UNIVERSITY OF Southampton

University of Southampton Research Repository
ePrints Soton

Copyright © and Moral Rights for this thesis are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holders.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given e.g.

AUTHOR (year of submission) "Full thesis title", University of Southampton, name of the University School or Department, PhD Thesis, pagination

http://eprints.soton.ac.uk

UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL AND HUMAN SCIENCES

Social Statistics

New Perspectives on Interviewer-Related Error in Surveys: Application of Survey Paradata

by

Malgorzata Magdalena Turner

Thesis for the degree of Doctor of Philosophy

April 2013

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL AND HUMAN SCIENCES

Social Statistics

Thesis for the degree of Doctor of Philosophy

NEW PERSPECTIVES ON INTERVIEWER-RELATED ERROR IN SURVEYS: APPLICATION OF SURVEY PARADATA

Malgorzata Magdalena Turner

The interviewer is a well-recognised source of various survey errors. Primary amongst these is measurement error and this is the main focus of this thesis. The interviewer, as perhaps the key source of measurement error, can introduce both bias and variability into survey estimates. This is referred to in survey methodological literature as the interviewer effect. However, measuring the interviewer effect is difficult, mostly as a result of the types of complex survey designs that are widely used by survey agencies.

The purpose of this three-paper thesis is to utilise new developments in survey methodology to estimate the interviewer effect and gain more insight into which interviewer characteristics are linked to the interviewer effect. In particular, this programme of research draws on the availability and understanding of survey paradata, a new and potentially powerful form of data for survey methodologists. Chapter One provides an overview of the theory behind survey errors and explains what a design effect is. It summarises three main approaches to estimating interviewer effect. It describes the most recent developments with regard to paradata. The chapter concludes with the data description including the National Travel Survey (NTS).

In Chapter Two (Paper One), a cross-classified multilevel model is used to disentangle area and interviewer contributions to the design effect with paradata variables included at the interviewer level. The objective is to better understand how variance in a range of survey outcomes is related to the characteristics of individual interviewers such as satisfaction with pay, determination, NTS experience or conscientiousness.

Chapter Three (Paper Two) investigates the theory of satisficing, which states that some respondents provide only an 'acceptable' rather than an 'optimal' response, thereby negatively affecting data quality. Four hypotheses are posed: that interviewers' attitudes towards interviewing, their personality traits and skills, their demographic characteristics and finally their own satisficing behaviour can affect their respondents' propensity to satisfice. To test the hypothesis, a range of cross-classified multilevel ordered logit models are fitted.

According to the theory of satisficing, those who satisfice should take less time to answer survey questions because they use 'cognitive shortcuts' when formulating responses. In Chapter Four (Paper Three) the effect of both interviewer and respondent satisficing on the length of time taken to produce and record an answer to a number of different questions is evaluated, using auto-generated time stamps from the audit trail of the NTS. Chapter Five presents a summary of key findings, as well as outlying scope for further research in this area.

Contents

ABSTRACT ..... **i**

Contents ..... **i**

List of tables ..... **v**

List of figures ..... **vii**

DECLARATION OF AUTHORSHIP ..... **xi**

Acknowledgements ..... **xiii**

Definitions and abbreviations ..... **xv**

1. Introduction **1**

1.1 Survey errors 3
1.1.1 Measurement error 4
1.2 Survey design effect 6
1.3 Interviewer effect 8
1.3.1 Separating area and interviewer design effects 9
1.3.2 Interviewer-related variance 10
1.3.3 Interviewer effect on data quality indicators 13
1.4 Paradata 14
1.5 Data 19
1.5.1 The National Travel Survey 19
1.5.2 The survey data on interviewers 21
1.5.3 Area-level variables 22
1.6 Contribution of this thesis 23

2. Using paradata to explain the interviewer contribution to the design effect in face-to-face surveys **25**

2.1 Introduction 25
2.2 The interviewer design effect 26
2.3 Paradata 30
2.4 Data 31
2.5 Measures 32
2.6 Model 35
2.7 Results 38
2.7.1 Satisfaction with pay 43
2.7.2 Determination to obtain an interview 46
2.7.3 Interviewer experience 48

2.7.4 Communication skills 51
2.7.5 Conscientiousness 53
2.7.6 Extroversion 55
2.8 Discussion 58
3. Survey satisficing: how interviewer personality traits, attitudes and satisficing behaviour influence respondents' propensity to satisfice 63
3.1 Introduction 63
3.2 Satisficing: theory and measurement 64
3.3 Sources of satisficing 67
3.3.1 The respondent 67
3.3.2 The task/instrument 68
3.3.3 The interviewer 69
3.4 Data 71
3.5 Measures 72
3.6 Model specification 77
3.7 Results 80
3.7.1 Transport 81
3.7.2 Amenities 85
3.8 Discussion 89
4. False economy? Testing the hypothesis that satisficing response styles reduce cognitive effort in survey respondents 93
4.1 Introduction 93
4.2 Survey satisficing 94
4.3 Response latencies as paradata 96
4.4 Response latencies and satisficing 97
4.5 Data 99
4.6 Measures 99
4.7 Model specification 104
4.8 Results 105
4.9 Discussion 110
5. Conclusions 115
5.1 Measurement error in surveys 115
5.2 Key findings 116
5.2.1 Interviewer-related variance 117
5.2.2 Interviewers' effect on satisficing 117
5.2.3 Satisficing and response latencies 118
5.2.4 The value of paradata 118

5.2.5 Statistical modelling 120
5.2.6 MSOA-level variables 120
5.3 Limitations of the study 120
5.4 Implications for survey practice 122
5.5 Further research directions 124
Appendices 127
Appendix 1: Who is behind the NatCen ID card? 129
Appendix 2