended to include area random effects.

Tables 2.6 and 2.7 show that in the 2001 UK LFS some areas are covered by more than one interviewer and some interviewers worked in more than one area, which means that interviewers and areas are not strictly nested, but they possibly have a cross–classified structure. In this case, individuals are nested within the cross–classification of both interviewers and areas. To analyse simultaneously the effects of interviewers and the effects of areas on the dependent variables of interest, cross–classified models are applied. Since the sample experiment design is not an interpenetrated one (O'Muircheartaigh and Campanelli, 1999), the idea is not to claim that interviewer and area effects will be entirely disentangled by applying these models, but to investigate this possibility.

Tables 2.10 and 2.11 present the cross–classified logistic models for employment and academic qualification respectively. In Table 2.10, although the response indicator coefficient is not significant, the interviewer–level variance is significant when the intercept is allowed only for random interviewer effects (Model E5). Once the model (Model E6) includes the cross–classification of interviewer and area random effects, the area variance is significant, whereas the interviewer variance is not significant anymore. The same pattern is still holding after allowing the coefficient of the response indicator to vary across interviewers (Model E7), and even after controlling for other individual–level characteristics (Model E8).

Table 2.10: Parameter estimates (with standard errors in brackets) for the models for employment (cross–classified models)

Fixed effect	Model E5	Model E6	Model E7	Model E8
Constant	0.804 (0.087) **	0.855 (0.093) **	0.861 (0.096) **	-0.263 (0.162)
Response indicator				
Response	0.106 (0.090)	0.103 (0.094)	0.096 (0.099)	0.147 (0.106)
Sex				
Female				-0.718 (0.071) **
Marital status				
First marriage				-0.166 (0.089) *
Re–married				-0.269 (0.138) *
Sep & legally married				-0.505 (0.218) **
Divorced				-0.516 (0.134) **
Widowed				-0.859 (0.249) **
Student indicator				
Not full–time				1.858 (0.144) **
Ethnic group				
Mixed group				0.360 (0.495)
Asian group				-0.852 (0.213) **
Black group				-0.779 (0.290) **
Other group				-0.585 (0.298) **
Random effect				
Random intercept:				
Interviewer variance $\sigma^2_{u_0}$	0.137 (0.039)	0.029 (0.027)	0.122 (0.076)	0.178 (0.114)
Area variance $\sigma^2_{v_0}$		0.247 (0.057)	0.230 (0.056)	0.297 (0.066)
Random coefficient:				
Interviewer coef. variance $\sigma^2_{u_1}$			0.148 (0.105)	0.210 (0.140)
Interv. inter.-coef. covariance $\sigma_{u_0u_1}$			-0.093 (0.082)	-0.162 (0.120)
Interv. inter.-coef. correlation $\rho_{u_0u_1}$				-0.840
BDIC	5696.60	5638.87	5637.75	5293.84

The base categories for the explanatory variables are Nonresponse, Male, Single, Full-time student and White group.
Model E5 is the two-level random intercept model with only the response indicator as the explanatory variable, Model E6 is the interviewer-area cross-classified model, Model E7 is the interviewer-area cross-classified model with random coefficient and Model E8 is Model E7 plus additional individual-level characteristics (explanatory variables). BDIC is the Bayesian Deviance Information Criterion.
** Significant at the 5% level
* Significant at the 10% level

Inspecting the random effects for the models in Table 2.11, one can see that when the intercept is allowed only for interviewer random effects (Model A5), the interviewer–level variance is significant. As the model includes the cross–classification of both interviewer and area random effects (Model A6), the interviewer–level variance decreases. However, both the area and interviewer variances are still significant. Allowing the response indicator coefficient to vary across interviewers (Model A7), the interviewer–level variance becomes less significant than it used to be in Model A6, while the significance of the area–level variance is still

about the same. As other individual–level characteristics are included in the model (Model A8), the interviewer–level variance becomes more inflated. Although this seems surprising, Snijders and Bosker (2012, p. 309) explain that for a multilevel model considering binary dependent variables, adding highly significant level–one (individual–level) explanatory variables in the model tend to increase estimated level–two (interviewer–level) variances.

Table 2.11: Parameter estimates (with standard errors in brackets) for the models for academic qualification (cross–classified models)

Fixed effect	Model A5	Model A6	Model A7	Model A8
Constant	-1.177 (0.101) **	-1.220 (0.108) **	-1.295 (0.125) **	-0.686 (0.144) **
Response indicator				
Response	0.288 (0.098) **	0.291 (0.102) **	0.359 (0.125) **	0.363 (0.136) **
Marital status				
First marriage				-0.130 (0.098)
Re–married				-0.752 (0.168) **
Sep & legally married				-0.439 (0.247) *
Divorced				-0.303 (0.153) **
Widowed				-0.426 (0.333)
Health				
Fairly good				-0.507 (0.095) **
Not good				-0.703 (0.156) **
Dependent child				
Yes				-2.787 (0.351) **
Age				
35 – 49				-0.211 (0.096) **
50 – 64				-0.816 (0.115) **
Urban/rural indicator				
Rural				0.357 (0.124) **
Random effect				
Random intercept:				
Interviewer variance $\sigma^2_{u_0}$	0.337 (0.067)	0.237 (0.067)	0.566 (0.246)	0.623 (0.270)
Area variance $\sigma^2_{v_0}$		0.341 (0.075)	0.347 (0.078)	0.336 (0.081)
Random coefficient:				
Interviewer coef. variance $\sigma^2_{u_1}$			0.532 (0.241)	0.653 (0.267)
Interv. inter.-coef. covariance $\sigma_{u_0u_1}$			-0.412 (0.224)	-0.492 (0.248)
Interv. inter.-coef. correlation $\rho_{u_0u_1}$				-0.771
BDIC	5563.82	5479.00	5466.26	5198.86

The base categories for the explanatory variables are Nonresponse, Single, Good, Not a dependent child, 16 – 34 and Urban. Model A5 is the two-level random intercept model with only the response indicator as the explanatory variable, Model A6 is the interviewer-area cross-classified model, Model A7 is the interviewer-area cross-classified model with random coefficient and Model A8 is Model A7 plus additional individual-level characteristics (explanatory variables). BDIC is the Bayesian Deviance Information Criterion.

** Significant at the 5% level

* Significant at the 10% level

Later on, in this section, a residual analysis will give further details about the final model for academic qualification.

2.4.4 Cross–classified models after reparametrization

The significant negative estimated value -0.69 of the correlation ρ_{u_0,u_1} between the interviewer random effects in Model A4 in Table 2.9 is not explained away by the inclusion of the area effect in Table 2.11. Indeed the estimated correlation of -0.77 is even more negative. To try to improve the understanding of the negative association between the random interviewer effects for the intercept and for the response indicator coefficient, a reparametrization, as in Equation (2.11), of the cross–classified models is considered. Tables 2.12 and 2.13 present models respectively for the dependent variables employment and academic qualification. In both tables the first model is the two–level random interviewer coefficient model, the second is the interviewer–area cross–classified model with random interviewer coefficient, the third is the cross–classified model with random area intercept, random interviewer coefficient for the response indicator and random interviewer coefficient for the nonresponse indicator and the last model is the third model plus individual–level characteristics.

For the models in Table 2.12 the focus is on the analysis of the random effects. In Models E9 and E7, the covariance between the two interviewer random effects is negative. This might indicate that the variation across interviewers differs for respondents and nonrespondents. According to the random effect estimates for the reparametrization in Model E7′, one can see that the variance of the interviewer random effects for the respondents is 0.085, whereas the variance of the interviewer random effects for the nonrespondents is 0.126. Therefore, the effects of interviewers (possibly combined with interviewer assignment effects) seem to

introduce more variation for the nonrespondents than for the respondents. In addition, after accounting for individual–level characteristics in Model E8′, the variance of the interviewer random effects for the nonrespondents seems to increase even more (0.171), whereas the variance of the interviewer random effects for the respondents decreases (0.062).

Table 2.12: Parameter estimates (with standard errors in brackets) for the models for employment (reparametrization)

Fixed effect	Model E9	Model E7	Model E7′	Model E8′
Constant	0.811 (0.090) **	0.861 (0.096) **	0.866 (0.097) **	-0.263 (0.166)
Response indicator				
Response	0.098 (0.096)	0.096 (0.099)	0.092 (0.100)	0.145 (0.106)
Sex				
Female				-0.717 (0.071) **
Marital status				
First marriage				-0.169 (0.089) *
Re–married				-0.269 (0.139) **
Sep & legally married				-0.509 (0.218) **
Divorced				-0.520 (0.135) **
Widowed				-0.861 (0.249) **
Student indicator				
Not full–time				1.861 (0.145) **
Ethnic group				
Mixed group				0.356 (0.496)
Asian group				-0.854 (0.213) **
Black group				-0.780 (0.291) **
Other group				-0.587 (0.298) **
Random effect				
Random intercept:				
Interviewer variance $\sigma^2_{u_0}$	0.143 (0.078)	0.122 (0.076)		
Area variance $\sigma^2_{v_0}$		0.230 (0.056)	0.228 (0.057)	0.298 (0.067)
Random coefficient:				
Interviewer (resp. coef.) variance $\sigma^2_{u_1}$	0.111 (0.083)	0.148 (0.105)		
Interviewer (resp. coef.) variance $\sigma^2_{u_1'}$			0.085 (0.034)	0.062 (0.031)
Interviewer (nonresp. coef.) variance $\sigma^2_{u_0'}$			0.126 (0.077)	0.171 (0.107)
Interv. resp.-nonresp coef. cov. $\sigma_{u_0'u_1'}$			0.033 (0.038)	0.020 (0.040)
Interv. intercept-coef. covariance $\sigma_{u_0u_1}$	-0.047 (0.068)	-0.093 (0.082)		
BDIC	5695.84	5637.75	5638.37	5293.95

The base categories for the explanatory variables are Nonresponse, Male, Single, Full-time student and White group.
Model E9 is the two-level random interviewer coefficient model, Model E7 is the interviewer-area cross-classified model with random interviewer coefficient, Model E7′ is the cross-classified model with random area intercept, random interviewer coefficient for the response indicator and random interviewer coefficient for the nonresponse indicator and Model E8′ is Model E7′ plus additional individual-level characteristics (explanatory variables). BDIC is the Bayesian Deviance Information Criterion.
** Significant at the 5% level
* Significant at the 10% level

In Models A9 and A7, from Table 2.13, the covariance between the two interviewer random effects is also negative. As in the models for employment, this

might also indicate that interviewer random variation is different between respondents and nonrespondents. As in the previous discussion, looking at the random effect estimates for the reparametrization in Model A7′, the variance of the interviewer random effects for the respondents is 0.271, whereas the variance of the interviewer random effects for the nonrespondents is 0.566. Therefore, the effects of interviewers (possibly combined with interviewer assignment effects) seem to introduce much higher variation for the nonrespondents than for the respondents.

Table 2.13: Parameter estimates (with standard errors in brackets) for the models for academic qualification (reparametrization)

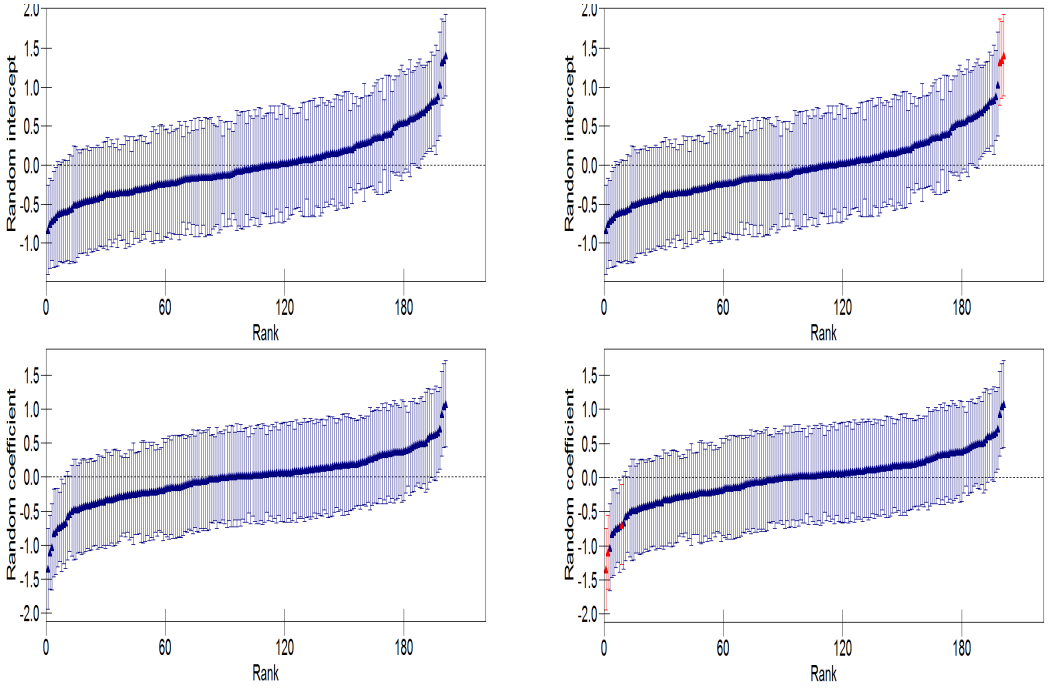
Fixed effect	Model A9	Model A7	Model A7′	Model A8′
Constant	-1.249 (0.121) **	-1.295 (0.125) **	-1.299 (0.132) **	-0.690 (0.143) **
Response indicator				
Response	0.356 (0.125) **	0.359 (0.125) **	0.363 (0.131) **	0.366 (0.136) **
Marital status				
First marriage				-0.129 (0.097)
Re–married				-0.753 (0.168) **
Sep & legally married				-0.438 (0.245) *
Divorced				-0.302 (0.152) **
Widowed				-0.431 (0.337)
Health				
Fairly good				-0.509 (0.096) **
Not good				-0.706 (0.155) **
Dependent child				
Yes				-2.783 (0.350) **
Age				
35 – 49				-0.210 (0.096) **
50 – 64				-0.816 (0.114) **
Urban/rural indicator				
Rural				0.354 (0.125) **
Random effect				
Random intercept:				
Interviewer variance $\sigma^2_{u_0}$	0.564 (0.224)	0.566 (0.246)		
Area variance $\sigma^2_{v_0}$		0.347 (0.078)	0.353 (0.079)	0.336 (0.079)
Random coefficient:				
Interviewer coef. variance $\sigma^2_{u_1}$	0.476 (0.213)	0.532 (0.241)		
Interviewer (resp. coef.) variance $\sigma^2_{u_1'}$			0.271 (0.073)	0.290 (0.077)
Interv. (nonresp. coef.) variance $\sigma^2_{u_0'}$			0.566 (0.252)	0.652 (0.267)
Interv. resp.-nonresp coef. cov. $\sigma_{u_0'u_1'}$			0.155 (0.098)	0.132 (0.105)
Interv. intercept-coef. covariance $\sigma_{u_0u_1}$	-0.332 (0.198)	-0.412 (0.224)		
BDIC	5552.91	5466.26	5466.80	5197.64

The base categories for the explanatory variables are Nonresponse, Single, Good, Not a dependent child, 16 – 34 and Urban.
Model A9 is the two-level random interviewer coefficient model, Model A7 is the interviewer-area cross-classified model with random interviewer coefficient, Model A7′ is the cross-classified model with random area intercept, random interviewer coefficient for the response indicator and random interviewer coefficient for the nonresponse indicator and Model A8′ is Model A7′ plus additional individual-level characteristics (explanatory variables). BDIC is the Bayesian Deviance Information Criterion.
** Significant at the 5% level
* Significant at the 10% level

With similar pattern to the one from the model for employment, after including individual–level characteristics in Model A8′, the variance of the interviewer random effects for the nonrespondents seems to increase even more (0.652), whereas the variance of the interviewer random effects for the respondents does not change much (0.290).

2.4.5 Residual analysis

Further analysis is needed to explain the significant interviewer assignment effects in Model A8 from Table 2.11. Figures 2.4(a)–(b) illustrate caterpillar plots of the residuals from the two–level random coefficient logistic model for the dependent variable academic qualification. In the caterpillar plot of the random effects for the intercept, there are three residuals (highlighted in Figure 2.4(b)), which represent interviewer random effects, with much higher departures from the overall average predicted by the fixed parameter compared to the others.

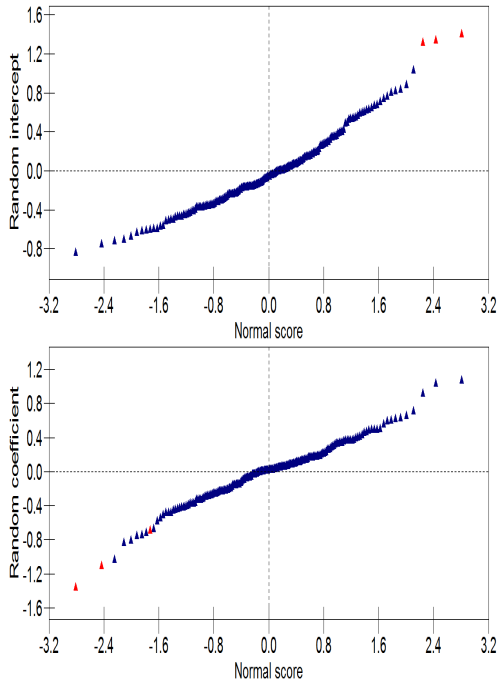


(a) Original plot

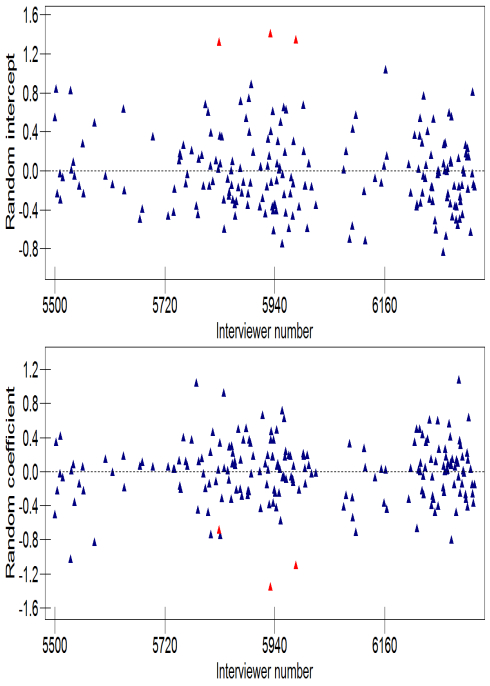
(b) Three residuals highlighted

Figure 2.4: Caterpillar plots of the interviewer–level residuals

By looking at the normal probability plots and the scatterplots respectively in Figures 2.5(a)–(b), the three residuals appear to be unusual. Since these three highlighted residuals seem to be extreme, it is advisable to investigate their influence on the results by repeating the analyses on Table 2.11 removing the cases of all individuals interviewed by the three corresponding interviewers.



(a) Normal probability plots with three residuals highlighted



(b) Scatterplots of the residuals against interviewer number

Figure 2.5: Diagnostic plots

Table 2.14 presents the new models for academic qualification. In the two–level random intercept logistic model (Model A3R), the interviewer assignment effect is significant at the 5% level. As the difference between response and nonresponse within an interviewer is allowed to vary across interviewers (Model A4R), the interviewer assignment effect is no longer significant at the 5% level. On the other hand, the interviewer–level variance for the coefficient of the response indicator is significant, which means that the nonresponse bias depends on interviewers. Furthermore, since interviewer effects are potentially confounded with area effects, the cross–classification of interviewers and areas is added (Model A8R). As it can

be noticed, the interviewer–level variance for the intercept is no longer significant at the 5% level, whereas the area–level variance is highly significant indicating that the level–two random intercept is representing an area effect. In addition, there is still evidence of interviewer effects on nonresponse bias. Note that if the nonsignificant random interviewer intercept is removed from Model A8R, the interviewer effect on nonresponse bias becomes much stronger (Model A10R).

Table 2.14: Parameter estimates (with standard errors in brackets) for the models for academic qualification (without cases from 3 interviewers)

Fixed effect	Model A3R	Model A4R	Model A8R	Model A10R
Constant	-0.671 (0.121) **	-0.671 (0.122) **	-0.717 (0.129) **	-0.700 (0.119) **
Response indicator				
Response	0.398 (0.105) **	0.391 (0.118) **	0.400 (0.122) **	0.339 (0.114) **
Marital status				
First marriage	-0.124 (0.092)	-0.121 (0.095)	-0.110 (0.098)	-0.126 (0.102)
Re–married	-0.779 (0.165) **	-0.776 (0.166) **	-0.776 (0.171) **	-0.825 (0.177) **
Sep & legally married	-0.554 (0.241) **	-0.553 (0.244) **	-0.498 (0.250) **	-0.463 (0.258) *
Divorced	-0.401 (0.151) **	-0.389 (0.153) **	-0.370 (0.158) **	-0.391 (0.160) **
Widowed	-0.388 (0.327)	-0.392 (0.325)	-0.353 (0.335)	-0.380 (0.342)
Health				
Fairly good	-0.475 (0.094) **	-0.485 (0.094) **	-0.497 (0.097) **	-0.460 (0.099) **
Not good	-0.732 (0.157) **	-0.745 (0.157) **	-0.725 (0.161) **	-0.746 (0.164) **
Dependent child				
Yes	-2.680 (0.344) **	-2.706 (0.345) **	-2.756 (0.347) **	-2.778 (0.354) **
Age				
35 – 49	-0.202 (0.092) **	-0.209 (0.095) **	-0.218 (0.097) **	-0.209 (0.101) **
50 – 64	-0.802 (0.110) **	-0.809 (0.111) **	-0.841 (0.115) **	-0.871 (0.120) **
Urban/rural indicator				
Rural	0.284 (0.116) **	0.307 (0.117) **	0.341 (0.124) **	0.399 (0.127) **
Random effect				
Random intercept:				
Interviewer variance $\sigma^2_{u_0}$	0.342 (0.072)	0.229 (0.136)	0.245 (0.154)	
Area variance $\sigma^2_{v_0}$			0.295 (0.073)	0.360 (0.138)
Random coefficient:				
Interviewer coef. variance $\sigma^2_{u_1}$		0.310 (0.154)	0.349 (0.181)	0.470 (0.172)
Interv. inter.-coef. covariance $\sigma_{u_0u_1}$		-0.064 (0.120)	-0.136 (0.149)	
Interv. inter.-coef. correlation $\rho_{u_0u_1}$		-0.240	-0.466	
BDIC	5135.04	5126.34	5063.56	5027.69

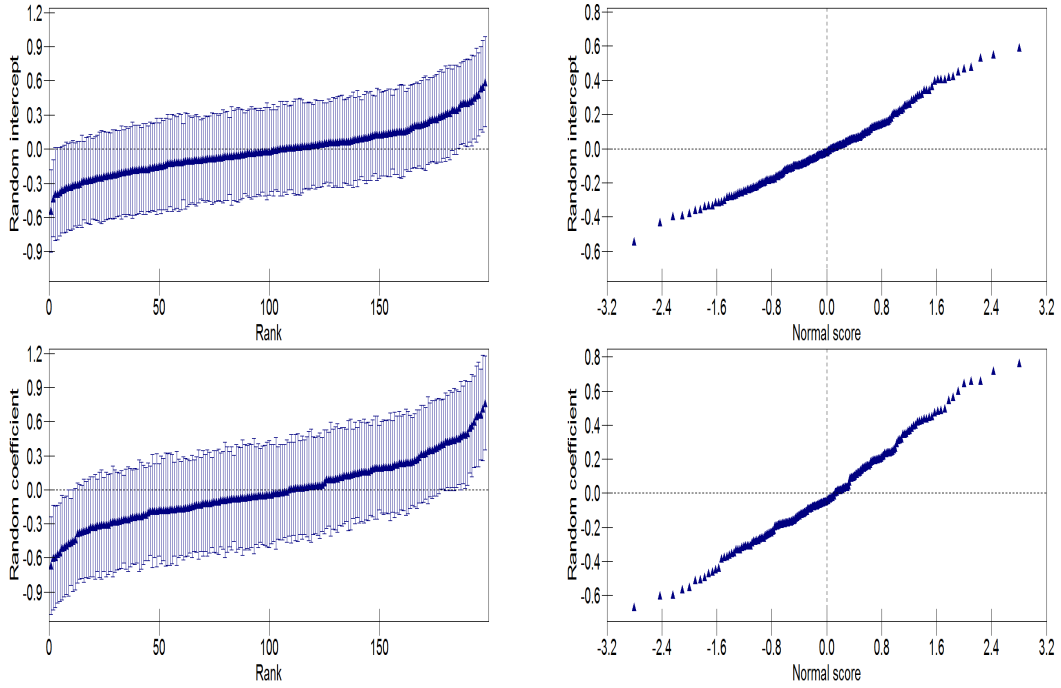
The base categories for the explanatory variables are Nonresponse, Single, Good, Not a dependent child, 16 – 34 and Urban. Model A3R is the two-level random intercept logistic model, Model A4R is the two-level random coefficient logistic model, Model A8R is the interviewer-area cross-classified model with random coefficient and Model A10R is Model A8R with no random interviewer intercept. The *R* in the models' labels refers to 'removing the cases from the three unusually assigned interviewers'. BDIC is the Bayesian Deviance Information Criterion.

** Significant at the 5% level

* Significant at the 10% level

Furthermore, considering the analysis without the three interviewers, the caterpillar plots in Figure 2.6(a) illustrate smaller variations than the ones in Figure

2.4(a). Also, the normal probability plots in Figure 2.6(b) look fairly linear, indicating that the normal assumption for the residuals seems reasonable. Thus, it seems that those three unusually assigned interviewers induced the significant interviewer assignment effects in Model A8 from Table 2.11.



(a) Caterpillar plots (b) Normal probability plots

Figure 2.6: Diagnostic plots without the three interviewers (residuals)

2.5 Conclusions

This study analysed the detection of interviewer effects on nonresponse bias. Multilevel logistic models were applied to model the interviewer effects on nonresponse bias on dependent variables of interest. The analysis considered data from the 2001 UK LFS linked dataset. A key advantage of these data is that census records are available for both respondents and nonrespondents. Unlike other studies, this research has considered the response indicator as an explanatory variable in the models rather than a dependent variable. The main contribution of this study is

to introduce a method to assess quite easily the nonresponse bias and the bias due to interviewer, when census linked data or auxiliary variables are available.

The estimated coefficients from the standard (single–level) logistic regressions (Models E1 and A1) can be used to test for possible nonresponse bias for the dependent variables of interest. Including other explanatory variables in the model and allowing the intercept to vary across interviewers, it is found that for academic qualification (Model A3) and with lesser extent (at the 10% level) for employment (Model E3), there is evidence that the nonresponse bias is significant for all interviewers, even after controlling for other explanatory variables. The final models all include explanatory variables at individual–level to control for the different cases allocated to interviewers.

In the random coefficient model for employment (Model E4), the coefficient of the response indicator is not significantly different from zero, which means that, on average, there is no nonresponse bias. The nonresponse interviewer–level variance is not significant either, meaning that there is no variation due to interviewer on the coefficient of the response indicator.

In the model for academic qualification (Model A4), the coefficient of the response indicator is significantly (p–value < 0.05) different from zero. This means that there is evidence that the odds of being highly academically qualified for respondents differs from the odds of being highly academically qualified for nonrespondents, which consequently indicates that, on average, there is nonresponse bias. In addition, the nonresponse interviewer–level variance is also significant (p–value < 0.05), meaning that the nonresponse bias varies by interviewer.

Since interviewer effects may be confounded with area effects due to interviewers being assigned to specific areas, the models for employment and for academic qualification also include the cross–classification of interviewer and area random effects.

By analysing the cross–classified models for employment, there is no interviewer assignment variation left, after including the cross–classification of interviewer and area random effects in the model (Model E6). On the other hand, for the cross–classified models for academic qualification, there is evidence that after including the crossclassification of interviewer and area random effects in the model (Model A6), the interviewer assignment variation reduces. This might mean that the interviewer assignment variation is only partially explained by the area random effects included in the model.

Considering the reparametrization of the cross–classified models for both employment and academic qualification, one can conclude that the variation of the interviewer random effects is larger for the nonrespondents than for the respondents, which might explain the negative correlation of the interviewer random effects between the intercept and the coefficient.

The cases from three extreme interviewers were removed from the sample after a residual analysis that identified these interviewers as unusually assigned. Thus, the models for academic qualification were refitted without these cases. For the refitted models, the interviewer assignment effect is significant at first (Model A3R). Then, allowing the difference between response and nonresponse within an interviewer to vary across interviewers (Model A4R), the interviewer assignment effect is no longer significant whereas the interviewer–level variance for the coefficient of the response indicator is significant. Finally, after controlling for the cross–classification of interviewers and areas (Model A8R), the interviewer–level variance for the intercept is no longer significant, whereas the area–level variance is highly significant. This indicates that the interviewer assignment effects are in fact attributed to areas rather than to interviewers. Also, there is evidence that the nonresponse bias for academic qualification might be driven by interviewers,

especially, after removing the nonsignificant random interviewer intercept (Model A10R).

One important practical implication for the detection of these interviewer effects is that it could be integrated an overall strategy to better allocate the survey resources in order to control the total survey error. For example, time and financial resources would only be invested in improving interviewer training given there is evidence that the interviewers affect the nonresponse biases significantly. Even if those effects are not found to be significant, the organizations could also report the application of the proposed methodology as part of their quality monitoring of the estimates produced, increasing the survey users trust on the survey results. In addition, lessons could be learnt from those interviewers that introduce small or no bias. So, survey agencies should pay attention to what these interviewers do in the field and use this knowledge to train the other interviewers accordingly. Furthermore, if interviewer characteristics are available in the dataset, these characteristics could be included in the models in order to examine whether the interviewer–level variation decreases. Also, the model could include an interaction between interviewer characteristics and the response indicator to investigate the interviewer effects within groups of interviewers.

Other binary dependent variables from the 2001 UK LFS linked dataset were considered in the model fitting process. However, the response indicator was not significant to explain these dependent variables. Groves (2006) and (Groves and Peytcheva, 2008) report that different estimates within a survey may vary with respect to nonresponse bias. As a general rule, the existence of bias depends if the likelihood of participation is related to this estimate.

Potential limitations in the research in this chapter may be due to a large number of deleted cases from the LFS dataset. Also, in the analysis, 80 cases were assigned as respondents as a working assumption. However, these cases could all belong to

the nonrespondent pool, or only part of them could belong to the nonrespondent pool or they all could actually belong to the respondent pool. Although they are only a small number of cases, further analysis assigning these 80 cases as nonresponse or discarding them from the sample is needed to analyse the impact of this assumption on the results. In the next chapter, the same methodology is applied to two other datasets, using a different type of data collection mode and auxiliary variables.

Chapter 3

Interviewer Effects on Nonresponse Bias: Further Investigations

3.1 Introduction

The literature on nonresponse has identified the survey interviewer as one influential factor on nonresponse in surveys (Durrant et al., 2010; O'Muircheartaigh and Campanelli, 1999; Snijkers et al., 1999). Whether and how interviewers vary in the way that they contribute to survey nonresponse is of some interest to survey organisations, since the recruitment, training and assignment of interviewers to respondents are matters about which they have some control. Understanding the variation between interviewers may therefore offer ways to reduce nonresponse error. However, a persistent question is whether interviewers could influence nonresponse bias.

In an analysis of nonresponse bias, it is essential to have auxiliary variables for both respondents and nonrespondents (Groves and Peytcheva, 2008). In general, these auxiliary variables are additional information collected apart from the main survey (Smith, 2011). This information may come from sampling frames, administrative records, follow up of nonrespondents, paradata, census records linked to each sample case or time invariant variables from previous waves in a longitudinal study, among other sources.

Loosveldt and Beullens (2014) argue that research on interviewer effects on nonresponse bias is quite rare, especially because of unavailability of relevant auxiliary information. In their paper, they propose a method to assess interviewer effects on nonresponse bias using quality indicators from an evaluation of the sample units' dwellings undertaken by trained experts prior to the actual survey. This evaluation provided rich auxiliary information for respondents and nonrespondents. The results from the application suggest that some interviewers tend to generate bias and others do not.

One of the interests in this chapter is to assess interviewer effects on nonresponse bias using a telephone survey. In the middle 1980's the percentage of surveys conducted over the telephone had increased relative to other survey modes. Thus, concerns regarding characteristics of interviewers that could negatively influence response rates and data quality in telephone surveys had received much attention from researchers. Groves and Fultz (1985) examine interviewer gender as a personal characteristic that may affect the cost and quality of telephone survey data. They find that male interviewers tend to have lower response rates and higher refusal rates than female interviewers as a consequence of male interviewers being less experienced. On one hand, males are less likely than females to interview female respondents, older respondents and poor people. On the other hand, they are more likely to interview people who are working compared to female interviewers.

In contrast, there are no real differences between males and females on the total per minute interview cost.

Even though telephone surveys became very popular in the 1990's, especially because of their lower costs and stricter interviewer control compared to face–to–face surveys, in more recent years this survey mode has been facing serious problems since unsolicited telephone calls from, for instance, goods sellers, products promoters and charitable causes have been driving away potential survey respondents that misinterpret a survey request for a sales offer (De Leeuw and Hox, 2004; Keeter et al., 2006).

Researchers have been debating about which interview mode yields smaller biases. For instance, Biemer (2001) compares the quality of data in terms of smaller biases meaning better quality. To maintain comparability, similar questionnaires are used to collect data from a face–to–face survey and a telephone interviewing from a Computer Assisted Telephone Interviewing (CATI) system. In the paper, he criticises comparison approaches that attribute the whole mode–related biases solely to measurement bias and suggests that other factors such as nonresponse bias may also make up the mode bias. Biemer proposes an approach to estimate these two sources of biases based on latent class analysis to obtain the measurement bias as a function of the estimates of the response probabilities. Using these estimates and information acquired from a telephone follow up survey of the face–to–face survey nonrespondents, the estimates of the nonresponse bias are obtained for the face–to–face and CATI modes. He concludes that the nonresponse bias is larger for the CATI mode compared to the face–to–face mode. Whilst, the measurement bias for the face–to–face mode is larger than the one for the CATI mode. In addition, although the telephone survey has smaller response rate, both interview modes yield similar data quality.

Nonresponse bias may arise as a consequence of the recruitment process of the survey respondents. Since there is a systematic decrease of response rates in sample surveys over the years, another useful investigation regarding the quality of survey estimates from different surveys could be to examine, for a particular survey mode, whether higher response rates yield less biased survey estimates. Keeter et al. (2000) compare estimates from two telephone surveys using identical questionnaires but different levels of interview effort. One of the studies, called "standard", is conducted in a short period of time (5 days) and selected either the younger adult male or the oldest adult female who is at home. The other study, called "rigorous", is carried out in a much longer period (8 weeks) and the respondents are randomly selected among all adults in the household. The standard study achieves a 36.0 percent response rate, whereas the rigorous one achieves 60.6 percent. They find that the two surveys yield almost similar results with most differences occurring among demographic items.

The literature has shown that interviewer effects may vary by survey mode. In face–to–face surveys there is more opportunity for the interaction between interviewers and respondents than in telephone surveys (West et al., 2013). As a consequence, interviewers perceived characteristics, such as gender, race and age, might have an influence on how respondents edit their answers before telling them (Davis et al., 2010). In this sense, respondents are more likely to provide socially desirable responses, especially when they are queried about sensitive issues. Groves and Magilavy (1986) emphasize that interviewer effects are smaller in telephone surveys than in face–to–face surveys. However, there is no evidence that these effects are negligible.

In this chapter, it is also of interest to investigate interviewer effects on nonresponse bias in a longitudinal survey. In the context of longitudinal designs, particularly for panel studies, to send the same interviewer back to the same respondent at later

waves is advisable to help avoiding refusals. Campanelli and O'Muircheartaigh (1999) investigate the impact of interviewer continuity on response rates. They make use of the interpenetrated sample design experiment in wave 2 of the British Household Panel Survey. Although they do not find any evidence of interviewer continuity effects on nonresponse, they strongly recommend that the continuity of interviewers in panel designs should not be ignored.

In another line of research, Pickery and Loosveldt (2002) discuss how interviewer effects in a panel survey can be analysed using multilevel models. In their approach, they consider the repeated measurements nested within respondents and, where there is interviewer continuity across the two waves, respondents nested within interviewers. When there is a change of interviewers across the two waves, they consider the repeated measurements nested within the cross–classification of respondents and interviewers since the structure is not purely hierarchical. The dependent variable used in the models is the number of "don't know" responses. They conclude that the interviewer effects are significant for all analyses. This gives evidence of the need to improve interviewer training on asking hard questions and dealing with "don't know" answers.

Chapter 2 proposed a methodology to analyse interviewer effects on nonresponse bias. This methodology consisted of fitting multilevel logistic models for binary dependent variables of interest using the response indicator as an explanatory variable. These models were applied to an initial dataset of a cross–sectional face–to–face survey, taking advantage of linked census data. However, past research has recognised that if interviewers have influence on nonresponse bias in sample estimates from face–to–face surveys, their influence on nonresponse bias for telephone surveys tends to be more pronounced. For example, Singer et al. (1983) report that in telephone surveys, compared to face–to–face, usually there is a smaller number of interviewers administering a much larger number of interviews. As a

consequence, their performance may have a negative effect on response rates and response quality.

This chapter aims to investigate further the assessment of interviewer effects on nonresponse bias. In contrast with Chapter 2, where the dataset used comes from the linkage of a cross–sectional face–to–face survey and census individual records, the key feature here is the use of a dataset from a different data collection mode, a telephone survey, and a dataset from a longitudinal study. Another distinction from Chapter 2 is that these two datasets are linked to alternative sources of auxiliary variables in order to provide information on both respondents and nonrespondents. It is important to consider these alternative sources of variables since census records are not always available. For instance, Smith (2011) argues that decennial census records may be outdated to be linked to survey data. Also, the access to linked government records may not be available to many statistical practitioners. For the analyses, data from a telephone survey from the Netherlands are linked to administrative records, whereas data from the 10th wave of the British Household Panel Survey (BHPS) are linked to time invariant variables from the first wave. Therefore, in addition to use data from different surveys, to take advantage of other types of auxiliary variables provides a deeper understanding of the interviewer effects on nonresponse bias. Multilevel logistic models similar to those considered in Chapter 2 are used to analyse the datasets.

This chapter is divided into five sections, including this introduction. A detailed explanation about the datasets and the variables considered in this study is provided in Section 3.2. Section 3.3 reviews the definition of the statistical models utilized in the data analysis. In Section 3.4, the main results are discussed and the conclusions are presented in Section 3.5.

3.2 Datasets

To continue the investigation of interviewer effects on nonresponse bias, two different datasets are considered in this chapter. The datasets come from a telephone survey, the Consumer Confidence Survey, linked to administrative data and from a mixed (mostly face–to–face) mode longitudinal survey, the British Household Panel Survey. The next two sections discuss these datasets further.

3.2.1 Consumer confidence survey

The Consumer Confidence Survey (CCS) is a Computer Assisted Telephone Interviewing (CATI) survey from the Netherlands. The CCS was conducted monthly throughout 2005 with equal sample sizes per month and equal sampling weights for all households. The cases contained in the CCS dataset are from one year of data collection. The data were collected by a total of 60 interviewers working in a centralised telephone unit. The experience and ages of interviewers varied. This survey was carried out by Statistics Netherlands.

This research makes use of the response indicator, the interviewer number and the distinction between contact and cooperation from the CCS records. Since the aim of this study is to investigate nonresponse bias induced by interviewers, information on respondents and nonrespondents is essential. The dataset also contains auxiliary variables at household–level from Dutch registry data.

The auxiliary variables available in the dataset are gender of the people in the household, marital status in the household, type of housing (whether it is owned or not), urbanization degree, type of household, size of household, number of jobs

in the household at the time of interview, mean age of household kernel[1], mean income of household kernel in the month of interview and house value.

The response indicator had been defined by assigning 1 to respondents to the CCS and 0 to nonrespondents. Table 3.1 presents the frequency distribution for the response indicator conditioning on contact.

Table 3.1: Frequency distribution for the CCS linked dataset response indicator

Response indicator	Frequency	Percent
Nonresponse	4741	29.1
Response	11524	70.9
Total	16265	100.0

The dependent variables used to investigate the possibility of nonresponse bias are number of jobs in the household at the time of interview, which is coded as 1 if there is at least one job in the household or as 0 otherwise; type of housing, which is coded as 1 if the property (house or flat) is owned by the householder or as 0 if it is rented; and size of household, which is coded as 1 if there is one occupier in the household or as 0 if there are at least two occupiers.

Table 3.2 presents the percentages for each dependent variable for nonrespondents and respondents. The differences in the percentages of the dependent variables between nonrespondents and respondents indicate nonresponse bias.

Table 3.2: Distributions (percentages) of the CCS linked dataset dependent variables.

Dependent variable	Nonrespondents	Respondents	Total
Jobs in the household	47.90	61.63	57.63
Type of housing	54.55	64.76	61.78
Size of household	35.71	24.09	27.48

[1] The kernel consists of the head of the household plus his/her partner, if there is one. Hence the kernel has one or two people.

3.2.2 British household panel survey

Another dataset considered in this study comes from the British Household Panel Survey (BHPS), which is an annual longitudinal survey. The BHPS has been carried out by the Economic and Social Research Council (ESRC) UK Longitudinal Study Centre together with the Institute for Social and Economic Research (ISER) at the University of Essex.

The survey design of the BHPS is a multistage stratified cluster design covering the UK (O'Muircheartaigh and Campanelli, 1999). The data collection occurs in different ways: face–to–face interview, telephone interview and self–completion questionnaire. Since the BHPS collects social–economic information from each sampled household member aged 16 or over, the main objective of this survey is to enable society to have a deeper understanding of social and economic changes in the UK over the years. Beginning in 1991, a total of 18 waves of the BHPS has been carried out so far.

In order to analyse the possibility of nonresponse bias on dependent variables of interest, time invariant variables from wave 1, which was carried out in 1991, are used as auxiliary variables for respondents and nonrespondents from wave 10, carried out in 2010. One variable of interest from the latter wave is the interviewer ID number. This variable is essential for the data analysis.

Regarding the process of the data linkage, the BHPS data files are merged using the following steps. The interviewer ID number variable is only present in the household sample files. Hence, firstly, the household sample file from wave 10 (jhhsamp) is matched with the individual sample file (jindsamp) from the same wave using a household identification number (JHID). This matching creates a new file (hh_ind_samp) containing household and individual records, as well as interviewer ID number, for all individual cases. Secondly, the created file is matched with the

individual respondent file from wave 10 (jindresp) through the JHID and a wave 10 person number (JPNO) to create another file (jBHPS). Finally, the jBHPS file is matched with the individual sample file from wave 1 (aindsamp) using a cross–wave person identifier (PID) to create the final data file (ajBHPS).

A response indicator is defined as follows: if a sample member of a household takes part in wave 10 of the BHPS, it is assigned the value 1 (response), whereas if a sample member of a household does not take part in wave 10, it is assigned the value 0 (nonresponse). Table 3.3 presents the frequency distribution for the response indicator.

Table 3.3: Frequency distribution for the response indicator of wave 10 of the BHPS

Response indicator	Frequency	Percent
Nonresponse	625	9.5
Response	5971	90.5
Total	6596	100.0

The actual nonresponse percentage in wave 10 of the BHPS is nearly 40%. However, many respondent and nonrespondent cases were deleted in the process of merging individual records from wave 10 with the initial records of the same individuals from wave 1. Most of these deleted cases were caused by item nonresponse and sample extensions, such as 1,500 households from Scotland and Wales joined the sample in 1999, new members become eligible after original sample members initiate new families, children become eligible as they reach the age of 16. In addition, many cases were deleted because they did not have an interviewer ID number associated with their records. Therefore, the nonresponse percentage for the analysis sample is 9.5%, as in Table 3.3. Possible implications of these deletions for the results of this research are discussed in Section 3.5.

There are only a few time invariant variables in the BHPS dataset. Therefore, in addition to the potential explanatory variables gender, age and ethnicity, the dependent variables considered in this study are mother working when interviewee was 14 years old and father working when interviewee was 14 years old. Both dependent variables are collected across all waves of the BHPS. However, except for wave 1, in subsequent waves these variables are collected only for new sampled cases. This is the main reason to use variables from wave 1 as the auxiliary variables.

The binary coding for the dependent variable mother working is 1 if the interviewee's mother was working when he/she was 14 and 0 otherwise. Similarly, the coding for father working is 1 if the interviewee's father was working when he/she was 14 and 0 otherwise. Table 3.4 presents the percentages for each dependent variable for nonrespondents and respondents. In both cases, the difference in the observed percentages between respondents and nonrespondents indicate possible nonresponse bias.

Table 3.4: Distributions (percentages) of the BHPS dependent variables.

Dependent variable	Nonrespondents	Respondents	Total
Mother working at age 14	43.28	45.50	45.29
Father working at age 14	89.15	91.44	91.23

Since the interview mode used in the BHPS is mostly face–to–face, interviewer effects are potentially confounded with area effects. According to Tables 3.5 and 3.6, interviewers are not strictly nested within areas since there are interviewers assigned to more than one area and some areas are covered by more than one interviewer. Thus, the multilevel analysis should take areas into account as random effects cross–classified with interviewer effects.

Table 3.5: Frequency distribution of number of interviewers per area (BHPS)

No. of interviewers per area	No. of areas
1	102
2	89
3	51
4	12
5	6
6	1
Total	261

Table 3.6: Frequency distribution of number of interviewers per area (BHPS)

No. of interviewers per area	No. of areas
1	195
2	66
3	20
4	13
5	1
6	3
55	1
Total	299

Respectively for the CCS linked dataset and for the BHPS dataset, household–level and individual–level characteristics (explanatory variables) are included in the models to take into account any variation due to different types of sample members assigned to interviewers. Frequency distributions for the explanatory variables are given in Appendix B.

3.3 Statistical models

The same types of multilevel models (Goldstein, 2011) used in Chapter 2 are also considered to model target binary dependent variables from the CCS and BHPS

datasets. One of the multilevel models of interest is the model in Equation (2.5)

$$\begin{aligned} y_{ij} \mid r_{ij}, \boldsymbol{x}_{ij}, \boldsymbol{u}_0 \;&\sim\; \text{indep Bernoulli}(\pi_{ij}), \quad i = 1, \ldots, n_j, j = 1, \ldots, J \;\text{ and} \\ &\qquad\qquad \boldsymbol{u}_0 = (u_{01}, ..., u_{0J})^\top, \\ \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) &= \beta_{0j} + \beta_1 r_{ij} + \boldsymbol{\beta}^\top \boldsymbol{x}_{ij}, \\ \beta_{0j} &= \beta_0 + u_{0j}. \end{aligned} \tag{3.1}$$

For the CCS analysis, the β_{0j} represent interviewer assignment effects because the interviewers of the survey were not involved in the collection of the administrative data, from which the dependent variables were taken. In the case of the BHPS analysis, the interviewer ID numbers are from wave 10, whereas the dependent variables are from wave 1. Therefore, the β_{0j} are also referred to as interviewer assignment effects because the interviewers from the two waves were unlikely to be the same or to interview the same respondent.

Another model of interest in this chapter is the one in Equation (2.7)

$$\begin{aligned} y_{ij} \mid r_{ij}, \boldsymbol{x}_{ij}, \boldsymbol{u}_1, ..., \boldsymbol{u}_J \;&\sim\; \text{indep Bernoulli}(\pi_{ij}), \quad i = 1, \ldots, n_j \;\text{ and} \\ &\qquad\qquad j = 1, \ldots, J, \\ \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) &= \beta_{0j} + \beta_{1j} r_{ij} + \boldsymbol{\beta}^\top \boldsymbol{x}_{ij}, \\ \beta_{0j} &= \beta_0 + u_{0j}, \\ \beta_{1j} &= \beta_1 + u_{1j}, \end{aligned} \tag{3.2}$$

An additional model of interest is the cross–classified model given in Equation (2.9), where the systematic component is given by

$$\log\left(\frac{\pi_{i(jk)}}{1-\pi_{i(jk)}}\right) = \beta_0 + \beta_1 r_{i(jk)} + u_{1j} r_{i(jk)} + \boldsymbol{\beta}^\top \mathbf{x}_{i(jk)} + u_{0j} + v_{0k}. \tag{3.3}$$

The modelling strategy applied here is similar to the one in Chapter 2.

In this chapter, the models are estimated using the second order predictive quasi–likelihood (PQL) for the starting values. Then, Markov Chain Monte Carlo (MCMC) with 150,000 iterations after a burn–in of 10,000 is used. As in Chapter 2, the models are fitted using `MLwiN` (Rasbash et al., 2012).

3.4 Results

This section discusses and presents the main results from the analyses of the two datasets (CCS linked dataset and BHPS).

3.4.1 Descriptive analysis for the CCS dataset

As previously mentioned the CCS is a telephone survey in which a CATI management system automatically assigns phone numbers to interviewers. Although Statistics Netherlands sends an individual pre–notification letter explaining about the content of the survey to the target population, there is no guarantee that the sampled units will respond to the survey request. To illustrate this, a histogram of the interviewers' response rate is provided in Figure 3.1. An inspection of the histogram suggests that the distribution of the interviewers' response rates is skewed to the left. However, no interviewer had a 100% response rate.

Based on a descriptive analysis (refer to Appendix C for the cross–tabulations) for the dependent variables of interest, it is found that people who take part in the survey, who live in a mixed gender household, who are registered partner or live in multiple household and who are 25 to 54 years old are more likely to have at least 1 job in the household. Also, people who respond to the survey, who live

in a mixed gender household, who are registered partner or married and who are 30 to 54 years old are more likely to own their properties. Whilst, as might be expected, people who do not respond to the survey are male, widowed, more than 69 years old and more likely to live alone.

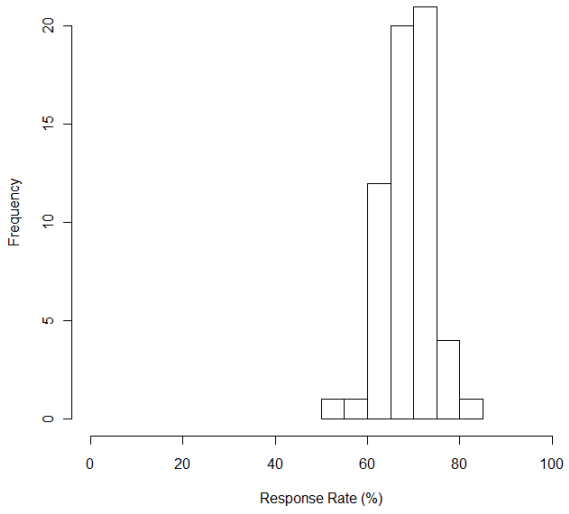


Figure 3.1: Histogram of the interviewers' response rates (CCS linked dataset)

3.4.2 Statistical modelling for the CCS linked dataset variables

Some of the models fitted in Chapter 2 are also fitted here. They are the standard (single–level) logistic model with only the response indicator as the explanatory variable, the single–level logistic model with additional individual–level characteristics (explanatory variables), the two–level random intercept logistic model with other explanatory variables and the two–level random coefficient logistic model with other explanatory variables. In addition to these, two other models are also

fitted: the two–level random intercept logistic model with only the response indicator as the explanatory variable and the two–level random coefficient logistic model also with only the response indicator as the explanatory variable.

The parameter estimates for the models for the dependent variables jobs in the household and type of housing are presented respectively in Tables 3.7 and 3.8. In Model J1 from Table 3.7, the odds of having at least one job in the household are different between respondents and nonrespondents, i.e. there is significant bias, since the coefficient of the response indicator is significantly different from zero. As other household–level characteristics (explanatory variables) are included into the model (Model J2), the coefficient of the response indicator is still significant at the 5% level, even though these explanatory variables explain part of the nonresponse bias.

The intercept in Model J1 is allowed to vary across interviewers (Model J3). In this model, there is practically no change in the coefficient of the response indicator or in its significance, compared to Model J1. Additionally, the Wald test statistic for the inclusion of the interviewer–dependent variance is 2.127 (on 1 degree of freedom) which is not significant (p–value = 0.145). This is a quite interesting finding, since it means that there are no interviewer assignment effects on the probability of having at least one job in the household. Also, the reduction on the Bayesian Deviance Information Criterion (BDIC) obtained with the inclusion of this parameter is quite small, suggesting that Model J1 should be preferable. The BDICs from Models J1 to J6 are presented in Table 3.7.

Table 3.7: Parameter estimates for the models for jobs in the household

Fixed effect	Model J1	Model J2	Model J3	Model J4	Model J5	Model J6
Constant	-0.083 (0.029) **	0.732 (0.206) **	-0.091 (0.031) **	-0.088 (0.032) **	0.725 (0.190) **	0.773 (0.200) **
Response indicator						
Response	0.557 (0.035) **	0.214 (0.057) **	0.558 (0.034) **	0.552 (0.044) **	0.214 (0.057) **	0.195 (0.068) **
Gender						
Female		-0.154 (0.092) *			-0.155 (0.093) *	-0.150 (0.093)
Mixed		0.899 (0.078) **			0.899 (0.079) **	0.906 (0.078) **
Age						
25 – 29		2.004 (0.302) **			2.012 (0.289) **	1.968 (0.294) **
30 – 34		1.269 (0.229) **			1.274 (0.216) **	1.236 (0.223) **
35 – 39		0.765 (0.214) **			0.774 (0.200) **	0.732 (0.206) **
40 – 44		0.649 (0.209) **			0.655 (0.195) **	0.611 (0.201) **
45 – 49		0.548 (0.207) **			0.556 (0.193) **	0.510 (0.199) **
50 – 54		0.203 (0.204)			0.211 (0.191)	0.168 (0.196)
55 – 59		-0.536 (0.199) **			-0.531 (0.185) **	-0.574 (0.191) **
60 – 64		-2.161 (0.201) **			-2.157 (0.187) **	-2.207 (0.192) **
65 – 69		-3.528 (0.212) **			-3.525 (0.198) **	-3.579 (0.204) **
70+		-5.434 (0.234) **			-5.433 (0.224) **	-5.489 (0.229) **
Random effect						
Random intercept:						
Interviewer variance $\sigma^2_{u_0}$			0.006 (0.004)	0.006 (0.005)	0.008 (0.007)	0.011 (0.010)
Random coefficient:						
Interviewer coef. variance $\sigma^2_{u_1}$				0.031 (0.013)		0.052 (0.035)
Interv. inter.-coef. covariance $\sigma_{u_0 u_1}$				-0.003 (0.007)		-0.017 (0.018)
Interv. inter.-coef. correlation $\rho_{u_0 u_1}$				-0.261		-0.681
BDIC	21914.69	10448.68	21910.42	21904.69	10448.06	10440.08

The base categories for the explanatory variables are Nonresponse, Male and 0 – 24.
Model J1 is the standard (single-level) logistic model with only the response indicator as the explanatory variable, Model J2 is the single-level logistic model with additional household-level characteristics (explanatory variables), Model J3 is the two-level random intercept logistic model with only the response indicator as the explanatory variable, Model J4 is the two-level random coefficient logistic model with only the response indicator as the explanatory variable, Model J5 is the two-level random intercept logistic model controlling for other explanatory variables and Model J6 is the two-level random coefficient logistic model controlling for other explanatory variables. BDIC is the Bayesian Deviance Information Criterion.
** Significant at the 5% level; * Significant at the 10% level

Allowing the difference between respondents and nonrespondents within an interviewer to vary across interviewers (Model J4), the coefficient of the response indicator is still significant at the 5% level. This indicates that there is nonresponse bias. Also, the Wald test (test statistic = 7.613 on 2 degrees of freedom) for the inclusion of the variance of the random interviewer effects for the coefficient of the response indicator and for the covariance between intercept and coefficient random effects is significant (p–value = 0.022). This means that there is evidence that the nonresponse bias varies by interviewer for the probability of having at least one job in the household. On the other hand, this evidence is smaller comparing the BDIC from Model J4 with the one from Model J3.

Model J6 extends Model J4 to control for additional explanatory variables. It can be seen that controlling for other explanatory variables in the model the nonresponse bias is still significant at the 5% level, even though this nonresponse bias is partially explained by the inclusion of gender and age in the model. In addition, the interviewer random effects are no longer significant after controlling for gender and age, since the Wald test (test statistic = 2.672 on 2 degrees of freedom) for the inclusion of the two extra parameters (variance of u_{1j} and covariance between u_{0j} and u_{1j}) is not significant (p–value = 0.263).

In Model H1 from Table 3.8, the odds of owning a property (house or flat) are different between respondents and nonrespondents, since the coefficient of the response indicator is significantly different from zero. This means that there is nonresponse bias. As other explanatory variables are included in the model (Model H2) the nonresponse bias decreases.

Table 3.8: Parameter estimates for the models for type of housing

Fixed effect	Model H1	Model H2	Model H3	Model H4	Model H5	Model H6
Constant	0.182 (0.029) **	-0.934 (0.170) **	0.178 (0.031) **	0.179 (0.031) **	-0.955 (0.168) **	-0.965 (0.182) **
Response indicator						
Response	0.426 (0.035) **	0.199 (0.039) **	0.424 (0.035) **	0.420 (0.042) **	0.199 (0.038) **	0.195 (0.043) **
Gender						
Female		-0.303 (0.062) **			-0.303 (0.062) **	-0.302 (0.062) **
Mixed		0.979 (0.054) **			0.981 (0.054) **	0.982 (0.054) **
Age						
25 – 29		0.667 (0.183) **			0.684 (0.182) **	0.694 (0.191) **
30 – 34		0.962 (0.176) **			0.980 (0.174) **	0.992 (0.186) **
35 – 39		1.180 (0.174) **			1.198 (0.172) **	1.209 (0.184) **
40 – 44		1.228 (0.173) **			1.246 (0.172) **	1.258 (0.183) **
45 – 49		1.059 (0.172) **			1.075 (0.171) **	1.085 (0.182) **
50 – 54		1.083 (0.172) **			1.101 (0.171) **	1.112 (0.183) **
55 – 59		0.864 (0.170) **			0.883 (0.169) **	0.894 (0.181) **
60 – 64		0.821 (0.173) **			0.838 (0.171) **	0.849 (0.183) **
65 – 69		0.333 (0.173) *			0.351 (0.171) **	0.361 (0.182) *
70+		0.024 (0.166)			0.041 (0.165)	0.051 (0.177)
Random effect						
Random intercept:						
Interviewer variance $\sigma^2_{u_0}$			0.005 (0.003)	0.004 (0.003)	0.003 (0.002)	0.005 (0.004)
Random coefficient:						
Interviewer coef. variance $\sigma^2_{u_1}$				0.024 (0.008)		0.014 (0.009)
Interv. inter.-coef. covariance $\sigma_{u_0u_1}$				-0.001 (0.003)		-0.005 (0.005)
Interv. inter.-coef. correlation $\rho_{u_0u_1}$				-0.071		-0.580
BDIC	21493.44	19171.29	21491.15	21492.49	19172.90	19173.19

The base categories for the explanatory variables are Nonresponse, Male and 0 – 24.
Model H1 is the standard (single-level) logistic model with only the response indicator as the explanatory variable, Model H2 is the single-level logistic model with additional household-level characteristics (explanatory variables), Model H3 is the two-level random intercept logistic model with only the response indicator as the explanatory variable, Model H4 is the two-level random coefficient logistic model with only the response indicator as the explanatory variable, Model H5 is the two-level random intercept logistic model controlling for other explanatory variables and Model H6 is the two-level random coefficient logistic model controlling for other explanatory variables. BDIC is the Bayesian Deviance Information Criterion.
** Significant at the 5% level; * Significant at the 10% level

Model H3 extends Model H1 to allow the intercept to vary across interviewers. The nonresponse bias is significant at the 5% level. However, the Wald test (test statistic = 1.889 on 1 degree of freedom) for the inclusion of the interviewer–dependent variance is not significant (p–value = 0.169). This means that there is no evidence that the probability of owning a property varies across interviewers. This also means that there are no interviewer assignment effects on the probability of owning a property.

Comparing the BDICs from Models H1 and H3, there is further evidence that the inclusion of the interviewer–dependent variance is not significant since the reduction of the BDIC from Model H3 is not substantial, as shown in Table 3.8.

Allowing the difference between respondents and nonrespondents within an interviewer to vary across interviewers (Model H4), the nonresponse bias is significant at the 5% level. Additionally, the Wald test (test statistic = 9.165 on 2 degrees of freedom) for the inclusion of the variance of the random interviewer effects for the coefficient of the response indicator and for the covariance between intercept and coefficient random effects is significant (p–value = 0.010). This indicates that the nonresponse bias varies by interviewer for the probability of owning a property.

As for the dependent variable jobs in household, once gender and age are included in the model (Model H6), there is a reduction in the nonresponse bias. However, the Wald test (test statistic = 2.807 on 2 degrees of freedom) for the inclusion of the variance of the random interviewer effects for the coefficient of the response indicator and for the covariance between intercept and coefficient random effects is not significant (p–value = 0.246).

Models for the dependent variable size of household are also fitted. These models provide similar interpretation as the models in Tables 3.7 and 3.8. The parameter estimates for size of household are presented in Appendix D.

3.4.3 Descriptive analysis for the BHPS dataset

Another type of auxiliary variable is a time invariant variable from a longitudinal survey. In the BHPS dataset, variables from wave 1 are used to provide information for respondents and nonrespondents from wave 10. The interviewers' response rates for wave 10 of the BHPS are explored by inspecting a histogram. The pattern of the interviewers' response rates suggests a distribution skewed to the left, as is illustrated in Figure 3.2.

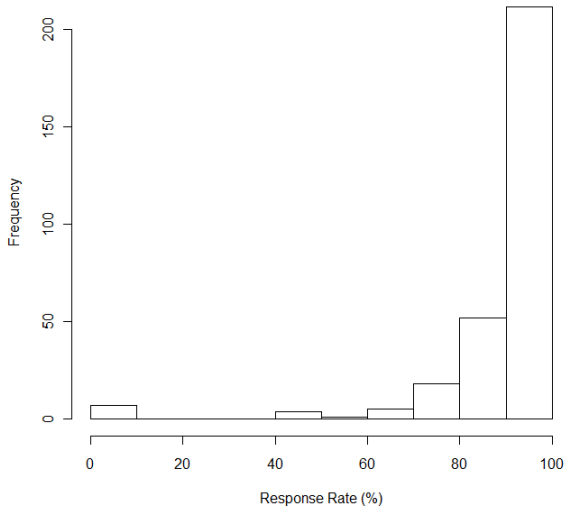


Figure 3.2: Histogram of the interviewers' response rates (BHPS dataset)

Considering the two dependent variables, it is found that having either one of the parents working when interviewees aged 14 does not differ whether the interviewee is male or female. Also, interviewees who respond to the survey and are white are more likely to have parents that were working when at age 14. Additionally, interviewees who are younger than 35 years old are more likely to have working mother at age 14, whereas interviewees who are 35 to 49 years old are more likely to have working father at age 14 (refer to Appendix C for the cross–tabulations).

3.4.4 Statistical modelling for the BHPS variables

Tables 3.9 and 3.10 present respectively the parameter estimates for the models for the dependent variables mother and father working when interviewee was 14 years old. In Model M1 from Table 3.9, since the coefficient of the response indicator is not significantly different from zero, the odds of mother working when interviewee was 14 for respondents is not significantly different from the odds for nonrespondents. In other words, there is no evidence of nonresponse bias. However, after controlling for age and ethnicity in the model (Model M2), the coefficient of the response indicator becomes significant at the 10% level. This can be explained by Simpson's paradox. The phenomenon occurs when a variable has no significant effect marginally, but the hidden effect becomes significant after conditioning on (or controlling for) other variables. Therefore, since the response indicator is significant at 10%, this indicates that there is some evidence that there is nonresponse bias, controlling for age and ethnicity in the model.

In Model M3 the intercept is allowed to vary across interviewers. Surprisingly, the Wald test (test statistic = 7.727 on 1 degree of freedom) for the inclusion of the interviewer–dependent variance is significant (p–value = 0.005), which means that there is evidence that the probability of mother working when interviewee was 14 varies across interviewers. These effects are referred to as interviewer assignment effects (see discussion that follows Equation (3.1)).

Model M4 extends Model M3 to allow the coefficient of the response indicator to vary across interviewers. The interviewer assignment effects are still significant. However, the Wald test (test statistic = 3.439 on 2 degrees of freedom) for the inclusion of the variance of the random interviewer effects for the coefficient of the response indicator and for the covariance between intercept and coefficient random effects is not significant (p–value = 0.179).

Table 3.9: Parameter estimates for the models for mother working when interviewee aged 14

Fixed effect	Model M1	Model M2	Model M3	Model M4	Model M5	Model M6
Constant	-0.284 (0.080) **	0.497 (0.102) **	-0.295 (0.084) **	-0.318 (0.105) **	0.499 (0.102) **	0.499 (0.119) **
Response indicator						
Response	0.090 (0.084)	0.173 (0.090) *	0.105 (0.087)	0.127 (0.107)	0.180 (0.090) **	0.188 (0.108) *
Age						
35 – 49		-0.305 (0.080) **			-0.308 (0.080) **	-0.316 (0.082) **
50+		-1.432 (0.076) **			-1.441 (0.077) **	-1.451 (0.079) **
Ethnicity						
Nonwhite		-0.675 (0.154) **			-0.678 (0.157) **	-0.678 (0.158) **
Random effect						
Random intercept:						
Interviewer variance $\sigma^2_{u_0}$			0.055 (0.020)	0.537 (0.262)	0.037 (0.018)	0.378 (0.171)
Random coefficient:						
Interviewer coef. variance $\sigma^2_{u_1}$				0.476 (0.269)		0.356 (0.168)
Interv. inter.-coef. covariance $\sigma_{u_0u_1}$				-0.481 (0.262)		-0.347 (0.165)
Interv. inter.-coef. correlation $\rho_{u_0u_1}$				-0.952		-0.946
BDIC	9078.99	8486.04	9054.75	9044.73	8474.81	8470.28

The base categories for the explanatory variables are Nonresponse, < 35 and White.
Model M1 is the standard (single-level) logistic model with only the response indicator as the explanatory variable, Model M2 is the single-level logistic model with additional individual-level characteristics (explanatory variables), Model M3 is the two-level random intercept logistic model with only the response indicator as the explanatory variable, Model M4 is the two-level random coefficient logistic model with only the response indicator as the explanatory variable, Model M5 is the two-level random intercept logistic model controlling for other explanatory variables and Model M6 is the two-level random coefficient logistic model controlling for other explanatory variables. BDIC is the Bayesian Deviance Information Criterion.
** Significant at the 5% level; * Significant at the 10% level

In Model M5, controlling for age and ethnicity, the nonresponse bias is significant at the 5% level. Although the interviewer assignment effects reduce, compared to Model M3, after controlling for age and ethnicity, they are still significant. The Wald test (test statistic = 4.314 on 1 degree of freedom) for the inclusion of the interviewer–dependent variance is significant (p–value = 0.038).

Even after allowing the coefficient of the response indicator to vary across interviewers (Model M6), the interviewer assignment effects are still highly significant. On the other hand, there is no evidence that the nonresponse bias varies by interviewer. This happens since the Wald test (test statistic = 4.526 on 2 degrees of freedom) for the inclusion of the variance of the random interviewer effects for the coefficient of the response indicator and for the covariance between intercept and coefficient random effects is not significant (p–value = 0.104). The small reduction on the BDIC from Model M6 compared to the one from Model M5 also indicates that the two extra parameters (variance of u_{1j} and covariance between u_{0j} and u_{1j}) are not significant. The BDICs are also presented in Table 3.9.

Since the BHPS is mostly a face–to–face survey, interviewer effects are potentially confounded with area effects. Therefore, since the interviewer assignment effect in Models M3 to M6 is significant, it is worth investigating if interviewer and area effects are entangled. To do this, the models in Table 3.10 include the cross–classification between interviewer and area random effects.

The area–dependent variance, in Models M7 to M9, is not significant. This might indicate that here the cross–classified models are not able to disentangle the effects of interviewers and areas. Therefore, further investigation may be needed to explain these significant interviewer assignment effects.

Table 3.10: Parameter estimates for the models for mother working when interviewee aged 14 (cross–classified models)

Fixed effect	Model M3	Model M7	Model M8	Model M9
Constant	-0.295 (0.084) **	-0.299 (0.084) **	-0.307 (0.102) **	0.496 (0.118) **
Response indicator				
Response	0.105 (0.087)	0.108 (0.086)	0.117 (0.105)	0.189 (0.108) *
Age				
35 – 49				-0.312 (0.081) **
50+				-1.450 (0.077) **
Ethnicity				
Nonwhite				-0.680 (0.158) **
Random effect				
Random intercept:				
Interviewer variance $\sigma^2_{u_0}$	0.055 (0.020)	0.033 (0.023)	0.381 (0.264)	0.388 (0.205)
Area variance $\sigma^2_{v_0}$		0.025 (0.020)	0.011 (0.011)	0.012 (0.012)
Random coefficient:				
Interviewer coef. variance $\sigma^2_{u_1}$			0.383 (0.256)	0.384 (0.211)
Interv. inter.-coef. covariance $\sigma_{u_0u_1}$			-0.348 (0.257)	-0.368 (0.205)
Interv. inter.-coef. correlation $\rho_{u_0u_1}$			-0.910	-0.954
BDIC	9054.75	9056.14	9048.36	8470.53

The base categories for the explanatory variables are Nonresponse, < 35 and White.
Model M3 is the two-level random intercept logistic model with only the response indicator as the explanatory variable, Model M7 is the cross-classified random intercept model with only the response indicator as the explanatory variable, Model M8 is the cross-classified random intercept model with random coefficient and only the response indicator as the explanatory variable and Model M9 is Model M8 plus other explanatory variables. BDIC is the Bayesian Deviance Information Criterion.
** Significant at the 5% level; * Significant at the 10% level

Considering the parameter estimates for the models in Table 3.11, the odds of father working when interviewee was 14 for respondents is significantly different from the odds for nonrespondents (Model F1). This happens because the coefficient of the response indicator is significantly (at the 5% level) different from zero. Including age and ethnicity in the model (Model F2) decreases the nonresponse bias.

Model F3 extends Model F1 to include a random intercept. The interviewer–dependent variance is significant (p–value = 0.024) based on the Wald test (test statistic = 5.102 on 1 degree of freedom). This means that there is evidence that the probability of father working when interviewee was 14 varies across interviewers. These are the so called interviewer assignment effects discussed in Section 3.3.

In Model F4 the coefficient of the response indicator is allowed to vary across interviewers. The Wald test (test statistic = 1.053 on 2 degree of freedom) for the inclusion of the variance of the random interviewer effects for the coefficient of the response indicator and for the covariance between intercept and coefficient random effects is not significant (p–value = 0.591). As might be expected, the BDIC from Model F4 compared to the one from Model F3 does not change, meaning that Model F3 should be preferred.

When other explanatory variables (age and ethnicity) are included in the model (Model F5), both the nonresponse bias and the interviewer assignment effects reduce a little. Therefore, it is recommended that the data analysis include individual–level characteristics to reduce the nonresponse bias and the possibility of interviewer assignment effects.

Table 3.11: Parameter estimates for the models for father working when interviewee aged 14

Fixed effect	Model F1	Model F2	Model F3	Model F4	Model F5	Model F6
Constant	2.098 (0.129) **	2.055 (0.157) **	2.141 (0.132) **	2.168 (0.148) **	2.120 (0.165) **	2.178 (0.179) **
Response indicator						
Response	0.272 (0.137) **	0.246 (0.137) *	0.285 (0.137) **	0.267 (0.154) *	0.247 (0.140) *	0.201 (0.159)
Age						
35 – 49		0.530 (0.140) **			0.524 (0.142) **	0.526 (0.143) **
50+		-0.071 (0.120)			-0.087 (0.122)	-0.087 (0.124)
Ethnicity						
Nonwhite		-0.932 (0.189) **			-0.942 (0.198) **	-0.944 (0.199) **
Random effect						
Random intercept:						
Interviewer variance $\sigma^2_{u_0}$			0.145 (0.064)	0.268 (0.181)	0.130 (0.064)	0.351 (0.222)
Random coefficient:						
Interviewer coef. variance $\sigma^2_{u_1}$				0.168 (0.165)		0.263 (0.199)
Interv. inter.-coef. covariance $\sigma_{u_0u_1}$				-0.137 (0.157)		-0.229 (0.194)
Interv. inter.-coef. correlation $\rho_{u_0u_1}$				-0.643		-0.752
BDIC	3927.54	3927.54	3927.54	3927.54	3927.54	3927.54

The base categories for the explanatory variables are Nonresponse, < 35 and White.
Model F1 is the standard (single-level) logistic model with only the response indicator as the explanatory variable, Model F2 is the single-level logistic model with additional individual-level characteristics (explanatory variables), Model F3 is the two-level random intercept logistic model with only the response indicator as the explanatory variable, Model F4 is the two-level random coefficient logistic model with only the response indicator as the explanatory variable, Model F5 is the two-level random intercept logistic model controlling for other explanatory variables and Model F6 is the two-level random coefficient logistic model controlling for other explanatory variables. BDIC is the Bayesian Deviance Information Criterion.
** Significant at the 5% level; * Significant at the 10% level

Table 3.12: Parameter estimates for the models for father working when interviewee aged 14 (cross–classified models)

Fixed effect	Model F3	Model F7	Model F8	Model F9	Model F10
Constant	2.141 (0.132) **	2.157 (0.134) **	2.128 (0.167) **	2.150 (0.141) **	2.170 (0.174) **
Response indicator					
Response	0.285 (0.137) **	0.287 (0.138) **	0.260 (0.141) *	0.304 (0.148) **	0.221 (0.155)
Age					
35 – 49			0.527 (0.145) **		0.528 (0.142) **
50+			-0.088 (0.124)		-0.089 (0.123)
Ethnicity					
Nonwhite			-0.932 (0.201) **		-0.937 (0.201) **
Random effect					
Random intercept:					
Interviewer variance $\sigma^2_{u_0}$	0.145 (0.064)	0.039 (0.052)	0.065 (0.068)	0.098 (0.139)	0.241 (0.158)
Area variance $\sigma^2_{v_0}$		0.150 (0.078)	0.115 (0.080)	0.095 (0.070)	0.081 (0.072)
Random coefficient:					
Interviewer coef. variance $\sigma^2_{u_1}$				0.129 (0.136)	0.227 (0.151)
Interv. inter.-coef. covariance $\sigma_{u_0u_1}$				-0.055 (0.131)	-0.181 (0.143)
Interv. inter.-coef. correlation $\rho_{u_0u_1}$				-0.488	-0.774
BDIC	3909.60	3902.08	3859.94	3905.01	3861.58

The base categories for the explanatory variables are Nonresponse, < 35 and White.
Model F3 is the two-level random intercept logistic model with only the response indicator as the explanatory variable, Model F7 is the cross-classified random intercept model with only the response indicator as the explanatory variable, Model F8 is Model F7 plus other explanatory variables, Model F9 is the cross-classified random intercept model with random coefficient and only the response indicator as the explanatory variable and Model F10 is Model F9 plus other explanatory variables. BDIC is the Bayesian Deviance Information Criterion.
** Significant at the 5% level; * Significant at the 10% level

As for the dependent variable mother working, cross–classified models are also fitted for father working to check if interviewer effects are confounded with area effects. The parameter estimates for the cross–classified models for the dependent variable father working are presented in Table 3.12.

Interestingly, in Model F7, according to the Wald test (test statistic = 3.713 on 1 degree of freedom), the area–dependent variance is significant (p–value = 0.053), whereas the interviewer–dependent variance is not (p–value = 0.449). This means that the assignment effects are mostly due to areas rather than to interviewers. However, after controlling for age and ethnicity (Model F8), the area–dependent variance is no longer significant.

3.5 Conclusions

In this chapter, some of the multilevel logistic models from Chapter 2 were used to investigate further the assessment of interviewer effects on nonresponse bias. These models were applied to data from a telephone (CATI) cross–sectional survey (CCS) and a face–to–face longitudinal survey (BHPS). Also, differently from Chapter 2, rather than census records, administrative data and time invariant variables from previous waves were linked to survey data respectively for the CCS and the BHPS in order to provide variables for respondents and nonrespondents. The main contribution of this chapter is to apply the methodology proposed in Chapter 2 to assess interviewer effects on nonresponse bias, when census linked data are not available and in the context of a telephone survey and a longitudinal survey.

In the models for the dependent variable jobs in the household, the significant coefficient of the response indicator in Model J1 indicates that there is nonresponse bias for the odds of having at least one job in the household. The nonresponse

bias is partially explained when other explanatory variables are included in the model (Model J2).

In contrast to the findings in Chapter 2, another interesting finding here is that there are no interviewer assignment effects for the odds of having at least one job in the household (Model J3). This effect results from differential assignment of interviewers with respect to the variable jobs in the household. Extending the model to include random coefficients (Model J4) the nonresponse bias varies by interviewer, i.e., there is evidence of interviewer effects. However, the interviewer effects are explained by controlling for gender and age (Model J6). The nonresponse bias seems rather pronounced, especially, in the Model J1. This agrees with the nonresponse literature that suggests that the nonresponse bias is larger for the telephone mode compared to the face–to–face one (Biemer, 2001).

The findings for the other two dependent variables from the CCS linked dataset, type of housing and size of household, are similar to the ones for jobs in the household.

Although the bias is still significant, it is clear that part of the nonresponse bias is explained by controlling for gender and age for all three dependent variables from the CCS linked dataset. Therefore, one should include in the data analysis the explanatory variables gender and age to try to reduce the nonresponse bias.

In the models for the dependent variable mother working indicator, there is no nonresponse bias for the odds of having mother working when interviewee was 14 at first (Model M1). Then, including other explanatory variables in the model (Model M2), the nonresponse bias is borderline significant. Additionally, an undesirable feature in the models (Models M3 to M6) is the presence of significant interviewer assignment effects.

The cross–classified models for this dependent variable did not provide an explanation for the significant interviewer assignment effects. Therefore, these models need deeper investigation.

The findings for the dependent variable father working indicator are quite similar to those for mother working indicator, except that the cross–classified models provided evidence that the assignment effects are due to areas rather than to interviewers (Model F7). Also the nonresponse bias is partially explained by including age and ethnicity in the model (Model F2).

The conclusions of the analyses in this chapter may be weakened due to the linkage of household–level instead of individual–level variables with the CCS data. Furthermore, the comparison between wave 1 and wave 10 of the BHPS may have implications for the analyses of interviewer effects. For instance, it is likely that some of those identified as nonresponders at wave 10 were also nonresponders at earlier waves. If they were assigned different interviewers at earlier waves, i.e., if there was not interviewer continuity (Campanelli and O'Muircheartaigh, 1999), then the cause of nonresponse would actually be these earlier interviewers, and the interviewer effects may be confounded with earlier interviewers.

Another limitation for the results in this chapter is because a number of respondent and nonrespondent cases were deleted from the BHPS dataset in the process of merging data from wave 1 with data from wave 10 as well as cases without interviewer ID. Thus, the way the analyses are conducted, based on a dataset having a 9.5% nonresponse rate, may be seen as a relatively straightforward approach to detect nonresponse bias. However, if the deleted cases were indeed selective cases, in the sense that they do have different characteristics than those of the respondents, then the analyses for BHPS variables may not be correct. In this case, a more complex approach, such as by imputing values for the cases that had been deleted to take into account their missing information, would be advisable to potentially

address the assessment of nonresponse bias. Such an approach, however, has the drawback of requiring more assumptions and models to be safely employed.

The importance of the findings of this thesis regarding the assessment of interviewer effects on nonresponse bias is that these effects are investigated under different survey conditions. Although the interviewer effects observed across the three different surveys in the analyses of nonresponse bias cannot be compared directly since these surveys are from different populations and survey topics, some of the findings in this research are consistent for a specific survey mode. The analyses undertaken provide therefore a type of sensitivity analysis for the results. For instance, for the two face–to–face surveys (LFS and BHPS) there are initially significant interviewer assignment effects, whereas for the CATI survey (CCS) these effects are not significant at all. One plausible explanation is that in the CATI survey interviewers are randomly assigned to sampled units, i.e., the survey design is approximately interpenetrated (O'Muircheartaigh and Campanelli, 1999). On the other hand, other findings are common across survey modes. For example, there is evidence of nonresponse bias and of interviewer effects on nonresponse bias for a dependent variable from the LFS linked dataset and from the dependent variables from the CCS linked dataset.

In the context of interviewer effects on nonresponse bias, it would be interesting for future work to apply multilevel multinomial model to variables of interest with more than two categories and also consider response indicators with more than two categories such as response, refusal and noncontact. In addition, interviewers characteristics could also be controlled in the models to check if the interviewer–level variation is explained by this inclusion. Furthermore, the inclusion of household effects in the models could be worthy pursuing since nonresponse seems to be largely a household feature rather than a property of individuals.

Chapter 4

Interviewer Effects on Measurement Error

4.1 Introduction

Measurement error is a common feature in nearly all surveys. It occurs when there is a discrepancy between an observed measure and its true value. This type of non-sampling error is undesirable since it harms the quality of the data to the extent of potentially biasing survey estimates and invalidating inferences. Measurement error mostly arises from four sources: interviewer, respondent, questionnaire and the mode of data collection (Groves, 2004). Examples of possible causes for this error can include, but are not limited to: differences that may occur in reactions of respondents to different interviewers (Ruddock, 1998), e.g., to interviewers of their own sex or own ethnic group; answers given by respondents may be influenced by their desire to impress an interviewer or their answers do not always reflect their true beliefs because they may feel under social pressure not to give an unpopular or socially undesirable answer, e.g., the use of illicit drugs may be underreported

(Mensch and Kandel, 1988); inadequate interviewer training; the wording of questions may be unclear, ambiguous or difficult to answer, e.g., it may require remembering past dates or facts; and respondents may answer questions differently depending on the manner in which the questionnaire is administered, whether in the presence of an interviewer, via telephone or self–administered (Lepkowski, 2004). For instance, in an internet survey on self–reported sexual behaviour, men reported significantly higher sociosexuality than women. However, there was no difference between men and women's reports when the concern about confidentiality was not compromised (Beaussart and Kaufman, 2013). There is an extensive literature on measurement error in surveys, for instance see Groves (2004), Biemer et al. (1991), Biemer and Lyberg (2003) and (Fuller, 1987) for earlier references of adjustments for measurement error.

Measurement error is generally associated with the data collection mode and reasons for the occurrence of such an error can be manyfold. For instance, telephone surveys are known to collect less accurate information than face–to–face surveys (Biemer and Lyberg, 2003, p. 42). Also, if the survey is conducted in the presence of other people, respondents may feel uncomfortable and misreport some answer, especially to sensitive questions. In another piece of research, Biemer (2001) compares the quality of data from a Computer Assisted Telephone Interviewing (CATI) and a face–to–face modes. He finds that the nonresponse bias is larger for the CATI compared to the face–to–face mode whereas the measurement bias is more pronounced for the face–to–face compared to the CATI mode.

Additionally, measurement error can arise from the interaction between interviewer and respondents. In such case, respondents may provide inaccurate answers, survey questionnaires may contain ambiguous questions and interviewers may misrecord some answers (Biemer et al., 1991, p. 2). Furthermore, survey interviewers can have an influence on measurement error when their personalities or attitudes

may induce respondents as, for example, to give socially desirable answers that may differ from the truth. Regardless of interviewer effects, Tourangeau et al. (2010) describe a study that involves sensitive topics and socially desirable behaviours where they expect that reluctant respondents are more likely to provide inaccurate answers related to these topics if they eventually become respondents. In the end Tourangeau and colleagues confirm their expectation, finding a strong link between nonresponse errors and reporting errors.

Amongst the sources of measurement errors, in interviewer–administered surveys, interviewers and questionnaires are factors that survey organizations may have some control. Therefore, improving interviewer training and questionnaire design increases the chance of gathering data as accurate as possible. Gideon (2012, p. 46) emphasizes that the efforts to improve questionnaire design and interviewer training aim mostly to minimize the measurement error in surveys. In his book, Gideon reports two standardized approaches for the interviewing protocol: a fully standardized and a flexible one. In the former, interviewers are recognized as a source of error since they may interpret questions differently. Hence, in this approach, the interviewers are not allowed to change words or terms in the questions to make sure that their intended meanings are preserved. In the latter, the respondents are seen as a source of error since they may misinterpret questions. In this case, the interviewers may intervene to provide clarification to gather accurate data. The choice between one of these approaches depends on the researcher's judgement of which one is more suitable for a specific survey.

In the past, some studies on measurement error have focussed on the evaluation of interaction effects between the gender of the interviewer and the gender of the respondent on several gender attitudes towards public policies, perceptions of men's and women's group interests (Kane and Macaulay, 1993) and on gender–sensitive items such as abortion and women's rights (Flores-Macias and Lawson,

2008). Other studies have focussed on race of interviewer effects, for instance, on political and racial attitudes (Davis, 1997) and on questions involving racial issues (Reese et al., 1986) in a telephone survey, where the respondent's race is asked in the end of the questionnaire (Cotter et al., 1982).

Recent research uses interviewers' observations to assess the quality of survey data, especially aiming to detect the presence of measurement error (Olson and Parkhurst, 2013). Sinibaldi et al. (2013) examine measurement error in interviewers' observations from six UK face–to–face surveys. In their analysis, they used census records to assess whether or not the interviewers' observations were accurate. As one of their main findings, the results suggested that interviewers significantly influenced measurement error.

Kreuter et al. (2010) question the quality of data, from reluctant respondents, obtained through an extra effort from interviewers as an attempt to reduce nonresponse. Since this practice involves persuasion, these observations may not be free from measurement error (Olson, 2006) and in fact they may lead to poorer quality data (Fricker and Tourangeau, 2010; Biemer, 2001). In their study, in addition to survey data from respondents, they also had administrative data for all sampled units. However, rather than linking survey data to administrative data, they linked administrative data to contact protocols. Then, they assess the measurement error from the difference between the estimates based on survey data and the estimates from the linked data.

In contrast, some studies have no auxiliary information from re–interviews, census records or administrative data to investigate measurement error in survey estimates. For example, Dixon (2010) assesses nonresponse and measurement error in a statistical matching framework by linking data from different surveys. Potential concerns in this approach are differences in sample composition and in question wording. Although progress on detecting measurement error have been

accomplished, the type of dataset used to assess measurement error on variables of interest ideally is one that contains the observed and the true values for these variables for all respondents.

This research aims to investigate the possibility of interviewers influence on measurement error in variables of interest. Assuming that the structure of the data is hierarchical, i.e. respondents are nested within interviewers, multilevel models are applied to a dataset consisting of data from the 2010 Norwegian sample of the European Social Survey linked to administrative data. The linked data contain the same type of variables of interest from the survey and from administrative data for all the respondents. A dataset with this feature is quite rare since survey practitioners try to avoid overburden sampled units, and therefore do not ask a question in a survey that is already captured by other sources, such as administrative or register data. Even though the measurement error literature is extensive, the dataset used in the application of the proposed models has the advantage of containing both observed and assumed true measures for variables of interest.

In this study, it is assumed that variables from a register represent the true measures, whereas variables from the survey are regarded as approximations for the true measures and are measured with measurement error. Similar assumptions have been made by Sinibaldi et al. (2013) and Kreuter et al. (2010). In the data analysis, the dependent variables are categorical measurement error indicators created based on the joint distribution from the pairs of observed and true values for essentially the same variables. Although the literature refers to the existing measurement error between the true and the observed measures, when both are categorical, as misclassification (Buonaccorsi, 2010, p. 1), the term measurement error will be used throughout this chapter.

The structure of this chapter is as follows. Section 4.2 provides a detailed description about the survey, the dataset and the variables considered in this study.

Section 4.3 defines the statistical models used in the data analysis, modelling strategy and estimation method. In Section 4.4, the main results are discussed and lastly Section 4.5 presents the conclusions, limitations of the study, implications for survey practice and possible future work.

4.2 Data

This research makes use of a rich dataset acquired by linking two data sources: a survey and administrative data. The key feature of this dataset is that it contains variables of interest collected in the survey that are also present in the administrative data, allowing for measurement error analysis. The following sections provide more information about the survey, the administrative data and the analysis sample.

4.2.1 Survey

The first data source is taken from the European Social Survey (ESS) (for more details on the ESS see: `http://www.europeansocialsurvey.org`). The ESS has been carried out every two years since 2002. The most recent round of this survey was in 2012 and around 30 countries took part in each round. The ESS was mainly initiated to provide a reliable source of cross–national data collected with high methodological standards to enable inferences to be drawn about changes over time in Europe. Thus, this survey aims to explain changes in attitude and behaviour patterns in Europe and improve methods of survey measurements across countries.

The ESS collects information on media use, social and public trust, political interest and participation, socio–political orientations, moral, political and social

values, national, ethnic and religious allegiances, well–being, health and security, demographics and socio–economics.

In this study, the analysis is restricted to the 2010 Norwegian sample (round 5) of the ESS (ESS–Norway). The ESS–Norway was conducted by Statistics Norway. The sample is stratified by three variables: age group, gender and region. Within each stratum, one–stage systematic random sampling is used such that each individual in the sampling frame has equal probability of being selected (EPSEM). The fieldwork for this survey took place between September 2010 and February 2011. To increase the response rate, before the actual interviews, the sampled units received a letter explaining the purpose of the survey and also unconditional monetary incentives. The ESS–Norway is a face–to–face survey whose data were collected using Computer Assisted Personal Interviewing (CAPI) by 110 experienced interviewers. These interviewers had received training on written ESS specific instructions, refusal conversion and how to fill in contact forms.

4.2.2 Administrative data

Administrative data are records, registers and databases collected for administrative purposes mainly to be used by government departments. Although their primary usage is not for statistical research, administrative data have increasingly been used for this purpose. Statistical offices usually have access to these data, which are statistically treated (e.g., clean) for various objectives that range from the release of summary statistics to advanced data analysis. Therefore, government organizations and the general public benefit from the work these offices do with these administrative data.

Statistics Norway receives different extracts of administrative data from various government sources. The variables from administrative data used in this study are

from the Central Population Register, Education Department and from Statistics Norway archive. Some of administrative data are updated continuously and some are updated from time to time. In order to link the ESS–Norway data to administrative data, these data were updated around the time of the ESS–Norway data collection takes place. This makes the information from administrative data close to the true measure.

Before the fieldwork for the survey takes place, Statistics Norway sends a cover letter explaining the purpose of the survey to the sampled units. In this letter there is a statement about the possibility of linking the survey data to specific sources of administrative data. Therefore, the sampled units who decide to take part in the survey also agree to have their data linked.

4.2.3 Linked dataset and analysis sample

For the application of the proposed model to investigate interviewer effects on measurement error, the analysis sample consists of data from the ESS–Norway linked to administrative data (register). These two sources of data are linked through a personal identifier by Statistics Norway. The linkage of the ESS–Norway and register data provided a dataset containing pairs of essentially the same variables of interest for each respondent. It is assumed that variables from the register represent the true measures, whereas variables from ESS–Norway are regarded as approximations for the true measures.

Two pairs of variables of interest are considered in this analysis: highest level of education (education) and number of people living regularly as members of a household (household size). These variables are the only ones that are available in both survey and register. There is no missing data for the variables used in this analysis that come from the register. However, the variable education from the

ESS–Norway has missing values for 6 cases. Therefore, since only the complete cases are considered in this analysis, these 6 cases were deleted. As a result, the analysis sample is left with information on a total of 1,540 individuals interviewed by 106 interviewers.

The variables education and household size from ESS–Norway and register are categorical with three categories each. Tables 4.1 and 4.2 present cross–tabulations respectively for the pair of variables education and the pair of variables household size. In these tables the misclassification (measurement error) is found in the off diagonal cells, where counts in these cells are different from zero.

According to Tables 4.1 and 4.2, the majority of people report in the survey the same level of education and the same household size as in the register. However, for both pairs of variables, people tend to over report their level of education and their household composition more often than they under report them.

Table 4.1: Cross–tabulation for education from ESS–Norway and from the register

Education from	Education from register			
ESS–Norway	Low	Middle	High	Total
Low	233	70	0	303
Middle	170	438	12	620
High	49	100	468	617
Total	452	608	480	1540

Table 4.2: Cross–tabulation for household size from ESS–Norway and from the register

Household size	Household size from register			
from ESS–Norway	1 person	2 people	3+ people	Total
1 person	227	20	25	272
2 people	132	367	73	572
3+ people	118	50	528	696
Total	477	437	626	1540

In this research, the dependent variables are measurement error indicators. These indicators are created based on the joint distribution of education (or household size) from the ESS–Norway and education (or household size) from the register. Hence, if there is an agreement between the level of education (or in the number of people in the household) from the survey and register, the measurement error indicator assumes the value 0. Alternatively, it assumes the value 1 if there is measurement error. The next section gives more details on the definition of these measurement error indicators.

The histograms in Figures 4.1 and 4.2 illustrate the distributions of measurement error per interviewer respectively for reporting education and household size. Inspecting the histograms, both distributions seems to be skewed to the right, which means that the percentage of measurement error per interviewer is in general small, where the spike in zero illustrates the percentage of no measurement error per interviewer. However, the percentage of measurement error is quite high for some interviewers. As a whole, the percentage of measurement error per interviewer varies from 0 to about 67% for reporting education and from 0 to 100% for reporting household size. However, the 67% comes from one interviewer who interviewed only three individuals, out of which two were misclassified, whereas the 100% comes from two interviewers: one interviewed only one individual and

the other one only two individuals; the three of them were misclassified. Without these extreme cases, the variation of measurement error per interviewer for reporting education is between 0 and 50%, whereas for reporting household size is between 0 and about 57%.

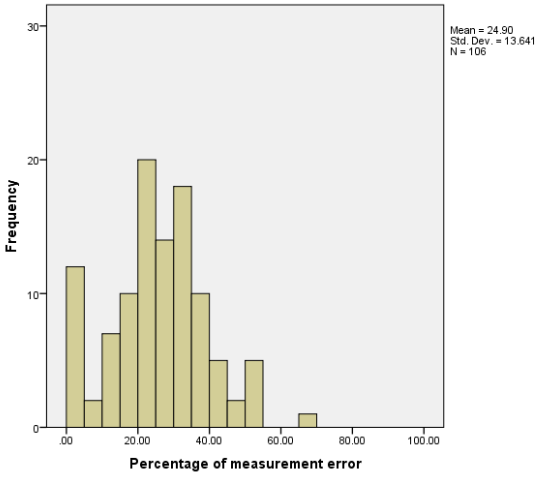


Figure 4.1: Histogram of the percentage of measurement error (for education) by interviewer

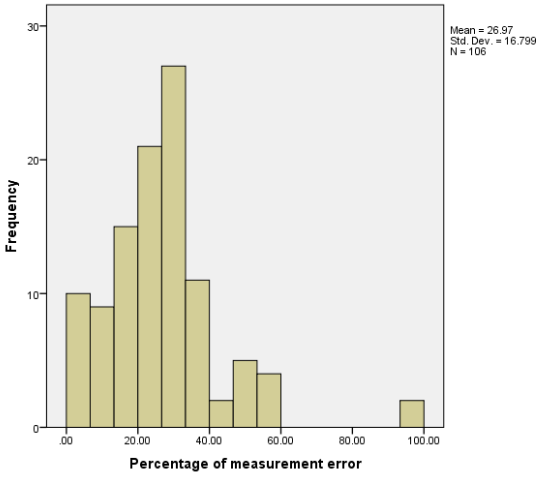


Figure 4.2: Histogram of the percentage of measurement error (for household size) by interviewer

For the ESS–Norway, interviewers are not randomly assigned to sampled units (i.e., it is not an interpenetrated assignment). Thus, interviewer effects are potentially confounded with area effects. Tables 4.3 and 4.4 show the frequency distribution of interviewers per area and the frequency distribution of the number of areas covered by interviewers, respectively. According to the frequency distributions in these tables, there is a non-nested structure between interviewers and areas.

Table 4.3: Frequency distribution of number of interviewers per areas

No. of interviewers per area	No. of areas
14	2
15	2
31	2
32	1
Total	7

Table 4.4: Frequency distribution of number of areas per interviewer

No. of areas per interviewer	No. of interviewers
1	72
2	27
3	4
4	1
5	2
Total	106

Since interviewers are not strictly nested within areas, one could apply multilevel cross–classified models as an attempt to separate interviewer and area effects. However, the ESS–Norway dataset does not contain a lower geographical unit area variable such as district or postcode sectors, which could be used as one of the level–two variables in the multilevel cross–classified models. Thus, the area variable is included in the models as a fixed effect. The area variable available in this dataset has only 7 categories (regions), which are Oslo and Akershus, Hedmark

and Oppland, South Eastern, Agder and Rogaland, Western Norway, Troendelag and Northern Norway (please refer to Appendix E to visualize Norway's regions). Therefore, in addition to this area variable, gender and age (categorized into 3 groups: 15 to 30, 31 to 66 and over 66 years old) of the respondents are included into the models as explanatory variables to control for the nonrandomized allocation of interviewers to sampled units. These explanatory variables are all taken from the register since they are not asked in the survey. Frequency distributions for these variables are displayed in Appendix F.

4.3 Methodology

Since the aim of this research is to investigate the effects of interviewers on measurement error, multilevel modelling techniques (Goldstein, 2011) are used in the analysis of the data. These types of models are widely applied in the literature to account for interviewer effects. Generally, interviewers can be seen as clusters, since the responses from individuals interviewed by the same interviewer are more alike than those from individuals interviewed by different interviewers (Biemer et al., 1991, p. 440). For the analysis, the variables of interest are those that have been collected in both the ESS–Norway and in the register. In the case of the ESS–Norway and register linked data these variables are education level and household composition.

Before presenting the statistical model used in the analysis of the data, the following notation is introduced: let $\widetilde{y}_{ij}$ be the variable of interest from the register for individual i interviewed by interviewer j, where $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$, and y^*_{ij} be the same variable of interest from the survey for the same individual and interviewer. It is assumed that $\widetilde{y}_{ij}$ is the true value, whereas y^*_{ij} is a proxy for $\widetilde{y}_{ij}$. Consider also that $\widetilde{y}_{ij}$ and y^*_{ij} are categorical variables, with three categories

each, such that $\widetilde{y}_{ij}$ and $y^*_{ij} \in \{1, 2, 3\}$. Their joint distribution is presented in Table 4.5.

Table 4.5: Joint distribution for $\widetilde{y}_{ij}$ and y^*_{ij}

	$\widetilde{y}_{ij}$		
y^*_{ij}	1	2	3
1	$\pi_{ij}^{(1,1)}$	$\pi_{ij}^{(1,2)}$	$\pi_{ij}^{(1,3)}$
2	$\pi_{ij}^{(2,1)}$	$\pi_{ij}^{(2,2)}$	$\pi_{ij}^{(2,3)}$
3	$\pi_{ij}^{(3,1)}$	$\pi_{ij}^{(3,2)}$	$\pi_{ij}^{(3,3)}$

The joint probabilities $\pi_{ij}^{(s,t)}$ from Table 4.5 are used to create the measurement error indicators. Thus, considering first the case where the measurement error indicator y_{ij} is binary, let

$$y_{ij} = \begin{cases} 0, & \text{if no measurement error is present } (y^*_{ij} = \widetilde{y}_{ij}) \\ 1, & \text{if measurement error is present } (y^*_{ij} \neq \widetilde{y}_{ij}). \end{cases}$$

To investigate interviewer effects on measurement error, the two–level random intercept logistic (empty) model may be applied

$$\begin{aligned} y_{ij} \mid \boldsymbol{u}_0 \ &\sim \ \text{indep Bernoulli}(\pi_{ij}), \quad i = 1, \ldots, n_j, j = 1, \ldots, J \ \text{ and} \\ &\qquad\qquad \boldsymbol{u}_0 = (u_{01}, ..., u_{0J})^\top, \qquad (4.1) \\ \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) &= \beta_0 + u_{0j}, \end{aligned}$$

where $\pi_{ij} = P(y_{ij} = 1 \mid \boldsymbol{u}_0) = P(y^*_{ij} \neq \widetilde{y}_{ij} \mid \boldsymbol{u}_0) = \sum_{s \neq t} \pi_{ij}^{(s,t)}$ is the probability of having measurement error, β_0 is the log–odds of having measurement error when the random effects are set to their zero mean and u_{0j} are random effects representing unexplained interviewer effects. These random effects are assumed to

follow a normal distribution, i.e. $u_{0j} \sim N\left(0, \sigma^2_{u_0}\right)$, where the variance parameter $\sigma^2_{u_0}$ is the variation in the intercept across interviewers.

Now, to investigate if there are effects of individual–level characteristics (explanatory variables), $\boldsymbol{x}_{ij}$, on measurement error, the model in (4.1) may be extended to

$$\begin{aligned} y_{ij} \mid \boldsymbol{x}_{ij}, \boldsymbol{u}_0 \ &\sim \ \text{indep Bernoulli}(\pi_{ij}), \quad i = 1, \ldots, n_j, j = 1, \ldots, J \ \text{ and} \\ &\qquad\qquad\qquad\qquad \boldsymbol{u}_0 = (u_{01}, ..., u_{0J})^\top, \qquad (4.2) \\ \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) &= \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}_{ij} + u_{0j}, \end{aligned}$$

where $\boldsymbol{\beta}$ is a vector of coefficients for the explanatory variables, indicating the effects of explanatory variables on measurement error. The other quantities are as in Equation (4.1).

Furthermore, the measurement error indicator may be refined. For example, one can define the measurement error indicator y_{ij} with three categories, indicating not only that there is measurement error but also the type of measurement error, such as over or under reports based on Table 4.5. Thus let

$$y_{ij} = \begin{cases} 1, & \text{if no measurement error is present } (y^*_{ij} = \widetilde{y}_{ij}) \\ 2, & \text{if the survey value is greater than the register value } (y^*_{ij} > \widetilde{y}_{ij}) \\ 3, & \text{if the survey value is smaller than the register value } (y^*_{ij} < \widetilde{y}_{ij}). \end{cases}$$

In this case, $y_{ij} = 1$ indicates no measurement error, whereas $y_{ij} = 2$ indicates that there is measurement error, with y^*_{ij} over reporting $\widetilde{y}_{ij}$, and $y_{ij} = 3$ also indicates the presence of measurement error, with y^*_{ij} under reporting $\widetilde{y}_{ij}$. Now, since y_{ij} has more than two categories, a two–level multinomial model may be applied. The

model may be defined using the systematic component

$$\log\left(\frac{\pi_{ij}^{(c)}}{\pi_{ij}^{(1)}}\right) = \beta_0^{(c)} + \boldsymbol{\beta}^{(c)\top}\boldsymbol{x}_{ij} + u_{0j}^{(c)}, \tag{4.3}$$

where $c = 2, 3$, $\pi_{ij}^{(1)} = P(y_{ij} = 1) = P(y_{ij}^* = \widetilde{y}_{ij}) = \sum_{s=t} \pi_{ij}^{(s,t)}$ is the baseline probability, $\pi_{ij}^{(2)} = P(y_{ij} = 2) = P(y_{ij}^* > \widetilde{y}_{ij}) = \sum_{s>t} \pi_{ij}^{(s,t)}$ and $\pi_{ij}^{(3)} = P(y_{ij} = 3) = P(y_{ij}^* < \widetilde{y}_{ij}) = \sum_{s<t} \pi_{ij}^{(s,t)}$. As before the vector of random interviewer effects $(u_{0j}^{(2)}, u_{0j}^{(3)})^\top$, $j = 1, \ldots, J$, are independent and identically distributed with normal distribution having zero mean vector and interviewer–level covariance matrix in the form

$$\Omega_u = \begin{bmatrix} \sigma^2_{u_0^{(2)}} & \sigma_{u_0^{(2)}u_0^{(3)}} \\ \sigma_{u_0^{(2)}u_0^{(3)}} & \sigma^2_{u_0^{(3)}} \end{bmatrix},$$

where the parameters $\sigma^2_{u_0^{(2)}}$ and $\sigma^2_{u_0^{(3)}}$ are respectively the random intercept variance for over reporting and the random intercept variance for under reporting. The term $\sigma_{u_0^{(2)}u_0^{(3)}}$ is the covariance between random intercept effects for over and under reporting.

Allowing y_{ij} to have even more than three categories, one can define the four category measurement error indicator for education, for example, as

$$y_{ij} = \begin{cases} 1, & \text{if there is agreement between survey and register values on lower education} \\ 2, & \text{if there is agreement between survey and register values on middle education} \\ 3, & \text{if there is agreement between survey and register values on high education} \\ 4, & \text{if there is measurement error.} \end{cases}$$

Equivalently, the five category measurement error indicator for education can be defined as

$$y_{ij} = \begin{cases} 1, & \text{if there is agreement between survey and register values on lower education} \\ 2, & \text{if there is agreement between survey and register values on middle education} \\ 3, & \text{if there is agreement between survey and register values on high education} \\ 4, & \text{if the survey value is greater than the register value} \\ 5, & \text{if the survey value is smaller than the register value.} \end{cases}$$

Similar definitions can also be applied for the measurement error indicators for household size. The two–level multinomial model in Equation (4.3) can be easily extended to handle measurement error indicators with four or more categories. In the case of four categories, $\pi_{ij}^{(1)} = P(y_{ij} = 1) = \pi_{ij}^{(1,1)}$ is the baseline probability, $\pi_{ij}^{(2)} = P(y_{ij} = 2) = \pi_{ij}^{(2,2)}$, $\pi_{ij}^{(3)} = P(y_{ij} = 3) = \pi_{ij}^{(3,3)}$ and $\pi_{ij}^{(4)} = P(y_{ij} = 4) = \sum_{s \neq t} \pi_{ij}^{(s,t)}$. Whilst for five categories, again $\pi_{ij}^{(1)} = P(y_{ij} = 1) = \pi_{ij}^{(1,1)}$ is the baseline probability, $\pi_{ij}^{(2)} = P(y_{ij} = 2) = \pi_{ij}^{(2,2)}$, $\pi_{ij}^{(3)} = P(y_{ij} = 3) = \pi_{ij}^{(3,3)}$. However, $\pi_{ij}^{(4)} = P(y_{ij} = 4) = \sum_{s>t} \pi_{ij}^{(s,t)}$ and $\pi_{ij}^{(5)} = P(y_{ij} = 5) = \sum_{s<t} \pi_{ij}^{(s,t)}$.

As interviewer effects are potentially confounded with area effects, the models should control for area effects to try distinguishing these confounding effects from the ones driven by interviewers. Since interviewers are not strictly nested within areas as shown in Tables 4.3 and 4.4 from the previous section, ideally a cross–classified multilevel model (which has been discussed in previous chapters) where individuals are nested within a cross–classification between interviewers and areas should be applied. However, the available area variable has only seven categories

(regions), which is not disaggregated enough to be used as a random effect. Therefore, the area effects are included into the multilevel models as fixed rather than random effects.

4.3.1 Modelling strategy and estimation

For the investigation of interviewer effects on measurement error, the measurement error indicators used as dependent variables in the models are created based on the joint distribution for education level (household composition) from ESS–Norway and education level (household composition) from the register.

The analysis starts by considering models for the binary dependent variables. In this case, the measurement error indicators are coded as 1 if there is measurement error and as 0 if there is no measurement error. Firstly, the empty two–level random intercept logistic model is fitted to the binary measurement error indicators. If the interviewer–level variance is significant, this is not necessarily interpreted as interviewer effects on measurement error since the models do not control for other factors that may be confounded with interviewer effects, such as area effects. Secondly, individual–level characteristics such as gender, age and area are entered into the models as a way to control for the fact that interviewers were not randomly assigned to sampled units. In particular, this aims to control for social and geographical differences in the sample that may be confounded with interviewer effects. Then, if after controlling for these individual–level characteristics, the interviewer–level variance is still significant, this may be interpreted as an approximation of a "true" interviewer effect.

The procedure described above can be also applied to measurement error indicators with more than two categories.

The two–level random intercept logistic models and the two–level random intercept multinomial models are estimated using Markov Chain Monte Carlo (MCMC) methodology with 80,000 iterations after a burn–in of 5,000. The models are fitted using `MLwiN` (Rasbash et al., 2012).

4.4 Results

Based on the cross–tabulation for the variables education (or household size) from survey and education (or household size) from register presented in Table 4.1 (and 4.2) from Section 4.2, for the cases with measurement error, people tend to over report their education level (or the number of people in the household) more often than under report it. This can be explained by the fact that people tend to give socially desirable answers that may differ from the truth. In this context, interviewers may aggravate this propensity from respondents.

4.4.1 Modelling the binary measurement error indicator

Tables 4.6 and 4.7 show the distributions of the binary measurement error indicators for education and household size, respectively.

Table 4.6: Coding for the cross–tabulation for education from ESS–Norway and from register with two categories

Education from	Education from register		
survey	Low	Middle	High
Low	0	1	1
Middle	1	0	1
High	1	1	0

Table 4.7: Coding for the cross–tabulation for household size from ESS–Norway and from register with two categories

Household size	Household size from register		
from survey	1 person	2 people	3+ people
1 person	0	1	1
2 people	1	0	1
3+ people	1	1	0

Fitting two–level random intercept logistic (empty) models to the binary measurement error indicators for the variables education and household size, the variance estimate (in Table 4.8) is not significant which indicates that, for both variables, the measurement error does not seem to vary by interviewer. However, if it was significant at this point, these level–two effects could be due not only to interviewers but also to other confounding effects. Therefore, the models in Table 4.9 are controlling for social and geographical characteristics to attempt to take into account the confounding effects.

The estimates from the fitted model for the binary measurement error indicator for education (Table 4.9) suggest that people aged 31 to 66 years old are slightly less likely to report a different education level than younger people. This could have several interpretations. Since the ESS–Norway is a face–to–face survey, a plausible example related to social desirability (Tourangeau et al., 2010; Sakshaug et al., 2010) could be that younger people may try more to impress interviewers by reporting a higher level of education than the more mature ones. Other explanations such as linking theory (Durrant et al., 2010) cannot be substantiated here since neither interviewer's age nor any other characteristics are available. Furthermore, people from Hedmark and Oppland are slightly more likely to report a different education level than people from Oslo and Akershus. Now, considering the estimates from the fitted model for the binary measurement error indicator for household size (also in Table 4.9), one may conclude that women are less likely

to report a different household size than men, people older than 30 are less likely to report a different household size than younger people. And people from Agder and Rogaland are less likely to report a different household size than people from Oslo and Akershus. In addition, for both measurement error indicators, there are no interviewer effects.

Table 4.8: Parameter estimates (with standard errors in parenthesis) for the random intercept logistic models (empty models)

Fixed effects	Education	Household size
Constant	-1.053 (0.061)	-0.996 (0.061)
Random effects		
$\sigma^2_{u_0}$	0.022 (0.025)	0.024 (0.029)

* $p < 0.10$; ** $p < 0.05$

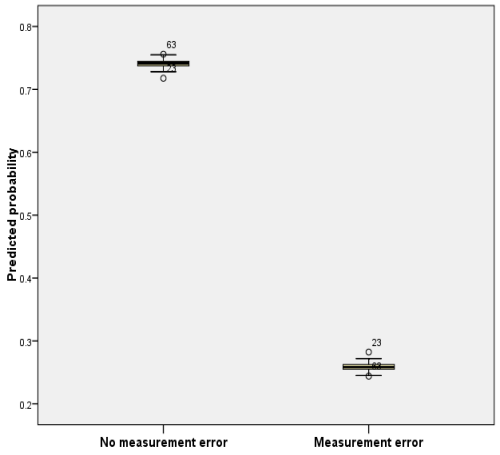
Table 4.9: Parameter estimates (with standard errors in parenthesis) for the random intercept logistic models controlling for explanatory variables

Fixed effects	Categories	Education	Household size
Constant		-0.847 (0.182)	-0.033 (0.178)
Gender	Female	-0.195 (0.120)	-0.600 (0.121) **
Age group	31 – 66	-0.238 (0.139) *	-0.623 (0.132) **
	Over 66	-0.122 (0.191)	-2.146 (0.271) **
Area	Hedmark and Oppland	0.461 (0.249) *	0.062 (0.264)
	South Eastern	0.183 (0.191)	0.019 (0.196)
	Agder and Rogaland	0.097 (0.203)	-0.441 (0.215) **
	Western Norway	-0.150 (0.203)	-0.111 (0.201)
	Troendelag	0.014 (0.235)	-0.149 (0.242)
	Northern Norway	-0.243 (0.250)	-0.030 (0.241)
Random effects			
$\sigma^2_{u_0}$		0.016 (0.020)	0.028 (0.034)

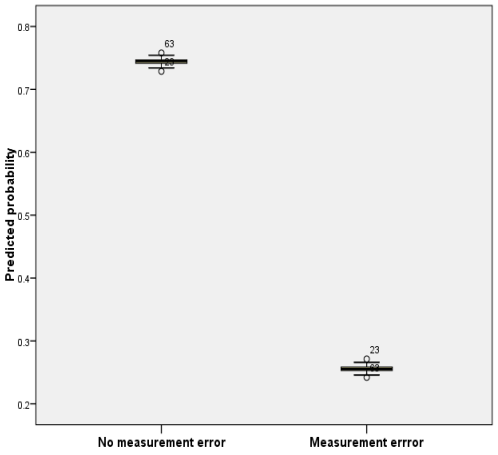
The baseline categories for the explanatory variables are Male, 15 – 30 and Oslo and Akershus.
* $p < 0.10$; ** $p < 0.05$

To assess the between–interviewer variation, box–plots of the predicted probabilities for the categories of the binary measurement error indicators for each interviewer are produced. Figures 4.3(a) and 4.3(b) illustrate the box–plots for the predicted probabilities respectively from the empty model and the model controlling for explanatory variables for education. The predicted probabilities from the latter model are computed by entering the sample proportions of each category of the explanatory variables. The same types of plots for household size are illustrated in Figures 4.3(c) and 4.3(d). Inspecting these plots, the predicted probabilities for the measurement error category of the indicators, for reporting education level and household size, vary around 25%, whereas for the no measurement error category, they vary around 75%. However, these predicted probabilities do not seem to vary much within the measurement error categories, neither for the empty models nor for the models controlling for gender, age and area.

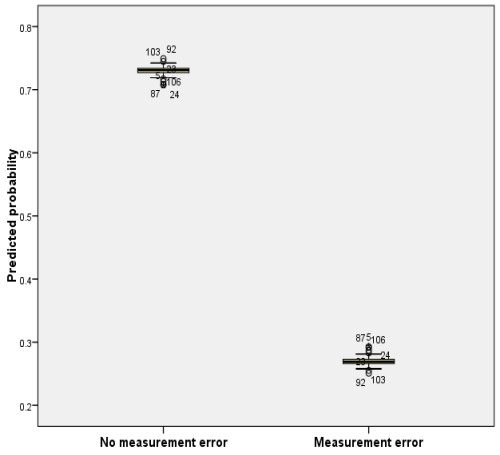
It is possible that this binary coding for the measurement error indicators is masking some potential interviewer effects on measurement error. For instance, some interviewers may be assigned to people who over report their education level and household size (positive effects), whereas some other interviewers may be assigned to people who under report their education level and household size (negative effects). Thus, on average these effects may cancel themselves out, resulting in negligible interviewer effects. Therefore, recoding these dependent variables to distinguish the measurement error category between over and under reporting education level and household size may be more informative.



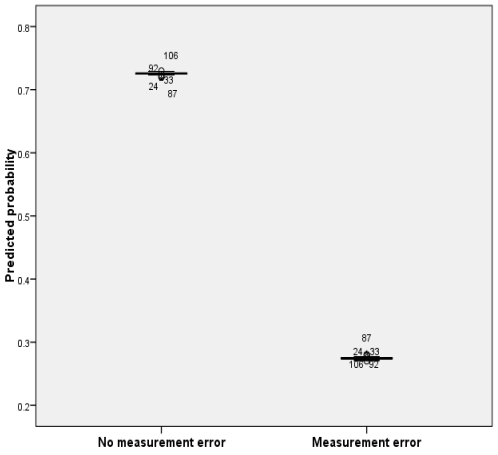
(a) Empty model for education



(b) Model controlling for explanatory variables for education



(c) Empty model for household size



(d) Model controlling for explanatory variables for household size

Figure 4.3: Box–plots of the predicted probabilities of logistic models for measurement error indicators for education and household size

4.4.2 Modelling the three category measurement error indicator

The codes for the measurement error indicators with three categories are displayed in Tables 4.10 and 4.11 for education and household size, respectively.

Interestingly, fitting two–level random intercept multinomial (empty) models for both measurement error indicators, the estimates in Table 4.12 indicate that the measurement error, for reporting education and household size, significantly varies by interviewers or by a combination of confounding level–two effects. This may

substantiate the hypothesis that some interviewers are assigned to people who under report education level and household size and others are assigned to people who over report education level and household size, so that the interviewer effects are no longer negligible.

Table 4.10: Coding for the cross–tabulation for education from ESS–Norway and from register with three categories

Education from	Education from register		
survey	Low	Middle	High
Low	1	3	3
Middle	2	1	3
High	2	2	1

Table 4.11: Coding for the cross–tabulation for household size from ESS–Norway and from register with three categories

Household size	Household size from register		
from survey	1 person	2 people	3+ people
1 person	1	3	3
2 people	2	1	3
3+ people	2	2	1

Table 4.12: Parameter estimates (with standard errors in parenthesis) for the random intercept multinomial model (empty model)

	Education		Household size	
Fixed effects	$\log(\pi_{2jk}/\pi_{1jk})$	$\log(\pi_{3jk}/\pi_{1jk})$	$\log(\pi_{2jk}/\pi_{1jk})$	$\log(\pi_{3jk}/\pi_{1jk})$
Constant	-1.344 (0.083)	-2.793 (0.147)	-1.388 (0.083)	-2.378 (0.128)
Random effects				
$\sigma^2_{u_0^{(2)}}$	0.201 (0.062) **		0.195 (0.060) **	
$\sigma^2_{u_0^{(3)}}$	0.392 (0.172) **		0.365 (0.141) **	
$\sigma_{u_0^{(2)}u_0^{(3)}}$	0.020 (0.067)		0.033 (0.063)	

* $p < 0.10$; ** $p < 0.05$

Controlling for gender, age and area (Table 4.13), women are less likely to over report education level than men, older people (over 30 years old) are less likely to over report education level than younger people and people from Hedmark and Oppland and Troendelag are more likely to under report education level than people from Oslo and Akershus. Now, for the measurement error indicator for household size (also in Table 4.13), the estimates suggest that women are less likely to over report household size than men and older people (over 30 years old) are less likely to over report household size than younger people. Furthermore, after controlling for these characteristics, the interviewer effects are only slightly significant for those who over report education level and household size.

Figures 4.4(a) and 4.4(b) illustrate the scatterplots for the predicted probabilities of over and under reporting respectively for education and household size. In Figure 4.4(a) and less visually obvious in Figure 4.4(b), the association between the two predicted probabilities seems to be negative. This means that some interviewers are assigned to people who over report their education level (or household size) whereas others are assigned to people who under report their education level (or household size). This strengthens the explanation given for the neglected interviewer effects from the two–level random intercept logistic models in Section 4.4.1.

Figures 4.5(a) and 4.5(b) illustrate the box–plots for the predicted probabilities for the three categories of the measurement error indicators for each interviewer from respectively the empty model and the model controlling for explanatory variables for education. Investigating these plots, it seems that there is a smaller interviewer–level variation for the predicted probabilities after controlling for gender, age and area (Figure 4.5(b)) compared to the variation in the plots in Figure 4.5(a). This could mean that these social and geographical characteristics explain

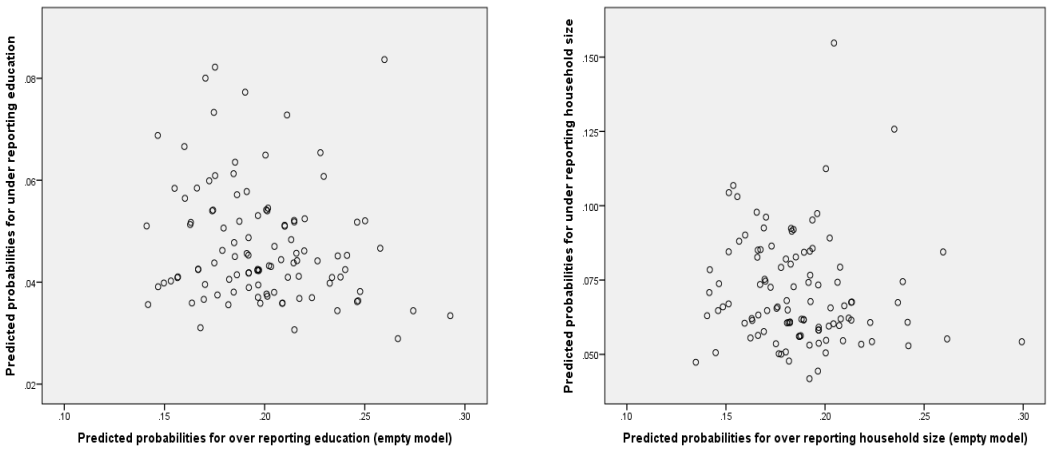
Table 4.13: Parameter estimates (with standard errors in parenthesis) for the random intercept multinomial model controlling for explanatory variables

		Education		Household size	
Fixed effects	Categories	$\log(\pi_{2jk}/\pi_{1jk})$	$\log(\pi_{3jk}/\pi_{1jk})$	$\log(\pi_{2jk}/\pi_{1jk})$	$\log(\pi_{3jk}/\pi_{1jk})$
Constant		-0.717 (0.201)	-5.282 (0.684)	-0.344 (0.205)	-1.539 (0.319)
Gender	Female	-0.324 (0.131) **	0.294 (0.242)	-0.731 (0.141) **	-0.295 (0.201)
Age group	31 – 66	-0.381 (0.145) **	1.356 (0.581) **	-0.550 (0.150) **	-0.799 (0.217) **
	Over 66	-1.189 (0.267) **	2.953 (0.587) **	-2.393 (0.355) **	-1.815 (0.416) **
Area	Hedmark and Oppland	0.346 (0.283)	1.093 (0.586) **	-0.154 (0.325)	0.446 (0.454)
	South Eastern	0.078 (0.217)	0.674 (0.488)	0.013 (0.223)	0.135 (0.363)
	Agder and Rogaland	-0.079 (0.233)	0.948 (0.512) *	-0.420 (0.250) *	-0.487 (0.429)
	Western Norway	-0.283 (0.232)	0.525 (0.502)	-0.225 (0.237)	0.166 (0.361)
	Troendelag	-0.283 (0.279)	1.075 (0.516) **	-0.032 (0.273)	-0.793 (0.529)
	Northern Norway	-0.419 (0.290)	0.532 (0.583)	-0.116 (0.284)	0.181 (0.435)
Random effects					
$\sigma^2_{u_0^{(2)}}$		0.052 (0.031) *		0.066 (0.039) *	
$\sigma^2_{u_0^{(3)}}$		0.261 (0.168)		0.307 (0.192)	
$\sigma_{u_0^{(2)}u_0^{(3)}}$		0.014 (0.048)		0.040 (0.061)	

The baseline categories for the explanatory variables are Male, 15 – 30 and Oslo and Akershus.
* $p < 0.10$; ** $p < 0.05$

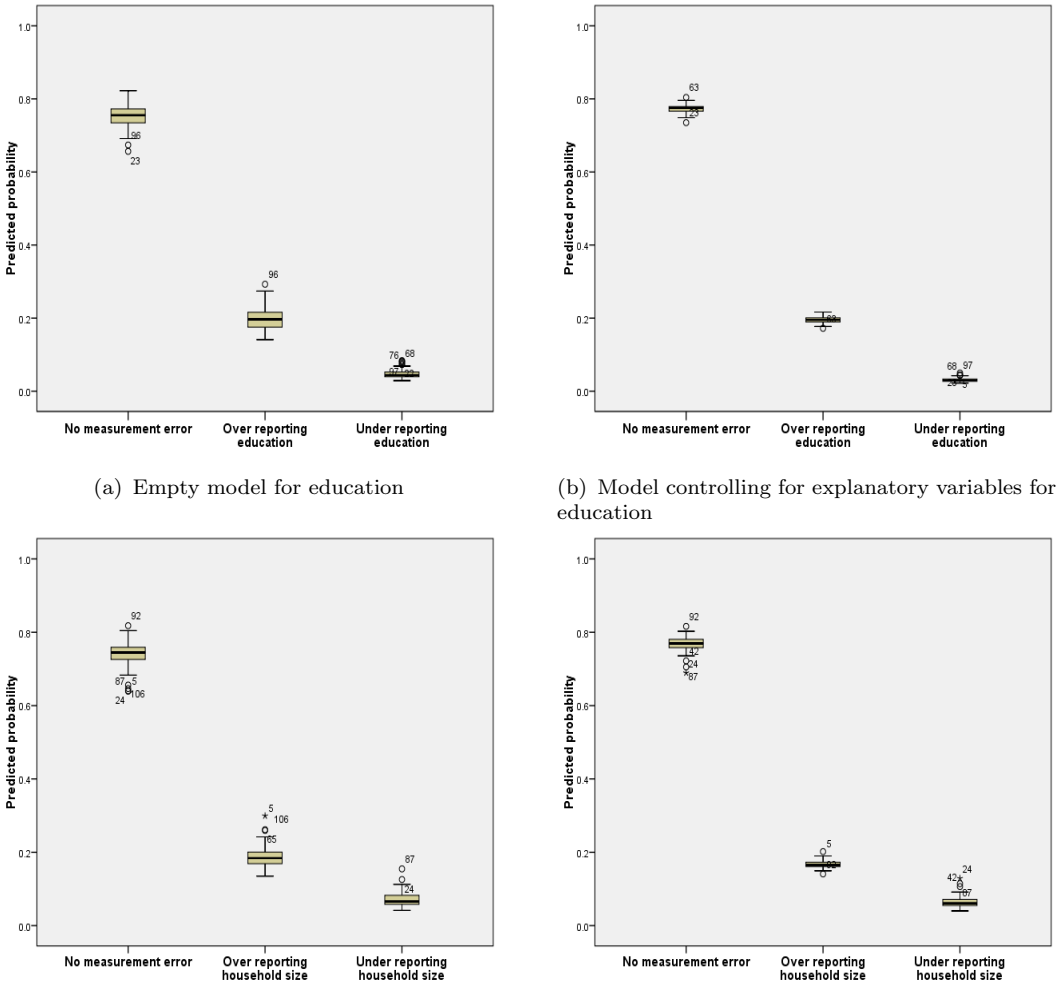
most of the level–two variation. Similar comments can also be made about the variation in the plots in Figures 4.5(c) and 4.5(d).

Clearly, coding the measurement error indicators into more categories provided more information about how the measurement error varies. In Figures 4.5(b) and 4.5(d), the box–plots on the upper left represent predicted probabilities for the no measurement error category. Apparently, even after controlling for gender, age and area, there is still a bit of unexplained variation left. Thus, one can disaggregate this no measurement error category into more categories and try to find other hidden effects.



(a) Education (b) Household size

Figure 4.4: Scatterplot for the predicted probabilities of under reporting versus over reporting for education and household size



(a) Empty model for education

(b) Model controlling for explanatory variables for education

(c) Empty model for household size

(d) Model controlling for explanatory variables for household size

Figure 4.5: Box–plots of the predicted probabilities of multinomial models for measurement error indicators for education and household size

4.4.3 Modelling the five category measurement error indicator

Now, the measurement error indicators are coded into five category variables as presented in Table 4.14 for education and in Table 4.15 for household size. The meaning for each category of these variables is the same as in Section 4.3.

The parameter estimates in Tables 4.16 and 4.17 for the fitted two–level random intercept multinomial (empty) models, for education and household size, suggest

that, surprisingly, there is evidence of level–two effects on the measurement error and no measurement error categories.

Table 4.14: Coding for the cross–tabulation for education from ESS–Norway and from register with five categories

Education from survey	Education from register		
	Low	Middle	High
Low	1	5	5
Middle	4	2	5
High	4	4	3

Table 4.15: Coding for the cross–tabulation for household size from ESS–Norway and from register with five categories

Household size from survey	Household size from register		
	1 person	2 people	3+ people
1 person	1	5	5
2 people	4	2	5
3+ people	4	4	3

Table 4.16: Parameter estimates (with standard errors in parenthesis) for the random intercept multinomial model (empty model) for reporting education

Fixed effects	$\log(\pi_{2jk}/\pi_{1jk})$	$\log(\pi_{3jk}/\pi_{1jk})$	$\log(\pi_{4jk}/\pi_{1jk})$	$\log(\pi_{5jk}/\pi_{1jk})$
Constant	0.628 (0.107)	0.638 (0.118)	0.278 (0.113)	-1.198 (0.165)
Random effects				
$\sigma^2_{u_0^{(2)}}$	0.310 (0.087) **			
$\sigma^2_{u_0^{(3)}}$	0.606 (0.181) **			
$\sigma^2_{u_0^{(4)}}$	0.354 (0.107) **			
$\sigma^2_{u_0^{(5)}}$	0.503 (0.183) **			
$\sigma_{u_0^{(2)}u_0^{(3)}}$	0.020 (0.093)			
$\sigma_{u_0^{(2)}u_0^{(4)}}$	0.061 (0.072)			
$\sigma_{u_0^{(2)}u_0^{(5)}}$	0.036 (0.087)			
$\sigma_{u_0^{(3)}u_0^{(4)}}$	0.173 (0.113)			
$\sigma_{u_0^{(3)}u_0^{(5)}}$	0.055 (0.136)			
$\sigma_{u_0^{(4)}u_0^{(5)}}$	0.052 (0.100)			

* $p < 0.10$; ** $p < 0.05$

Table 4.17: Parameter estimates (with standard errors in parenthesis) for the random intercept multinomial model (empty model) for reporting household size

Fixed effects	$\log(\pi_{2jk}/\pi_{1jk})$	$\log(\pi_{3jk}/\pi_{1jk})$	$\log(\pi_{4jk}/\pi_{1jk})$	$\log(\pi_{5jk}/\pi_{1jk})$
Constant	0.486 (0.112)	0.829 (0.103)	0.239 (0.112)	-0.780 (0.147)
Random effects				
$\sigma^2_{u_0^{(2)}}$	0.421 (0.133) **			
$\sigma^2_{u_0^{(3)}}$	0.324 (0.097) **			
$\sigma^2_{u_0^{(4)}}$	0.312 (0.093) **			
$\sigma^2_{u_0^{(5)}}$	0.456 (0.154) **			
$\sigma_{u_0^{(2)}u_0^{(3)}}$	0.112 (0.087)			
$\sigma_{u_0^{(2)}u_0^{(4)}}$	0.100 (0.086)			
$\sigma_{u_0^{(2)}u_0^{(5)}}$	0.060 (0.100)			
$\sigma_{u_0^{(3)}u_0^{(4)}}$	0.089 (0.073)			
$\sigma_{u_0^{(3)}u_0^{(5)}}$	0.008 (0.083)			
$\sigma_{u_0^{(4)}u_0^{(5)}}$	0.043 (0.083)			

* $p < 0.10$; ** $p < 0.05$

Controlling for gender, age and area in the model for reporting education level (Table 4.18), women are less likely to over report education level than men, people aged 31 to 66 year old are more likely to over report education level, whereas older people (over 66 years old) are less likely to over report education level than younger people and people from Troendelag and Northern Norway are less likely to over report education level than people from Oslo and Akershus. Whilst, in the model for reporting household size (Table 4.19), women are less likely to over report the household size than men and older people (over 31 years old) are less likely to over report household size than younger people. The above interpretations are based on the predicted probabilities for the measurement error indicators of reporting education level and household size for each category of the explanatory variables in Tables 4.20 and 4.21. Note further that even after controlling for these confounding effects, there is still evidence of significant interviewer effects on measurement error, especially for the over reporting category. Therefore, one may conclude that the over reporting measurement error may be driven by interviewer.

Table 4.18: Parameter estimates (with standard errors in parenthesis) for the random intercept multinomial model controlling for explanatory variables for reporting education

Fixed effects	Categories	$\log(\pi_{2jk}/\pi_{1jk})$	$\log(\pi_{3jk}/\pi_{1jk})$	$\log(\pi_{4jk}/\pi_{1jk})$	$\log(\pi_{5jk}/\pi_{1jk})$
Constant		0.086 (0.287)	0.373 (0.306)	0.810 (0.286)	-3.832 (0.668)
Gender	Female	-0.590 (0.172) **	-0.145 (0.178)	-0.599 (0.181) **	0.009 (0.273)
Age group	31 – 66	1.541 (0.200) **	2.296 (0.212) **	1.084 (0.198) **	2.798 (0.570) **
	Over 66	0.551 (0.241) **	0.405 (0.274)	-0.868 (0.301) **	3.308 (0.575) **
Area	Hedmark and Oppland	-0.556 (0.422)	-1.202 (0.466) **	-0.455 (0.419)	0.368 (0.645)
	South Eastern	-0.012 (0.319)	-1.154 (0.342) **	-0.505 (0.324)	0.134 (0.524)
	Agder and Rogaland	0.073 (0.342)	-1.250 (0.393) **	-0.658 (0.355) *	0.436 (0.554)
	Western Norway	0.420 (0.325)	-1.011 (0.357) **	-0.615 (0.346) *	0.267 (0.541)
	Troendelag	0.043 (0.369)	-1.178 (0.418) **	-0.830 (0.399) **	0.593 (0.571)
	Northern Norway	-0.421 (0.355)	-1.745 (0.414) **	-1.370 (0.393) **	-0.336 (0.628)
Random effects					
$\sigma^2_{u_0^{(2)}}$	0.140 (0.065) **				
$\sigma^2_{u_0^{(3)}}$	0.500 (0.161) **				
$\sigma^2_{u_0^{(4)}}$	0.192 (0.081) **				
$\sigma^2_{u_0^{(5)}}$	0.345 (0.178) *				
$\sigma_{u_0^{(2)}u_0^{(3)}}$	0.031 (0.075)				
$\sigma_{u_0^{(2)}u_0^{(4)}}$	0.069 (0.056)				
$\sigma_{u_0^{(2)}u_0^{(5)}}$	0.016 (0.073)				
$\sigma_{u_0^{(3)}u_0^{(4)}}$	0.202 (0.098) **				
$\sigma_{u_0^{(3)}u_0^{(5)}}$	0.091 (0.123)				
$\sigma_{u_0^{(4)}u_0^{(5)}}$	0.060 (0.085)				

The baseline categories for the explanatory variables are Male, 15 – 30 and Oslo and Akershus.
* $p < 0.10$; ** $p < 0.05$

Table 4.19: Parameter estimates (with standard errors in parenthesis) for the random intercept multinomial model controlling for explanatory variables for reporting household size

Fixed effects	Categories	$\log(\pi_{2jk}/\pi_{1jk})$	$\log(\pi_{3jk}/\pi_{1jk})$	$\log(\pi_{4jk}/\pi_{1jk})$	$\log(\pi_{5jk}/\pi_{1jk})$
Constant		-0.579 (0.369)	1.258 (0.307)	1.355 (0.309)	0.175 (0.407)
Gender	Female	-0.113 (0.177)	0.223 (0.174)	-0.642 (0.192) **	-0.213 (0.243)
Age group	31 – 66	0.917 (0.309) **	-0.583 (0.239) **	-0.745 (0.253)**	-0.994 (0.296) **
	Over 66	0.730 (0.326) **	-6.764 (1.306) **	-3.435 (0.413) **	-2.851 (0.461) **
Area	Hedmark and Oppland	0.288 (0.416)	-0.249 (0.409)	-0.188 (0.420)	0.395 (0.522)
	South Eastern	0.566 (0.318) *	0.473 (0.307)	0.421 (0.311)	0.542 (0.428)
	Agder and Rogaland	-0.071 (0.364)	0.460 (0.328)	-0.181 (0.344)	-0.269 (0.490)
	Western Norway	0.585 (0.335)	0.795 (0.320) **	0.367 (0.336)	0.726 (0.433) *
	Troendelag	0.320 (0.384)	0.254 (0.369)	0.182 (0.373)	-0.676 (0.574)
	Northern Norway	1.286 (0.428) **	0.852 (0.422) **	0.728 (0.439) *	0.998 (0.548) *
Random effects					
$\sigma^2_{u_0^{(2)}}$	0.292 (0.127) **				
$\sigma^2_{u_0^{(3)}}$	0.191 (0.084) **				
$\sigma^2_{u_0^{(4)}}$	0.135 (0.064) **				
$\sigma^2_{u_0^{(5)}}$	0.322 (0.166) *				
$\sigma_{u_0^{(2)}u_0^{(3)}}$	0.124 (0.081)				
$\sigma_{u_0^{(2)}u_0^{(4)}}$	0.096 (0.070)				
$\sigma_{u_0^{(2)}u_0^{(5)}}$	0.045 (0.097)				
$\sigma_{u_0^{(3)}u_0^{(4)}}$	0.086 (0.060)				
$\sigma_{u_0^{(3)}u_0^{(5)}}$	0.001 (0.079)				
$\sigma_{u_0^{(4)}u_0^{(5)}}$	0.038 (0.069)				

The baseline categories for the explanatory variables are Male, 15 – 30 and Oslo and Akershus.
* $p < 0.10$; ** $p < 0.05$

Table 4.20: Predicted probabilities for the measurement error indicator for reporting education (5 categories) for each category of the explanatory variables

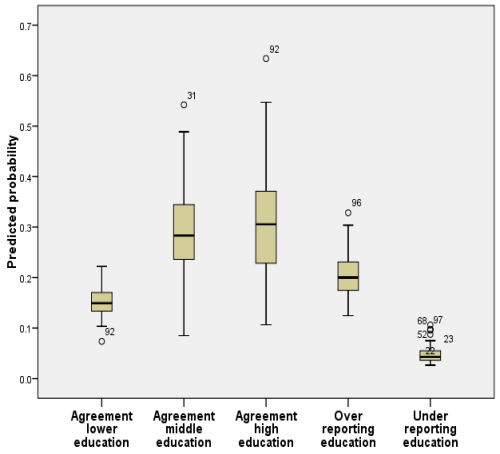
Explanatory variables	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_4$	$\hat{p}_5$
Male	0.11	0.35	0.27	0.25	0.03
Female	0.16	0.27	0.33	0.19	0.04
Age group 1 (15 30)	0.30	0.25	0.15	0.29	0.01
Age group 2 (31 66)	0.08	0.29	0.38	0.22	0.03
Age group 3 (over 66)	0.23	0.33	0.18	0.09	0.17
Region 1 (Oslo and Akershus)	0.08	0.19	0.48	0.23	0.02
Region 2 (Hedmark and Oppland)	0.16	0.21	0.29	0.29	0.05
Region 3 (South Eastern)	0.14	0.32	0.26	0.24	0.03
Region 4 (Agder and Rogaland)	0.14	0.36	0.24	0.21	0.04
Region 5 (Western Norway)	0.12	0.41	0.25	0.18	0.03
Region 6 (Troendelag)	0.15	0.35	0.27	0.18	0.05
Region 7 (Northern Norway)	0.22	0.34	0.23	0.17	0.03

Table 4.21: Predicted probabilities for the measurement error indicator for reporting household size (5 categories) for each category of the explanatory variables

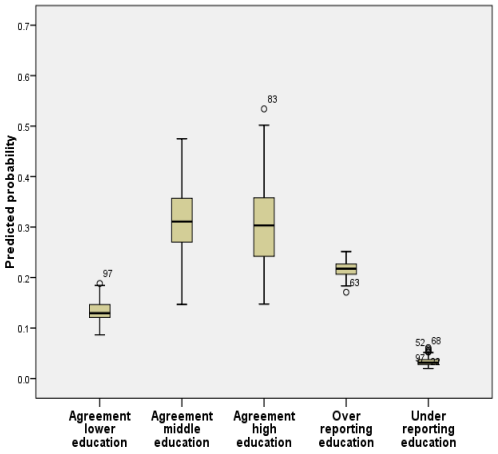
Explanatory variables	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_4$	$\hat{p}_5$
Male	0.16	0.26	0.21	0.28	0.09
Female	0.18	0.26	0.30	0.17	0.08
Age group 1 (15 30)	0.08	0.06	0.47	0.28	0.11
Age group 2 (31 66)	0.12	0.23	0.39	0.20	0.06
Age group 3 (over 66)	0.36	0.57	0.00	0.04	0.03
Region 1 (Oslo and Akershus)	0.22	0.23	0.22	0.24	0.08
Region 2 (Hedmark and Oppland)	0.22	0.30	0.17	0.19	0.12
Region 3 (South Eastern)	0.15	0.27	0.24	0.24	0.10
Region 4 (Agder and Rogaland)	0.21	0.21	0.33	0.19	0.06
Region 5 (Western Norway)	0.14	0.25	0.30	0.21	0.11
Region 6 (Troendelag)	0.19	0.28	0.25	0.25	0.04
Region 7 (Northern Norway)	0.10	0.37	0.22	0.21	0.10

The box–plots for the predicted probabilities for the five categories of the measurement error indicators for each interviewer from the models controlling for explanatory variables in Figures 4.6(b) and 4.6(d) show reduction of interviewer–level variation within all the categories compared to the plots for the empty models

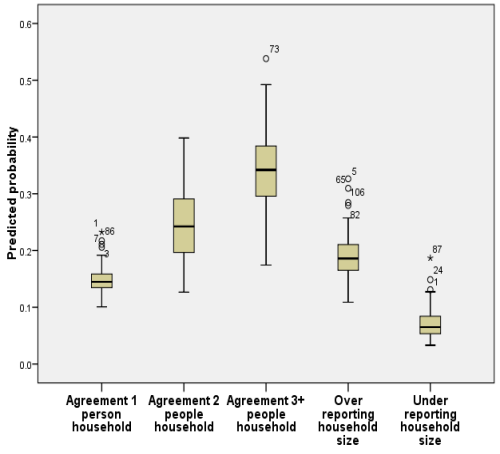
(Figures 4.6(a) and 4.6(c)). These reductions are especially smaller for the categories representing measurement error.



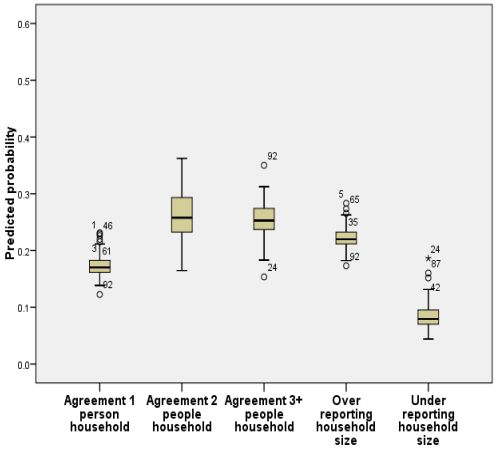
(a) Empty model for education



(b) Model controlling for explanatory variables for education



(c) Empty model for household size

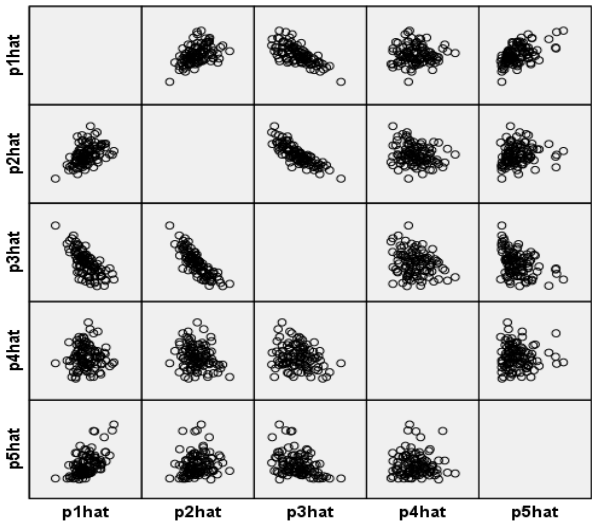


(d) Model controlling for explanatory variables for household size

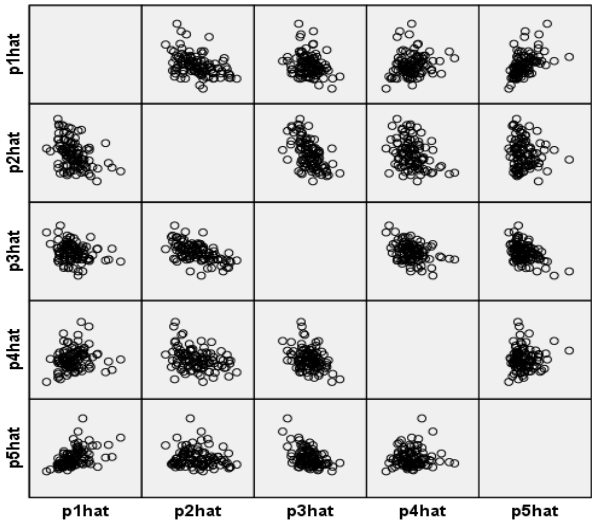
Figure 4.6: Box–plots of the predicted probabilities of multinomial models for five category measurement error indicators for education and household size

In the box–plots in Figures 4.6(a) to 4.6(d), one can perceive that there is a considerable variability for some of the categories that represent no measurement error, i.e. agreement on category 2 and category 3 of education and household size. This variability may be due to interviewer assignment effects, meaning that some interviewers are more assigned to sampled units that agreed on category 2 of the variables on both ESS–Norway and register, whereas some other interviewers are more assigned to sample units that agreed on category 3 of the variables. The

scatterplots matrices in Figures 4.7(a) and 4.7(b) show the association between the predicted probabilities for these two categories (*p2hat* and *p3hat*).



(a) Education



(b) Household size

Figure 4.7: Scatterplots matrix of the predicted probabilities for the measurement error indicator for education and household size

Other categorizations of the measurement error indicators for education and household size, as well as corresponding fitted models can be found in Appendix G.

In the residual analysis (please refer to Appendix H to check the residual plots) for the two–level random intercept multinomial models for three and five measurement error indicators for reporting education and household size, the normality assumption for the level–two residuals seems reasonable. In addition, the presence of outliers is not visually detected.

4.5 Conclusions

This study aims to investigate the effects of interviewers on measurement error. Multilevel models are applied to a dataset that contains data from the 2010 Norwegian sample of the ESS linked to administrative data. Unlike the case in many other studies on measurement error, the dataset used in the application of the proposed models contains information on essentially the same variables from both the survey and the register. Thus, the main contribution of this chapter is to propose an approach to assess the interviewer effects on measurement error.

The multilevel models considered here are applied to measurement error indicators created based on the joint distributions of education and household size from the ESS–Norway and a register. There are three main findings from this research. Firstly, the estimates from the fitted empty two–level logistic models on binary measurement error indicators (for education and household size) indicate that the significant measurement error does not vary by interviewer. In fact, even if the variance of the random effects was significant, these effects may not be purely interviewer effects but a combination of interviewer and other confounding effects such as area effects.

Secondly, refining the measurement error indicators to have three categories (no measurement error, over and under reporting) and fitting empty two–level multinomial models (for education and household size), interestingly, the estimates suggest significant level–two effects on measurement error. However, after controlling for gender, age and area, the interviewer effects are only border line significant for those who over report education and household size.

Lastly, considering the measurement error indicators with five categories and fitting empty two–level multinomial models (for education and household size), the estimates suggest significant level–two effects on measurement error. In addition, after controlling for gender, age and area, there is evidence of significant interviewer effects. Also, irrespective of the number of categories for the measurement error indicator, the interviewer effects should not be significant if there is no measurement error. Other categorizations for this indicator are considered in the model fitting process. The analyses for the measurement error indicator with 4 up to 9 different categories, all provide significant interviewer effects on measurement error.

One possible limitation of the research in this chapter is the assumption that the register variables are true measures for the variables observed in the survey. Although this assumption is often valid, there is the possibility for part of the register–based information not to correspond to the truth, for instance when such information is outdated or affected by measurement error as well. In such cases, one can treat the effects found in this study as interviewer effects on the joint distribution of the variables, as opposed to the treatment of the variables in the register as true measures.

Another limitation might be that the identified interviewer effects may not be true interviewer effects since it was not possible to properly control for area effects, which are potentially confounded with interviewer effects (Durrant et al., 2010;

O'Muircheartaigh and Campanelli, 1999). The ideal model to control for these two effects would be a cross–classified multilevel model. However, the available area variable consists of seven large regions, which is not disaggregated enough to try disentangling area and interviewer effects by applying this approach.

The research in this chapter provides a useful way for survey organizations to assess interviewer effects on measurement error. However, some of these organization usually avoid asking some questions in the survey when information about these questions could be acquired from an external (e.g, a register) source. In these cases, the reduction of the number of questions to be measured in the survey may be a strategy to lessen the interviewer workload and to avoid that individuals have to respond to a long questionnaire. However, at least for key variables that may be susceptible to measurement error, the findings from this study suggest that survey practitioners should attempt to measure these variables in the survey as well, so that by having data from the two sources, the models proposed in the chapter could be applied. This methodology is therefore a useful tool to, for example, monitor interviewers performance as well as to identify where to improve interviewer training to avoid the occurrence of measurement error.

On one hand, if the measurement error was detected as a result of a monitoring process of the performance of interviewers, a next step would be to intensify interviewer training as an attempt to minimize the occurrence of measurement error on variables of interest in a future survey. On the other hand, given that some estimates are affected by measurement error, a next step would be to apply a suitable adjustment method for measurement error. These methods are well documented in the literature as, for instance, in the general references Fuller (1987), Carroll et al. (2006) and Buonaccorsi (2010).

As a suggestion for further work considering the ESS–Norway, it would be interesting to have access to another area variable consisting of, for instance, the allocation of interviewers by districts or postcode sectors to enable using this variable as one of the level–two variables in cross–classified multilevel models to properly separate interviewer and area effects. Alternatively, the application of the proposed models to another linked dataset containing this type of area variable could also be interesting to improve the models discussed in this chapter. Furthermore, measurement error and nonresponse bias may be connected in a survey since interviewer additional efforts to reduce nonresponse may exacerbate the occurrence of measurement error (West and Olson, 2010; Olson, 2006; Olson and Kennedy, 2006). However, the methodology discussed in this chapter does not take into account the nonrespondents. Thus, in order to investigate the role of interviewers in this scenario, a different model should be proposed.

Another idea for future research in the context of Chapter 4 is when continuous dependent variables from the survey (y^*_{ij}) and the same variables from administrative data ($\tilde{y}_{ij}$) are available. The analysis of the interviewer effects on measurement error on variables of interest could be performed by considering multilevel models (with random interviewer effects) for continuous data. In these models, the interviewer effects can be assessed by defining the response variable as the difference between y^*_{ij} and $\tilde{y}_{ij}$. Alternatively, depending on the distribution of the variables of interest, the dependent variable can be defined as the ratio $y^*_{ij}/\tilde{y}_{ij}$ or $\log(y^*_{ij}/\tilde{y}_{ij})$.

Chapter 5

General Conclusions

This study aims to detect empirically interviewer effects on nonresponse bias and measurement error in sample surveys. It is recognized that nonresponse and measurement error are threats to data quality because of selection bias and inaccurate reportage. In interviewer–administered surveys, interviewers may have an effect on nonrespondents and respondents. If the respondent pool has characteristics that differ from those in the nonrespondent pool, this can lead to nonresponse bias. Whilst if respondents misreport the required information or interviewers fail to convey the respondent answers accurately, this may lead to measurement error. This chapter provides a summary of the main findings from Chapters 2, 3 and 4. In addition, it presents a discussion of potential limitations of the study as well as recommendations for future work.

In Chapter 2, the main contributions are to introduce a novel approach to assess nonresponse bias and in particular to assess the interviewer effects on nonresponse bias using multilevel models. In this approach, a binary response indicator is used as one of the explanatory variables in the model. In addition to a random intercept, a random coefficient is introduced for the response indicator to investigate the effects of interviewers on nonresponse bias. This approach contrasts with

some applications of multilevel models in the literature to investigate interviewer effects. For instance, Loosveldt and Beullens (2014) consider models having a random intercept and random slopes for each explanatory variable using the response indicator as the dependent variable of the model. Chapter 2 also discusses an application to support the proposed method using a cross–sectional dataset from the 2001 UK Labour Force Survey (LFS) linked to census records. The key advantage of having census linked data in this study is that information on respondents and nonrespondents are available, which is essential for the analysis of nonresponse bias.

The proposed models in Chapter 2 are applied to the dependent variables employment indicator and academic qualification indicator. In the analyses for the indicator of employment, it is found that, on average, there is no nonresponse bias or evidence of interviewer effects on nonresponse bias. On the other hand, in the analyses for the indicator of academic qualification, on average, there is nonresponse bias and also evidence of interviewer effects on nonresponse bias. The fact that the nonresponse bias is not significant for employment but it is significant for academic qualification is supported in the literature. For example, Groves (2006) and Groves and Peytcheva (2008) report that different estimates within a survey may vary with respect to nonresponse bias. Another finding from the research in Chapter 2 is that the interviewer assignment effect (the effect resulting from differential assignment of interviewer with respect to the yet to be collected dependent variable) is significant. It is possible that interviewer and area effects may be confounded. Then, as an attempt to separate these two effects, a cross–classified model is considered and both the interviewer assignment effect and the area effect are significant. This may mean that the cross–classified model is not successful in disentangling these effects. However, an investigation of the interviewer–level residuals suggests that three interviewers are unusually assigned. After that, the

cross–classified models are refitted without the cases from these three interviewers. As a result, the interviewer assignment effects initially attributed to interviewers are actually due to areas. This finding highlights the importance of a thorough residual analysis to identify possible unusual behaviour in the data. Finally, after removing these cases, even though the assignment of interviewers to sampled units is not at random (interpenetration), the cross–classified model seems to correctly disentangle these two random effects. Most importantly, it is found evidence that the nonresponse bias may be driven by interviewers.

Potential limitations in the research in Chapter 2 may be due to a large number of deleted cases from the LFS dataset. Also, in the analysis, 80 cases were assigned as respondents as a working assumption. However, these cases could all belong to the nonrespondent pool, or only part of them could belong to the nonrespondent pool or they all could actually belong to the respondent pool. Although they are only a small number of cases, further analysis assigning these 80 cases as nonresponse or discarding them from the sample is needed to analyse the impact of this assumption on the results.

In Chapter 3, the main contribution is to investigate further the assessment of interviewer effects on nonresponse bias under different survey conditions than those in Chapter 2. In contrast to Chapter 2, where the dataset used comes from the linkage of a cross–sectional face–to–face survey and individual census records, the focus in this chapter is on datasets from different data collection modes: a CATI survey, the Dutch Consumer Confidence Survey (CCS) and a longitudinal study, the British Household Panel Survey (BHPS). Another distinction from Chapter 2 is that these datasets are linked to other sources of auxiliary variables, since census records are not available. For the CCS, the source of auxiliary variables linked to the survey data in order to acquire information on respondents and nonrespondents is administrative data. Whilst for the BHPS, time invariant variables from its first

wave are linked to respondents and nonrespondents from wave 10. This further investigation in Chapter 3 provides then a way to shed some light to the restricted literature that nonresponse biases are larger for telephone than to face–to–face surveys (Biemer, 2001).

Multilevel logistic models similar to those considered in Chapter 2 are used to analyse the datasets. The models are fitted for the dependent variables jobs in the household, type of housing and size of household from the CCS linked dataset, and father working and mother working indicators from the BHPS dataset. One important finding from the analysis for the dependent variables from the CCS linked dataset is that there are no interviewer assignment effects. Additionally, on average, there is nonresponse bias and initially there is evidence of interviewer effects on nonresponse bias. However, after controlling for gender and age of the respondents not only the random interviewer effects are no longer significant, but also a reduction in the nonresponse bias is observed. Although the bias is still significant, it is clear that part of the nonresponse bias is explained by controlling for gender and age for all three dependent variables from the CCS linked dataset. Therefore, one should include in the data analysis the explanatory variables gender and age to try to reduce the nonresponse bias. For the dependent variables from the BHPS dataset, there are significant interviewer assignment effects. Then, after controlling for the cross–classification of interviewer and area random effects, for father working indicator, the interviewer assignment effects are no longer significant. Also, the nonresponse bias reduces after controlling for age and ethnicity of the respondents. By contrast, for mother working indicator, even allowing for the cross–classification between interviewer and area effects, these effects are still entangled. In addition, there is some evidence of nonresponse bias.

The conclusions of the analyses in Chapter 3 may be weakened due to the linkage

of household–level instead of individual–level variables with the CCS data. Furthermore, the comparison between wave 1 and wave 10 of the BHPS may have implications for the analyses of interviewer effects. For instance, it is likely that some of those identified as nonresponders at wave 10 were also nonresponders at earlier waves. If they were assigned different interviewers at earlier waves, i.e., if there was not interviewer continuity (Campanelli and O'Muircheartaigh, 1999), then the cause of nonresponse would actually be these earlier interviewers, and the interviewer effects may be confounded with earlier interviewers.

Another limitation for the results in Chapter 3 is because a number of respondent and nonrespondent cases were deleted from the BHPS dataset in the process of merging data from wave 1 with data from wave 10 as well as cases without interviewer ID. Thus, the way the analyses are conducted, based on a dataset having a 9.5% nonresponse rate, may be seen as a relatively straightforward approach to detect nonresponse bias. However, if the deleted cases were indeed selective cases, in the sense that they do have different characteristics than those of the respondents, then the analyses for BHPS variables may not be correct. In this case, a more complex approach, such as by imputing values for the cases that had been deleted to take into account their missing information, would be advisable to potentially address the assessment of nonresponse bias. Such an approach, however, has the drawback of requiring more assumptions and models to be safely employed.

The importance of the findings of this thesis regarding the assessment of interviewer effects on nonresponse bias is that these effects are investigated under different survey conditions. Although the interviewer effects observed across the three different surveys in the analyses of nonresponse bias cannot be compared directly since these surveys are from different populations and survey topics, some of the findings in this research are consistent for a specific survey mode. The analyses undertaken provide therefore a type of sensitivity analysis for the results. For

instance, for the two face–to–face surveys (LFS and BHPS) there are initially significant interviewer assignment effects, whereas for the CATI survey (CCS) these effects are not significant at all. One plausible explanation is that in the CATI survey interviewers are randomly assigned to sampled units, i.e., the survey design is approximately interpenetrated (Campanelli and O'Muircheartaigh, 1999). On the other hand, other findings are common across survey modes. For example, there is evidence of nonresponse bias and of interviewer effects on nonresponse bias for a dependent variable from the LFS linked dataset and from the dependent variables from the CCS linked dataset.

One important practical implication for the detection of these interviewer effects is that it could be integrated an overall strategy to better allocate the survey resources in order to control the total survey error. For example, time and financial resources would only be invested in improving interviewer training given there is evidence that the interviewers affect the nonresponse biases significantly. Even if those effects are not found to be significant, the organizations could also report the application of the proposed methodology as part of their quality monitoring of the estimates produced, increasing the survey users trust on the survey results. In addition, lessons could be learnt from those interviewers that introduce small or no bias. So, survey agencies should pay attention to what these interviewers do in the field and use this knowledge to train the other interviewers accordingly. Furthermore, if interviewer characteristics are available in the dataset, these characteristics could be included in the models in order to examine whether the interviewer–level variation decreases. Also, the model could include an interaction between interviewer characteristics and the response indicator to investigate the interviewer effects within groups of interviewers.

In the context of interviewer effects on nonresponse bias, it would be interesting for future work to apply multilevel multinomial model to variables of interest

with more than two categories and also consider response indicators with more than two categories such as response, refusal and noncontact. In addition, interviewer's characteristics could also be controlled in the models to check if the interviewer–level variation is explained by this inclusion. Furthermore, the inclusion of household effects in the models could be worthy pursuing since nonresponse seems to be largely a household feature rather than a property of individuals.

In Chapter 4, the main contribution is to introduce an intuitive method to assess interviewer effects on measurement error. Some studies in the literature investigate interviewer effects on measurement error on survey estimates by looking at the proportion of the total interviewer variance that is due to this source of error (West and Olson, 2010; West et al., 2013). In this study, multilevel logistic models with interviewer random effects are applied to measurement error indicators that are created based on the joint distributions for the pairs of essentially the same variables from a survey and from administrative data. This chapter also describes an application to support the proposed method making use of a rich dataset that contains the observed variables from the 2010 Norwegian sample of the European Social Survey (ESS–Norway) linked to the "true" variables from administrative records.

The proposed models in Chapter 4 are applied to measurement error indicators for the variables education level and household size. The fitted models for the binary measurement error indicators suggest that the measurement errors do not vary by interviewers, i.e., there is no evidence of interviewer effects. Interestingly, the estimates from the empty models considering the three category measurement error indicators suggest that interviewer effects are significant for both under and over reporting of education and household size. However, after controlling for gender, age and area in the models, the interviewer effects are only borderline significant for those who over report education and household size. On the other

hand, for the five category measurement indicators, the model estimates indicate evidence of interviewer effects on measurement error. Generally, for the three different fitted models, the estimates, considering the measurement indicators for both variables, education level and household size, behave roughly the same. Also, irrespective of the number of categories for the measurement error indicator, the interviewer effects should not be significant if there is no measurement error. Other categorizations for this indicator are considered in the model fitting process. The analyses for the measurement error indicator with 4 up to 9 different categories, all provide significant interviewer effects on measurement error.

One possible limitation of the research in Chapter 4 is the assumption that the register variables are true measures for the variables observed in the survey. Although this assumption is often valid, there is the possibility for part of the register–based information not to correspond to the truth, for instance when such information is outdated or affected by measurement error as well. In such cases, one can treat the effects found in this study as interviewer effects on the joint distribution of the variables, as opposed to the treatment of the variables in the register as true measures.

The results in Chapter 4 might be also limited because the identified interviewer effects may not be true interviewer effects since it was not possible to properly control for area effects, which are potentially confounded with interviewer effects (Durrant et al., 2010; O'Muircheartaigh and Campanelli, 1999). The ideal model to control for these two effects would be a cross–classified multilevel model. However, the available area variable consists of seven large regions, which is not disaggregated enough to try disentangling area and interviewer effects by applying this approach.

The research in Chapter 4 provides a useful way for survey organizations to assess interviewer effects on measurement error. However, some of these organization

usually avoid asking some questions in the survey when information about these questions could be acquired from an external (e.g, a register) source. In these cases, the reduction of the number of questions to be measured in the survey may be a strategy to lessen the interviewer workload and to avoid that individuals have to respond to a long questionnaire. However, at least for key variables that may be susceptible to measurement error, the findings in Chapter 4 suggest that survey practitioners should attempt to measure these variables in the survey as well, so that by having data from the two sources, the models proposed in the chapter could be applied. This methodology is therefore a useful tool to, for example, monitor interviewers performance as well as to identify where to improve interviewer training to avoid the occurrence of measurement error.

As a suggestion for further work considering the ESS–Norway, it would be interesting to have access to another area variable consisting of, for instance, the allocation of interviewers by districts or postcode sectors to enable using this variable as one of the level–two variables in cross–classified multilevel models to properly separate interviewer and area effects. Alternatively, the application of the proposed models to another linked dataset containing this type of area variable could also be interesting to improve the models discussed in this chapter. Furthermore, measurement error and nonresponse bias may be connected in a survey since interviewer additional efforts to reduce nonresponse may exacerbate the occurrence of measurement error (West and Olson, 2010; Olson, 2006; Olson and Kennedy, 2006). However, the methodology discussed in this chapter does not take into account the nonrespondents. Thus, in order to investigate the role of interviewers in this scenario, a different model should be proposed.

Another idea for future research in the context of Chapter 4 is when continuous dependent variables from the survey (y^*_{ij}) and the same variables from administrative data ($\tilde{y}_{ij}$) are available. The analysis of the interviewer effects on measurement

error on variables of interest could be performed by considering multilevel models (with random interviewer effects) for continuous data. In these models, the interviewer effects can be assessed by defining the response variable as the difference between y_{ij}^* and $\tilde{y}_{ij}$. Alternatively, depending on the distribution of the variables of interest, the dependent variable can be defined as the ratio $y_{ij}^*/\tilde{y}_{ij}$ or $\log(y_{ij}^*/\tilde{y}_{ij})$.

The findings regarding interviewer effects on nonresponse bias and measurement error from the applications of models discussed in this thesis should be seen as a guidance for possible decisions. This limitation results from the general issue that causal–and–effect relationships are harder to be established from the statistical analyses solely, when the data come from observational studies. Thus, extension of the findings to other surveys may not follow directly. However, the methodologies proposed to assess the interviewer effects can be applied to any other survey, as long as it contains the essential variables for these analyses.

Appendices

Appendix A

Additional Tables for the LFS Linked Dataset

Table A.1: Frequency distribution for gender (LFS linked dataset)

Gender	Frequency	Percent
1 Male	2323	48.9
2 Female	2425	51.1
Total	4748	100.0

Table A.2: Frequency distribution for marital status (LFS linked dataset)

Marital status	Frequency	Percent
1 Single (Never married)	1555	32.8
2 Married (First marriage)	2175	45.8
3 Re–married	412	8.7
4 Separated (but still legally married)	123	2.6
5 Divorced	399	8.4
6 Widowed	84	1.8
Total	4748	100.0

Table A.3: Frequency distribution for student indicator (LFS linked dataset)

Student indicator	Frequency	Percent
1 Full–time	335	7.1
2 Not full–time	4413	92.9
Total	4748	100.0

Table A.4: Frequency distribution for health (LFS linked dataset)

Health	Frequency	Percent
1 Good	3376	71.1
2 Fairly good	976	20.6
3 Not good	396	8.3
Total	4748	100.0

Table A.5: Frequency distribution for carer (LFS linked dataset)

Carer	Frequency	Percent
1 Not a carer	4111	86.6
2 1-19 hrs care a week	434	9.1
3 20-49 hrs care a week	76	1.6
4 50+ hrs care a week	127	2.7
Total	4748	100.0

Table A.6: Frequency distribution for pensioner indicator (LFS linked dataset)

Pensioner indicator	Frequency	Percent
0 Not of pensionable age	4543	95.7
1 Of pensionable age	205	4.3
Total	4748	100.0

Table A.7: Frequency distribution for dependent child (LFS linked dataset)

Dependent child	Frequency	Percent
0 Not a dependent child	4563	96.1
1 Dependent child	185	3.9
Total	4748	100.0

Table A.8: Frequency distribution for highest qualification (LFS linked dataset)

Highest qualification	Frequency	Percent
0 No academic qualification	1222	25.7
1 O levels GCSEs A levels other	2252	47.4
2 First degree Higher degree NVQ levels	965	20.3
3 Other qualifications	309	6.5
Total	4748	100.0

Table A.9: Frequency distribution for age (LFS linked dataset)

Age	Frequency	Percent
2 16 – 34	1716	36.1
3 35 – 49	1645	34.6
4 50 – 64	1387	29.2
Total	4748	100.0

Table A.10: Frequency distribution for urban/rural indicator (LFS linked dataset)

Urban/rural indicator	Frequency	Percent
1 Urban	4235	89.2
2 Rural	513	10.8
Total	4748	100.0

Table A.11: Frequency distribution for ethnic group (LFS linked dataset)

Ethnic group	Frequency	Percent
1 White group	4462	94.0
2 Mixed group	28	0.6
3 Asian group	129	2.7
4 Black Group	66	1.4
5 Other group	63	1.3
Total	4748	100.0

Appendix B

Additional Tables for the CCS and BHPS Datasets

Table B.1: Frequency distribution for gender (CCS linked dataset)

Gender	Frequency	Percent
1 Male	1766	10.9
2 Female	3523	21.7
3 Mixed	10976	67.5
Total	16265	100.0

Table B.2: Frequency distribution for age (CCS linked dataset)

Age group		Frequency	Percent
1	0 – 24	171	1.1
2	25 – 29	661	4.1
3	30 – 34	1118	6.9
4	35 – 39	1441	8.9
5	40 – 44	1618	9.9
6	45 – 49	1666	10.2
7	50 – 54	1559	9.6
8	55 – 59	1802	11.1
9	60 – 64	1373	8.4
10	65 – 69	1252	7.7
11	70 and older	3604	22.2
Total		16265	100.0

Table B.3: Frequency distribution for gender (BHPS dataset)

Gender	Frequency	Percent
1 Male	3017	45.7
2 Female	3579	54.3
Total	6596	100.0

Table B.4: Frequency distribution for age (BHPS dataset)

Age	Frequency	Percent
1 16 – 34	1003	15.2
2 35 – 49	2116	32.1
3 50 +	3477	52.7
Total	6596	100.0

Table B.5: Frequency distribution for ethnicity (BHPS dataset)

Ethnic group	Frequency	Percent
1 White	6392	96.9
2 Nonwhite	204	3.1
Total	6596	100.0

Appendix C

Cross–tabulations of CCS Variables

Table C.1: Cross tabulation of jobs in household and response indicator (CCS linked dataset)

		Jobs in household		
Response indicator		No job	At least 1 job	Total
Nonresponse	Count	2470	2271	4741
	%	52.1	47.9	100.0
Response	Count	4422	7102	11524
	%	38.4	61.6	100.0
Total	Count	6892	9373	16265
	%	42.4	57.6	100.0

Table C.2: Cross tabulation of jobs in household and gender (CCS linked dataset)

		Jobs in household		
Gender		No job	At least 1 job	Total
Male	Count	888	878	1766
	%	50.3	49.7	100.0
Female	Count	2542	981	3523
	%	72.2	27.8	100.0
Mixed	Count	3462	7514	10976
	%	31.5	68.5	100.0
Total	Count	6892	9373	16265
	%	42.4	57.6	100.0

Table C.3: Cross tabulation of jobs in household and marital status (CCS linked dataset)

		Jobs in household		
Marital status		No job	At least 1 job	Total
Not married	Count	644	1952	2596
	%	24.8	75.2	100.0
Married	Count	3396	6228	9624
	%	35.3	64.7	100.0
Widowed	Count	2221	160	2381
	%	93.3	6.7	100.0
Divorced	Count	523	656	1179
	%	44.4	55.6	100.0
Registered partner	Count	10	42	52
	%	19.2	80.8	100.0
Multiple household	Count	98	335	433
	%	22.6	77.4	100.0
Total	Count	6892	9373	16265
	%	42.4	57.6	100.0

Table C.4: Cross tabulation of jobs in household and age (CCS linked dataset)

		Jobs in household		
Age		No job	At least 1 job	Total
0 – 24	Count	39	132	171
	%	22.8	77.2	100.0
25 – 29	Count	20	641	661
	%	3.0	97.0	100.0
30 – 34	Count	69	1049	1118
	%	6.2	93.8	100.0
35 – 39	Count	132	1309	1441
	%	9.2	90.8	100.0
40 – 44	Count	167	1451	1618
	%	10.3	89.7	100.0
45 – 49	Count	189	1477	1666
	%	11.3	88.7	100.0
50 – 54	Count	243	1316	1559
	%	15.6	84.4	100.0
55 – 59	Count	494	1308	1802
	%	27.4	72.6	100.0
60 – 64	Count	890	483	1373
	%	64.8	35.2	100.0
65 – 69	Count	1100	152	1252
	%	87.9	12.1	100.0
70 and older	Count	3549	55	3604
	%	98.5	1.5	100.0
Total	Count	6892	9373	16265
	%	42.4	57.6	100.0

Table C.5: Cross tabulation of type of housing and response indicator (CCS linked dataset)

		Type of housing		
Response indicator		Rental	Owned	Total
Nonresponse	Count	2155	2586	4741
	%	45.5	54.5	100.0
Response	Count	4061	7463	11524
	%	35.2	64.8	100.0
Total	Count	6216	10049	16265
	%	38.2	61.8	100.0

Table C.6: Cross tabulation of type of housing and gender (CCS linked dataset)

		Type of housing		
Gender		Rental	Owned	Total
Male	Count	927	839	1766
	%	52.5	47.5	100.0
Female	Count	2308	1215	3523
	%	65.5	34.5	100.0
Mixed	Count	2981	7995	10976
	%	27.2	72.8	100.0
Total	Count	6216	10049	16265
	%	38.2	61.8	100.0

Table C.7: Cross tabulation of type of housing and marital status (CCS linked dataset)

		Type of housing		
Marital status		Rental	Owned	Total
Not married	Count	1209	1387	2596
	%	46.6	53.4	100.0
Married	Count	2584	7040	9624
	%	26.8	73.2	100.0
Widowed	Count	1548	833	2381
	%	65.0	35.0	100.0
Divorced	Count	714	465	1179
	%	60.6	39.4	100.0
Registered partner	Count	13	39	52
	%	25.0	75.0	100.0
Multiple household	Count	148	285	433
	%	34.2	65.8	100.0
Total	Count	6216	10049	16265
	%	38.2	61.8	100.0

Table C.8: Cross tabulation of type of housing and age (CCS linked dataset)

		Type of housing		
Age		Rental	Owned	Total
0 – 24	Count	100	71	171
	%	58.5	41.5	100.0
25 – 29	Count	234	427	661
	%	35.4	64.6	100.0
30 – 34	Count	337	781	1118
	%	30.1	69.9	100.0
35 – 39	Count	357	1084	1441
	%	24.8	75.2	100.0
40 – 44	Count	394	1224	1618
	%	24.4	75.6	100.0
45 – 49	Count	459	1207	1666
	%	27.6	72.4	100.0
50 – 54	Count	434	1125	1559
	%	27.8	72.2	100.0
55 –59	Count	582	1220	1802
	%	32.3	67.7	100.0
60 – 64	Count	471	902	1373
	%	34.3	65.7	100.0
65 – 69	Count	582	670	1252
	%	46.5	53.5	100.0
70 and older	Count	2266	1338	3604
	%	62.9	37.1	100.0
Total	Count	6216	10049	16265
	%	38.2	61.8	100.0

Table C.9: Cross tabulation of size of household and response indicator (CCS linked dataset)

		Size of household		
Response indicator		2 or more	1	Total
Nonresponse	Count	3048	1693	4741
	%	64.3	35.7	100.0
Response	Count	8748	2776	11524
	%	75.9	24.1	100.0
Total	Count	11796	4469	16265
	%	72.5	27.5	100.0

Table C.10: Cross tabulation of size of household and gender (CCS linked dataset)

		Size of household		
Gender		2 or more	1	Total
Male	Count	244	1522	1766
	%	13.8	86.2	100.0
Female	Count	649	2874	3523
	%	18.4	81.6	100.0
Mixed	Count	10903	73	10976
	%	99.3	0.7	100.0
Total	Count	11796	4469	16265
	%	72.5	27.5	100.0

Table C.11: Cross tabulation of size of household and marital status (CCS linked dataset)

		Size of household		
Marital status		2 or more	1	Total
Not married	Count	1166	1430	2596
	%	44.9	55.1	100.0
Married	Count	9457	167	9624
	%	98.3	1.7	100.0
Widowed	Count	262	2119	2381
	%	11.0	89.0	100.0
Divorced	Count	458	721	1179
	%	38.8	61.2	100.0
Registered partner	Count	47	5	52
	%	90.4	9.6	100.0
Multiple household	Count	406	27	433
	%	93.8	6.2	100.0
Total	Count	11796	4469	16265
	%	72.5	27.5	100.0

Table C.12: Cross tabulation of size of household and age (CCS linked dataset)

		Size of household		
Age		2 or more	1	Total
0 – 24	Count	124	47	171
	%	72.5	27.5	100.0
25 – 29	Count	540	121	661
	%	81.7	18.3	100.0
30 – 34	Count	881	237	1118
	%	78.8	21.2	100.0
35 – 39	Count	1228	213	1441
	%	85.2	14.8	100.0
40 – 44	Count	1401	217	1618
	%	86.6	13.4	100.0
45 – 49	Count	1448	218	1666
	%	86.9	13.1	100.0
50 – 54	Count	1310	249	1559
	%	84.0	16.0	100.0
55 – 59	Count	1449	353	1802
	%	80.4	19.6	100.0
60 – 64	Count	1035	338	1373
	%	75.4	24.6	100.0
65 – 69	Count	887	365	1252
	%	70.8	29.2	100.0
70 and older	Count	1493	2111	3604
	%	41.4	58.6	100.0
Total	Count	11796	4469	16265
	%	72.5	27.5	100.0

Table C.13: Cross tabulation of mother working and response indicator (BHPS dataset)

		Mother working indicator		
Response indicator		Not working	Working	Total
Nonresponse	Count	357	268	625
	%	57.1	42.9	100.0
Response	Count	3275	2696	5971
	%	54.8	45.2	100.0
Total	Count	3632	2964	6596
	%	55.1	55.1	100.0

Table C.14: Cross tabulation of mother working and response indicator (BHPS dataset)

		Mother working indicator		
Gender		Not working	Working	Total
Male	Count	1679	1338	3017
	%	55.7	44.3	100.0
Female	Count	1953	1626	3579
	%	54.6	45.4	100.0
Total	Count	3632	2964	6596
	%	55.1	44.9	100.0

Table C.15: Cross tabulation of mother working and age (BHPS dataset)

		Mother working indicator		
Age		Not working	Working	Total
< 35	Count	353	650	1003
	%	35.2	64.8	100.0
35 – 49	Count	889	1227	2116
	%	42.0	58.0	100.0
50+	Count	2390	1087	3477
	%	68.7	31.3	100.0
Total	Count	3632	2964	6596
	%	55.1	44.9	100.0

Table C.16: Cross tabulation of mother working and ethnicity (BHPS dataset)

		Mother working indicator		
Ethnicity		Not working	Working	Total
White	Count	3501	2891	6392
	%	54.8	45.2	100.0
Nonwhite	Count	131	73	204
	%	64.2	35.8	100.0
Total	Count	3632	2964	6596
	%	55.1	44.9	100.0

Table C.17: Cross tabulation of father working and response indicator (BHPS dataset)

		Father working indicator		
Response indicator		Not working	Working	Total
Nonresponse	Count	69	556	625
	%	11.0	89.0	100.0
Response	Count	511	5460	5971
	%	8.6	91.4	100.0
Total	Count	580	6016	6596
	%	8.8	91.2	100.0

Table C.18: Cross tabulation of father working and gender (BHPS dataset)

		Father working indicator		
Gender		Not working	Working	Total
Male	Count	258	2759	3017
	%	8.6	91.4	100.0
Female	Count	322	3257	3579
	%	9.0	91.0	100.0
Total	Count	580	6016	6596
	%	8.8	91.2	100.0

Table C.19: Cross tabulation of father working and age (BHPS dataset)

		Father working indicator		
Age		Not working	Working	Total
< 35	Count	100	903	1003
	%	10.0	90.0	100.0
35 – 49	Count	127	1989	2116
	%	6.0	94.0	100.0
50+	Count	353	3124	3477
	%	10.2	89.8	100.0
Total	Count	580	6016	6596
	%	8.8	91.2	100.0

Table C.20: Cross tabulation of father working and ethnicity (BHPS dataset)

		Father working indicator		
Ethnicity		Not working	Working	Total
White	Count	543	5849	6392
	%	8.5	91.5	100.0
Nonwhite	Count	37	167	204
	%	18.1	81.9	100.0
Total	Count	580	6016	6596
	%	8.8	91.2	100.0

Appendix D

Additional Table for Chapter 3

Table D.1: Parameter estimates for the model for the size of household

Fixed effect	Model S1	Model S2	Model S3	Model S4	Model S5	Model S6
Constant	-0.588 (0.030) **	-0.722 (0.174) **	-0.586 (0.032) **	-0.586 (0.033) **	-0.723 (0.179) **	-0.728 (0.251) **
Response indicator						
Response	-0.560 (0.037) **	-0.334 (0.041) **	-0.561 (0.037) **	-0.564 (0.047) **	-0.335 (0.041) **	-0.337 (0.048) *
Age						
25 – 29		-0.526 (0.199) **			-0.530 (0.203) **	-0.526 (0.201) **
30 – 34		-0.341 (0.187) *			-0.343 (0.191) *	-0.339 (0.188) *
35 – 39		-0.779 (0.187) **			-0.780 (0.191) **	-0.775 (0.188) **
40 – 44		-0.892 (0.187) **			-0.895 (0.191) **	-0.891 (0.188) **
45 – 49		-0.930 (0.186) **			-0.933 (0.190) **	-0.926 (0.187) **
50 – 54		-0.704 (0.185) **			-0.706 (0.189) **	-0.698 (0.185) **
55 – 59		-0.448 (0.182) **			-0.449 (0.186) **	-0.442 (0.182) **
60 – 64		-0.158 (0.183)			-0.158 (0.187)	-0.154 (0.184)
65 – 69		0.070 (0.183)			0.070 (0.187)	0.074 (0.183)
70+		1.265 (0.175) **			1.266 (0.180) **	1.273 (0.175) **
Random effect						
Random intercept:						
Interviewer variance $\sigma^2_{u_0}$			0.004 (0.003)	0.006 (0.005)	0.004 (0.003)	0.009 (0.007)
Random coefficient:						
Interviewer coef. variance $\sigma^2_{u_1}$				0.031 (0.013)		0.025 (0.014)
Interv. inter.-coef. covariance $\sigma_{u_0 u_1}$				-0.004 (0.007)		-0.010 (0.009)

The base categories for the explanatory variables are Nonresponse and 0 – 24.
Model S1 is the standard (single-level) logistic model with only the response indicator as the explanatory variable, Model S2 is the single-level logistic model with additional household-level characteristics (explanatory variables), Model S3 is the two-level logistic random intercept model with only the response indicator as the explanatory variable, Model S4 is the two-level logistic random coefficient model with only the response indicator as the explanatory variable, Model S5 is the two-level logistic random intercept model controlling for other explanatory variables and Model S6 is the two-level logistic random coefficient model controlling for other explanatory variables.
** Significant at the 5% level; * Significant at the 10% level

Appendix E

Map of Norwegian Geographical Units (areas)

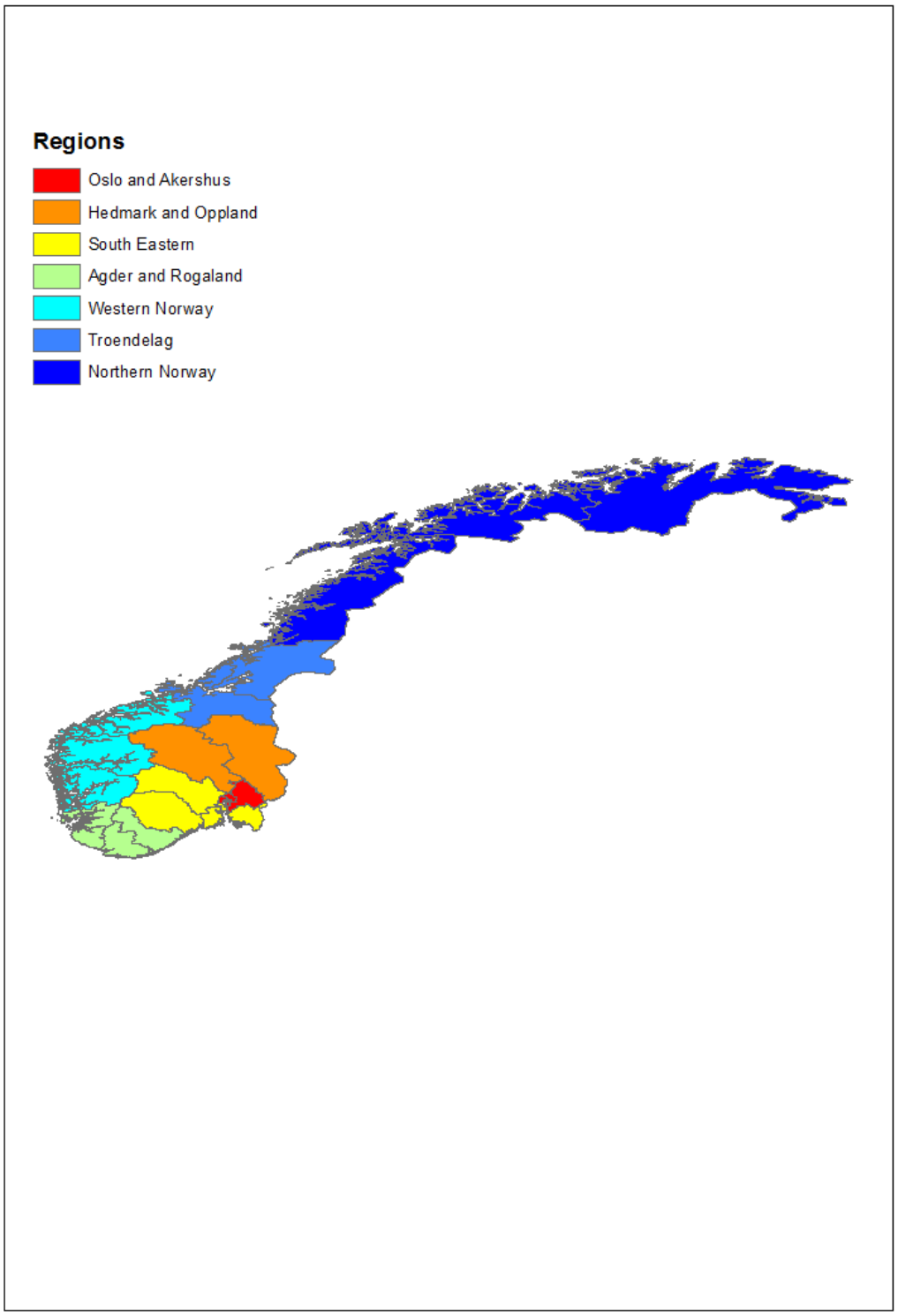


Figure E.1: Map of Norwegian geographical units (areas)

Appendix F

Distributions of ESS–Norway Explanatory Variables

Table F.1: Frequency distribution for gender (ESS–Norway)

Gender	Frequency	Percent
Male	803	52.1
Female	737	47.9
Total	1540	100.0

Table F.2: Frequency distribution for age (ESS–Norway)

Age	Frequency	Percent
15 - 30	364	23.6
31 - 66	950	61.7
Over 66	226	14.7
Total	1540	100.0

Table F.3: Frequency distribution for area (ESS–Norway)

Area	Frequency	Percent
Oslo and Akershus	314	20.4
Hedmark and Oppland	110	7.1
South Eastern	296	19.2
Agder and Rogaland	255	16.6
Western Norway	271	17.6
Troendelag	152	9.9
Northern Norway	142	9.2
Total	1540	100.0

Appendix G

Additional Tables for Chapter 4

Table G.1: Coding for the cross-tabulation for education from ESS-Norway and from register with four categories

Education from survey	Education from register		
	Low	Middle	High
Low	1	4	4
Middle	4	2	4
High	4	4	3

Table G.2: Parameter estimates (with standard errors in parenthesis) of the random intercept multinomial model (empty model) for measurement error on Education

Fixed effects	$\log(\pi_{2jk}/\pi_{1jk})$	$\log(\pi_{3jk}/\pi_{1jk})$	$\log(\pi_{4jk}/\pi_{1jk})$
Constant	0.626 (0.104) **	0.641 (0.119) **	0.528 (0.105) **
Random effects			
$\sigma^2_{u_0^{(2)}}$	0.278 (0.084) **		
$\sigma^2_{u_0^{(3)}}$	0.576 (0.179) **		
$\sigma^2_{u_0^{(4)}}$	0.284 (0.090) **		
$\sigma_{u_0^{(2)}u_0^{(3)}}$	0.015 (0.089)		
$\sigma_{u_0^{(2)}u_0^{(4)}}$	0.059 (0.065)		
$\sigma_{u_0^{(3)}u_0^{(4)}}$	0.153 (0.104)		

* $p < 0.10$; ** $p < 0.05$

Table G.3: Parameter estimates (with standard errors in parenthesis) of the random intercept multinomial model for measurement error on Education controlling for explanatory variables

Fixed effects	Categories	$\log(\pi_{2jk}/\pi_{1jk})$	$\log(\pi_{3jk}/\pi_{1jk})$	$\log(\pi_{4jk}/\pi_{1jk})$
Constant		0.082 (0.279)	0.348 (0.312)	0.645 (0.272) **
Gender	Female	-0.598 (0.170) **	-0.146 (0.175)	-0.476 (0.171) **
Age group	31 – 66	1.522 (0.197) **	2.284 (0.215) **	1.214 (0.196) **
	Over 66	0.562 (0.238) **	0.449 (0.279)	0.236 (0.242)
Area	Hedmark and Oppland	-0.531 (0.418)	-1.162 (0.467) **	-0.295 (0.393)
	South Eastern	-0.010 (0.305)	-1.143 (0.337) **	-0.391 (0.301)
	Agder and Rogaland	0.097 (0.331)	-1.216 (0.394) **	-0.441 (0.332)
	Western Norway	0.443 (0.313)	-0.982 (0.360) **	-0.439 (0.322)
	Troendelag	0.061 (0.368)	-1.156 (0.417) **	-0.506 (0.372)
	Northern Norway	-0.413 (0.349)	-1.719 (0.412) **	-1.161 (0.364) **
Random effects				
$\sigma^2_{u_0^{(2)}}$	0.107 (0.057) *			
$\sigma^2_{u_0^{(3)}}$	0.461 (0.156) **			
$\sigma^2_{u_0^{(4)}}$	0.133 (0.065) **			
$\sigma_{u_0^{(2)}u_0^{(3)}}$	0.010 (0.072)			
$\sigma_{u_0^{(2)}u_0^{(4)}}$	0.044 (0.048)			
$\sigma_{u_0^{(3)}u_0^{(4)}}$	0.161 (0.087) *			

The baseline categories for the explanatory variables are Male, 15 - 30 and Oslo and Akershus.
* $p < 0.10$; ** $p < 0.05$

Table G.4: Coding for the cross–tabulation for household size from ESS-Norway and from register with four categories

Household size	Household size from register		
from survey	1 person	2 people	3+ people
1 person	1	4	4
2 people	4	2	4
3+ people	4	4	3

Table G.5: Parameter estimates (with standard errors in parenthesis) of the random intercept multinomial model (empty model) for measurement error on Household Size

Fixed effects	$\log(\pi_{2jk}/\pi_{1jk})$	$\log(\pi_{3jk}/\pi_{1jk})$	$\log(\pi_{4jk}/\pi_{1jk})$
Constant	0.487 (0.110) **	0.832 (0.101) **	0.596 (0.103) **
Random effects			
$\sigma^2_{u_0^{(2)}}$	0.383 (0.128) **		
$\sigma^2_{u_0^{(3)}}$	0.276 (0.085) **		
$\sigma^2_{u_0^{(4)}}$	0.249 (0.075) **		
$\sigma_{u_0^{(2)}u_0^{(3)}}$	0.098 (0.081)		
$\sigma_{u_0^{(2)}u_0^{(4)}}$	0.093 (0.078)		
$\sigma_{u_0^{(3)}u_0^{(4)}}$	0.062 (0.061)		

* $p < 0.10$; ** $p < 0.05$

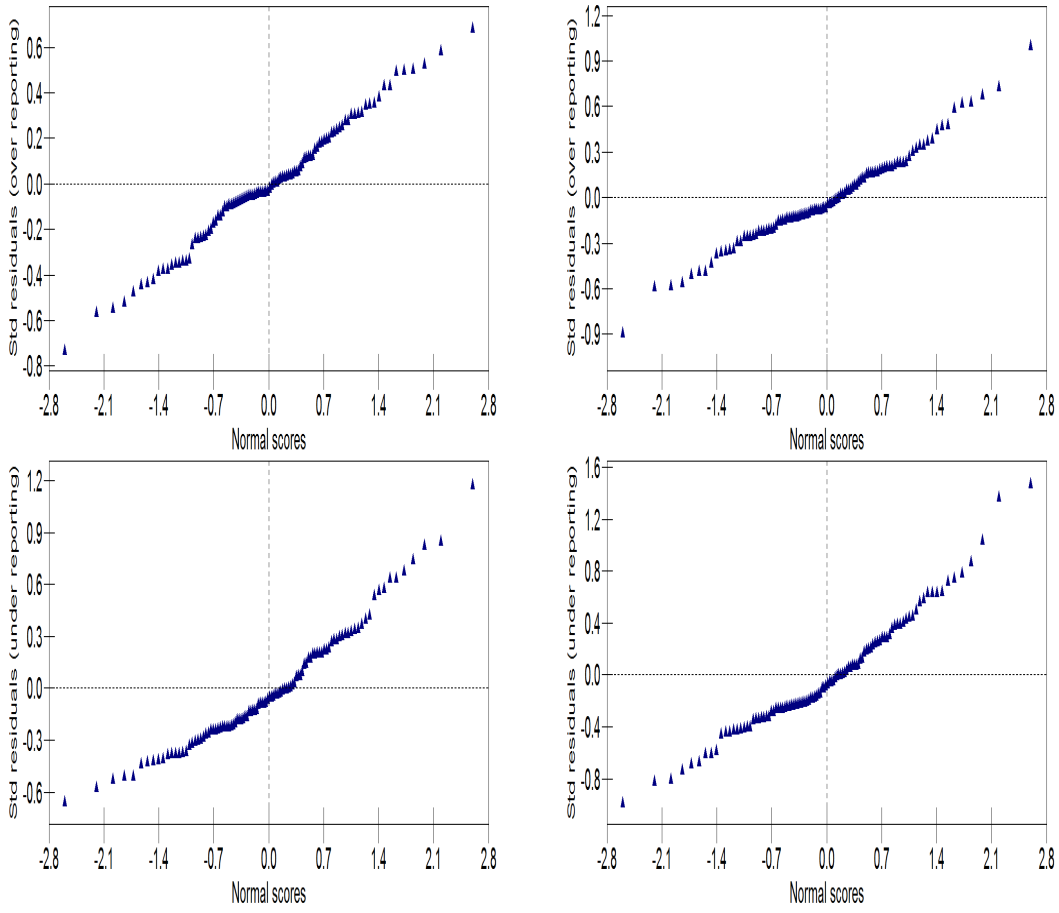
Table G.6: Parameter estimates (with standard errors in parenthesis) of the random intercept multinomial model for measurement error on Household Size controlling for explanatory variables

Fixed effects	Categories	$\log(\pi_{2jk}/\pi_{1jk})$	$\log(\pi_{3jk}/\pi_{1jk})$	$\log(\pi_{4jk}/\pi_{1jk})$
Constant		-0.611 (0.354) *	1.236 (0.286) **	1.632 (0.280) **
Gender	Female	-0.112 (0.174)	0.223 (0.173)	-0.516 (0.178) **
Age group	31 – 66	0.939 (0.303) **	-0.558 (0.223) **	-0.794 (0.229) **
	Over 66	0.750 (0.320) **	-6.708 (1.281) **	-3.168 (0.335) **
Area	Hedmark and Oppland	0.323 (0.405)	-0.233 (0.402)	0.046 (0.379)
	South Eastern	0.605 (0.313) *	0.506 (0.301) *	0.467 (0.299)
	Agder and Rogaland	-0.048 (0.353)	0.465 (0.311)	-0.189 (0.314)
	Western Norway	0.599 (0.342) *	0.798 (0.323) **	0.485 (0.318)
	Troendelag	0.297 (0.374)	0.191 (0.357)	-0.001 (0.350)
	Northern Norway	1.286 (0.429) **	0.841 (0.419) **	0.808 (0.412) **
Random effects				
$\sigma^2_{u_0^{(2)}}$	0.255 (0.121) **			
$\sigma^2_{u_0^{(3)}}$	0.159 (0.078) **			
$\sigma^2_{u_0^{(4)}}$	0.096 (0.053) *			
$\sigma_{u_0^{(2)}u_0^{(3)}}$	0.105 (0.081)			
$\sigma_{u_0^{(2)}u_0^{(4)}}$	0.071 (0.065)			
$\sigma_{u_0^{(3)}u_0^{(4)}}$	0.053 (0.051)			

The baseline categories for the explanatory variables are Male, 15 - 30 and Oslo and Akershus.
* $p < 0.10$; ** $p < 0.05$

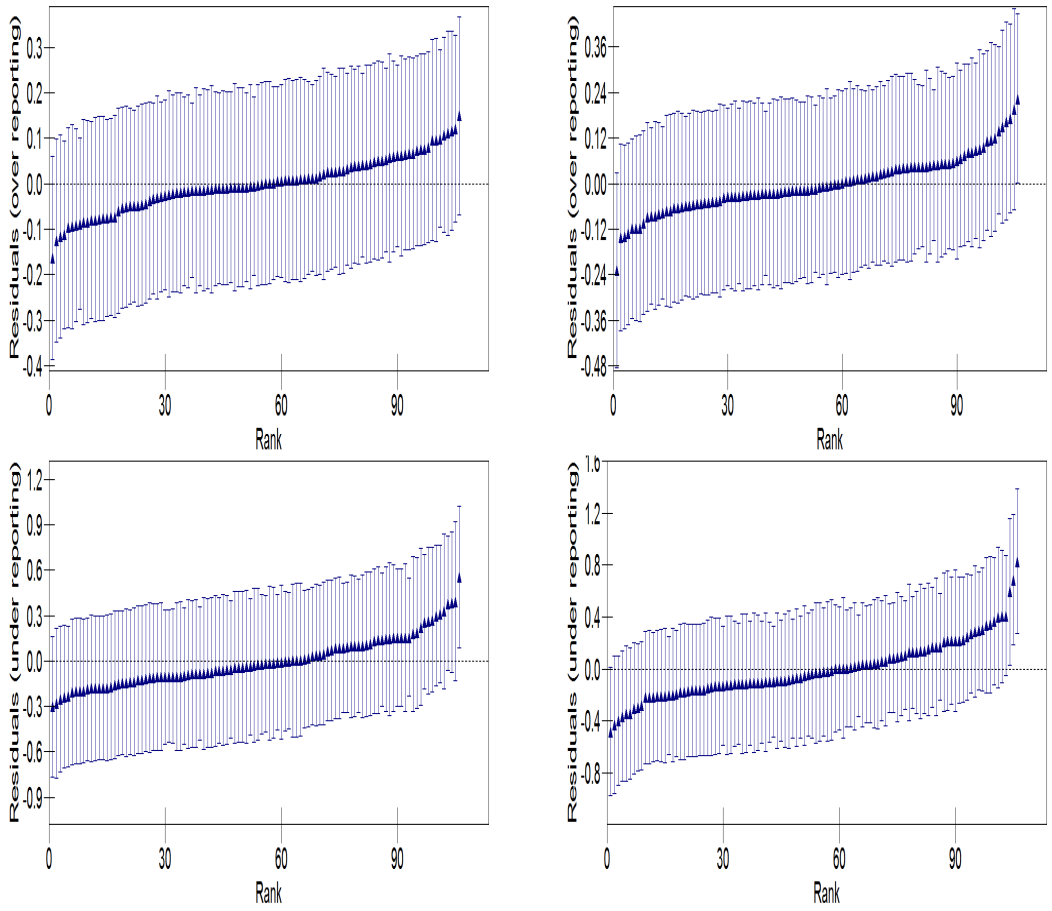
Appendix H

Additional Plots for Chapter 4



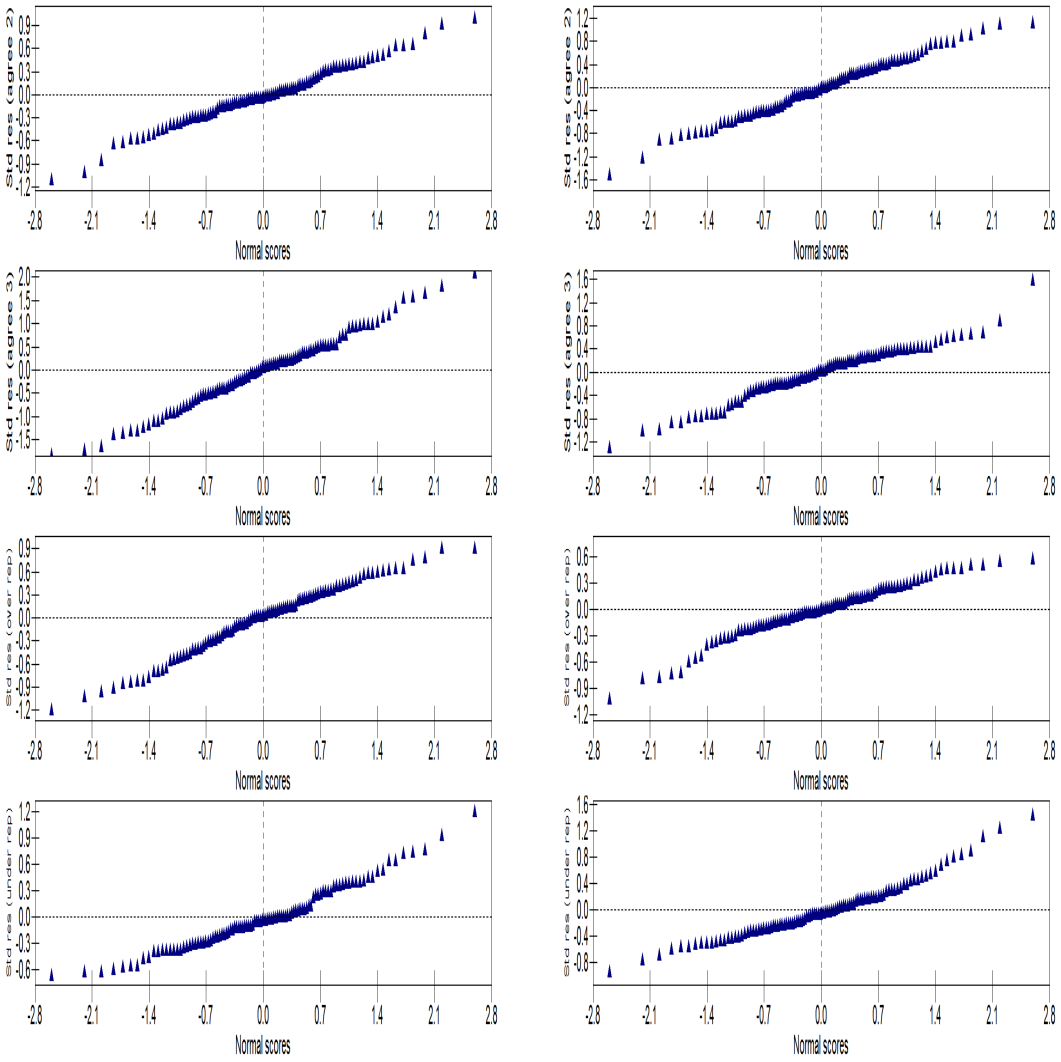
(a) Education (b) Household size

Figure H.1: Normal probability plots for the standardized residuals for the three category measurement error model for reporting education and household size



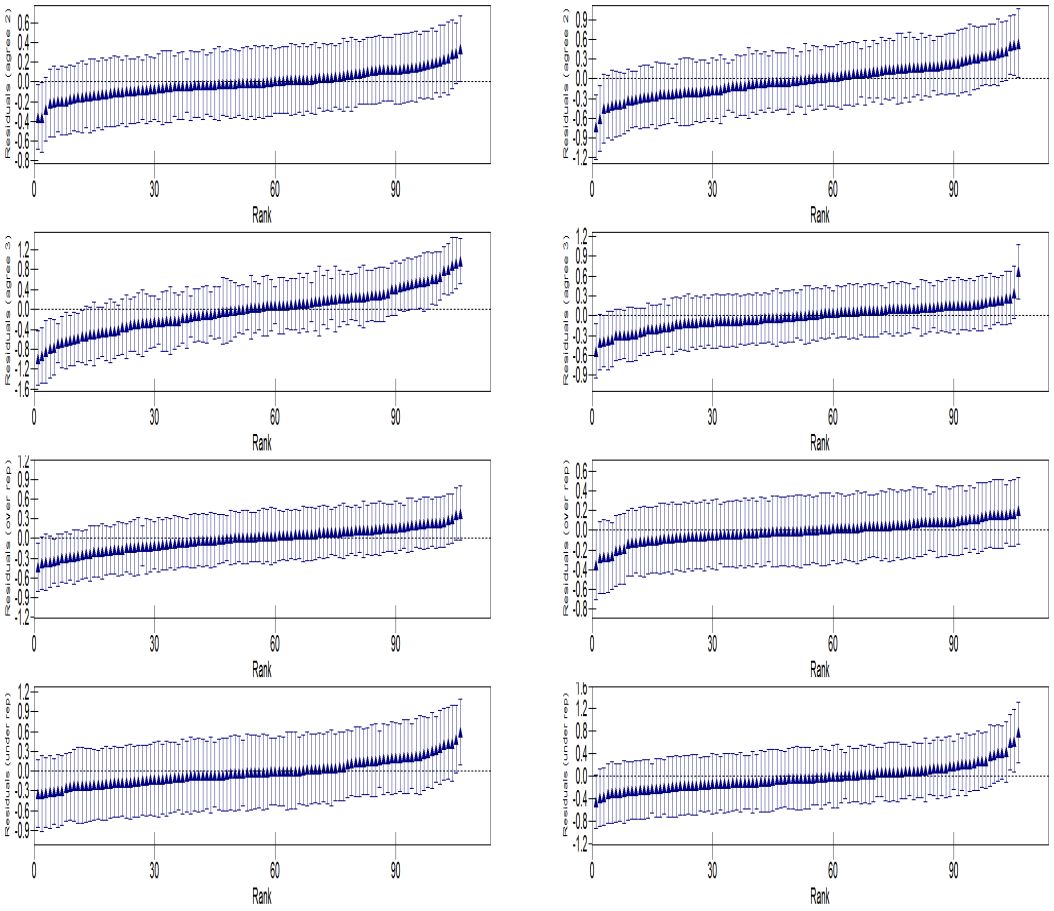
(a) Education (b) Household size

Figure H.2: Caterpillar plots for the three category measurement error model for reporting education and household size



(a) Education (b) Household size

Figure H.3: Normal probability plots for the standardized residuals for the five category measurement error model for reporting education and household size



(a) Education (b) Household size

Figure H.4: Caterpillar plots for the five category measurement error model for reporting education and household size

References

Afshartous, D. and Wolf, M. (2007). Avoiding 'data snooping' in multilevel and mixed effects models. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, 170(4):1035–1059.

Agresti, A. (2013). *Categorical Data Analysis*. 3rd ed. Wiley, Hoboken, NJ.

Aitkin, M., Anderson, D., and Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society. Series A (General)*, 144(4):148–161.

Aitkin, M. and Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society. Series A (General)*, 149(1):1–43.

Beaussart, M. L. and Kaufman, J. C. (2013). Gender differences and the effects of perceived internet privacy on self-reports of sexual behavior and sociosexuality. *Computers in Human Behavior*, 29(6):2524–2529.

Bethlehem, J. (2002). Weighting nonresponse adjustments based on auxiliary information. In Groves, R. M., Dillman, D. A., Eltinge, J. L., and Little, R. J. A., (eds.), *Survey Nonresponse*, page 275 288. Wiley, New York.

Bethlehem, J., Cobben, F., and Schouten, B. (2011). *Handbook of Nonresponse in Household Surveys*. Wiley, Hoboken, NJ.

Biemer, P. P. (2001). Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing. *Journal of Official Statistics*, 17(2):295–320.

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, 74(5):817–848.

Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., and Sudman, S., (eds.) (1991). *Measurement Errors in Surveys*. Wiley, New York.

Biemer, P. P. and Lyberg, L. (2003). *Introduction to survey quality*. Wiley, Hoboken, NJ.

Blom, A. G., de Leeuw, E. D., and Hox, J. J. (2011). Interviewer effects of nonresponse in the European social survey. *Journal of Official Statistics*, 27(2):359–377.

Bradburn, N., Sudman, S., and Blair, E. (1979). *Improving Interview Method and Questionnaire Design*. National Opinion Research Center: Series in social research. Jossey-Bass, Incorporated, San Francisco, CA.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25.

Browne, W. J. and Draper, D. (2006). A comparison of Bayesian and likehood-based methods for fitting mutilevel models. *Bayesian Analysis*, 1(3):473–514.

Brunton-Smith, I., Sturgis, P., and Williams, J. (2012). Is Success in Obtaining Contact and Cooperation Correlated With the Magnitude of Interviewer Variance? *Public Opinion Quarterly*, 76(2):265–286.

Buonaccorsi, J. P. (2010). *Measurement Error: Models, Methods, and Applications*. Chapman and Hall/CRC, Boca Raton, FL.

Burstein, L., Linn, R. L., and Capell, F. J. (1978). Analyzing multilevel data in the presence of heterogeneous within-class regressions. *Journal of Educational Statistics*, 3:347–383.

Campanelli, P. and O'Muircheartaigh, C. (1999). Interviewers, interviewer continuity, and panel survey nonresponse. *Quality and Quantity*, 33(1):59–76.

Campanelli, P., Sturgis, P., and Purdon, S. (1997). *Can You Hear Me Knocking?: An Investigation Into the Impact of Interviewers on Survey Response Rates*. Social and Community Planning Research, London.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. 2nd ed. Chapman and Hall, Boca Raton, FL.

Cochran, W. (1977). *Sampling techniques*. 3rd ed. Wiley, New York.

Cotter, P. R., Cohen, J., and Coulter, P. B. (1982). Race-of-interviewer effects in telephone interviews. *The Public Opinion Quarterly*, 46(2):278–284.

David, M. H., Little, R., Samuhel, M., and Triest, R. (1983). Imputation models based on the propensity to respond. In *ASA Proceedings of the Business and Economic Statistics Section*, pages 168–173. American Statistical Association.

Davis, D. W. (1997). Nonrandom Measurement Error and Race of Interviewer Effects Among African Americans. *The Public Opinion Quarterly*, 61(1):183–207.

Davis, J. A., Spaeth, J. L., and Huson, C. (1961). A technique for analyzing the effects of group composition. *American Sociological Review*, 26:215–225.

Davis, R. E., Couper, M. P., Janz, N. K., Caldwell, C. H., and Resnicow, K. (2010). Interviewer effects in public health surveys. *Health Education Research*, 25(1):14–26.

De Leeuw, E. D. and Hox, J. J. (2004). I am not selling anything: 29 experiments in telephone introductions. *International Journal of Public Opinion Research*, 16(4):464–473.

Dixon, J. (2010). Assessing nonresponse bias and measurement error using statistical matching. In *ASA Proceedings of the Section on Survey Research Methods*, pages 3388–3396. American Statistical Association.

Durrant, G. B., Groves, R. M., Staetsky, L., and Steele, F. (2010). Effects of Interviewer Attitudes and Behaviors on Refusal in Household Surveys. *Public Opinion Quarterly*, 74(1):1–36.

Durrant, G. B. and Steele, F. (2009). Multilevel modelling of refusal and non-contact in household surveys: Evidence from six UK Government surveys. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, 172(2):361–381.

Flores-Macias, F. and Lawson, C. (2008). Effects of interviewer gender on survey responses: Findings from a household survey in mexico. *International Journal of Public Opinion Research*, 20(1):100–110.

Freeth, S., Kane, C., and Cowie, A. (2002). Survey interviewer attitudes and demographic profile, preliminary results from the 2001 ONS interviewer attitudes survey. In *Office for National Statistics working paper*, pages 1–18, London.

Fricker, S. and Tourangeau, R. (2010). Examining the relationship between nonresponse propensity and data quality in two national household surveys. *Public Opinion Quarterly*, 74(5):934–955.

Fuller, W. A. (1987). *Measurement Error Models*. Wiley, New York.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical methods for social research. Cambridge University Press, Cambridge.

Gideon, L. (2012). *Handbook of Survey Methodology for the Social Sciences*. Springer, New York.

Goldstein, H. (1987). *Multilevel Models in Educational and Social Research*. Griffin, London.

Goldstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data. *Biometrika*, 78:45–51.

Goldstein, H. (1995). *Multilevel Statistical Models*. 2nd ed. Edward Arnold Publishers Ltd, London.

Goldstein, H. (2003). *Multilevel Statistical Models*. 3rd ed. Edward Arnold Publishers Ltd, London.

Goldstein, H. (2011). *Multilevel Statistical Models*. 4th ed. Wiley, Chichester, UK.

Goldstein, H. and Healy, M. J. R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, 158:175–177.

Goldstein, H. and Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, 159:505–513.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review*, 55:245–259.

Groves, R. M. (2004). *Survey Errors and Survey Costs*. 2nd ed. Wiley, Hoboken, NJ.

Groves, R. M. (2006). Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly*, 70(5):646–675.

Groves, R. M. and Couper, M. (1998). *Nonresponse in Household Interview Surveys*. Wiley, New York.

Groves, R. M. and Fultz, N. H. (1985). Gender effects among telephone interviewers in a survey of economic attitudes. *Sociological Methods & Research*, 14(1):31–52.

Groves, R. M. and Magilavy, L. J. (1986). Measuring and explaining interviewer effects in centralized telephone surveys. *Public Opinion Quarterly*, 50(2):251–266.

Groves, R. M. and Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly*, 72(2):167–189.

Groves, R. M., Presser, S., Tourangeau, R., West, B. T., Couper, M. P., Singer, E., and Toppe, C. (2012). Support for the survey sponsor and nonresponse bias. *Public Opinion Quarterly*, 76(3):512–524.

Groves, R. M., Singer, E., and Corning, A. (2000). Leverage-saliency theory of survey participation: Description and an illustration. *Public Opinion Quarterly*, 64(3):299–308.

Hatchett, S. and Schuman, H. (1975). White respondents and race-of-interviewer effects. *Public Opinion Quarterly*, 39(4):523–527.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70.

Hox, J. and De Leeuw, E. (2002). The influence of interviewers attitude and behavior on household survey nonresponse: An international comparison. In

Groves, R. M., Dillman, D. A., Eltinge, J. L., and Little, R. J. A., (eds.), *Survey Nonresponse*, pages 103–120. Wiley, New York.

Hox, J. J. (2010). *Multilevel Analysis: Techniques and Applications*. Routledge, New York.

Jäckle, A., Lynn, P., Sinibaldi, J., and Tipping, S. (2013). The effect of interviewer experience, attitudes, personality and skills on respondent co-operation with face-to-face surveys. *Survey Research Methods*, 7(1):1–15.

Kane, E. W. and Macaulay, L. J. (1993). Interviewer gender and gender attitudes. *Public Opinion Quarterly*, 57(1):1–28.

Keeter, S., Kennedy, C., Dimock, M., Best, J., and Craighill, P. (2006). Gauging the impact of growing nonresponse on estimates from a national rdd telephone survey. *Public Opinion Quarterly*, 70(5):759–779.

Keeter, S., Miller, C., Kohut, A., Groves, R. M., and Presser, S. (2000). Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Quarterly*, 64(2):125–148.

Kennickell, A. (2002). Interviewers and data quality: Evidence from the 2001 Survey of Consumer Finances. In *ASA Proceedings of the Joint Statistical Meetings*, pages 1807–1812. American Statistical Association.

Kreuter, F., Müller, G., and Trappmann, M. (2010). Nonresponse and measurement error in employment research: Making use of administrative data. *Public Opinion Quarterly*, 74(5):880–906.

Lepkowski, J. (2004). Non-observation error in household surveys in developing countries. In *An Analysis of Operating Characteristics of Household Surveys in Developing and Transition Countries: Survey Costs, Design Effects and Non-Sampling Errors*, pages 171–198. United Nations.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society, Series B: Methodological*, 34:1–41.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88:125–134.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data*. 2nd ed. Wiley, Chichester.

Loosveldt, G. and Beullens, K. (2014). A procedure to assess interviewer effects on nonresponse bias. *SAGE Open*, 4(1):1–12.

Lynn, P., Beerten, R., Laiho, J., and Martin, J. (2002). Towards standardisation of survey outcome categories and response rates calculation. *Research in Official Statistics*, 1:63–86.

Maas, C. J. M. and Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, 1(3):86–92.

Mahalanobis, P. C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109:325–378.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. 2nd ed. Chapman & Hall, London.

Mensch, B. S. and Kandel, D. B. (1988). Underreporting of substance use in a national longitudinal youth cohort: Individual and interviewer effects. *Public Opinion Quarterly*, 52(1):100–124.

Merkle, D. and Edelman, M. (2002). Nonresponse in exit polls: A comprehensive analysis. In Groves, R. M., Dillman, D. A., Eltinge, J. L., and Little, R. J. A., (eds.), *Survey Nonresponse*, pages 243–258. Wiley, New York.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A, General*, 135:370–384.

Olson, K. (2006). Survey participation, nonresponse bias, measurement error bias, and total bias. *Public Opinion Quarterly*, 70(5):737–758.

Olson, K. and Kennedy, C. (2006). Examination of the relationship between nonresponse and measurement error in a validation study of alumni. In *ASA Proceedings of the Survey Research Methods Section*, pages 4181–4188. American Statistical Association.

Olson, K. and Parkhurst, B. (2013). Collecting paradata for measurement error evaluations. In Kreuter, F., (ed.), *Improving Surveys with Paradata: Analytic Uses of Process Information*, pages 43–72. Wiley, Hoboken, NJ.

O'Muircheartaigh, C. and Campanelli, P. (1999). A multilevel exploration of the role of interviewers in survey non-response. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(3):437–446.

Pickery, J. and Loosveldt, G. (2002). A multilevel multinomial analysis of interviewer effects on various components of unit nonresponse. *Quality and Quantity*, 36(4):427–437.

Pickery, J. and Loosveldt, G. (2004). A simultaneous analysis of interviewer effects on various data quality indicators with identification of exceptional interviewers. *Journal of Official Statistics*, 20(1):77–89.

Rasbash, J., Steele, F., Browne, W., and Goldstein, H. (2012). *A User's Guide to MLwiN*. v2.26. Centre for Multilevel Modelling, University of Bristol.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications Inc.

Reese, S. D., Danielson, W. A., Shoemaker, P. J., Chang, T.-K., and Hsu, H.-L. (1986). Ethnicity-of-interviewer effects among mexican-americans and anglos. *Public Opinion Quarterly*, 50(4):563–572.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3):351–357.

Romano, J. P. and Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Ruddock, V. (1998). Measuring and improving data quality. UK Government Statistical Service Methodology Series No. 14. `http://www.statistics.gov.uk/methodsquality/publications.asp`.

Sakshaug, J. W., Yan, T., and Tourangeau, R. (2010). Nonresponse error, measurement error, and mode of data collection: Tradeoffs in a multi-mode survey of sensitive and non-sensitive items. *Public Opinion Quarterly*, 74(5):907–933.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78:719–727.

Schouten, B., van den Brakel, J., Buelens, B., van der Laan, J., and Klausch, T. (2013). Disentangling mode-specific selection and measurement bias in social surveys. *Social Science Research*, 42(6):1555–1570.

Scott, A. and Davis, P. (2001). Estimating interviewer effects for binary responses. In *Proceedings of Statistics Canada Symposium*, pages 1–8, Canada.

Singer, E., Frankel, M. R., and Glassman, M. B. (1983). The effect of interviewer characteristics and expectations on response. *Public Opinion Quarterly*, 47(1):68–83.

Sinibaldi, J., Durrant, G. B., and Kreuter, F. (2013). Evaluating the measurement error of interviewer observed paradata. *Public Opinion Quarterly*, 77(S1):173–193.

Smith, A. F. M. (1973). A general Bayesian linear model. *Journal of the Royal Statistical Society, Series B: Methodological*, 35:67–75.

Smith, T. W. (2011). The report of the international workshop on using multi-level data from sample frames, auxiliary databases, paradata and related sources to detect and adjust for nonresponse bias in surveys. *International Journal of Public Opinion Research*, 23(3):389–402.

Snijders, T. A. B. and Bosker, R. J. (2012). *Multilevel Analysis: an Introduction to Basic and Advanced Multilevel Modeling*. Sage Publications Inc, London.

Snijkers, G., Hox, J., and de Leeuw, E. D. (1999). Interviewers' tactics for fighting survey nonresponse. *Journal of Official Statistics*, 15:185–198.

Tourangeau, R., Groves, R. M., and Redline, C. D. (2010). Sensitive topics and reluctant respondents: Demonstrating a link between nonresponse bias and measurement error. *Public Opinion Quarterly*, 74(3):413–432.

West, B. T., Kreuter, F., and Jaenichen, U. (2013). "Interviewer" effects in face-to-face surveys: A function of sampling, measurement error, or nonresponsive? *Journal of Official Statistics*, 29(2):277–297.

West, B. T. and Olson, K. (2010). How much of interviewer variance is really nonresponse error variance? *Public Opinion Quarterly*, 74(5):1004–1026.

White, A., Freeth, S., and Martin, J. (2001). Evaluation of survey data quality using matched census-survey records. In *International Conference on Quality in Official Statistics*, Stockholm, Sweden.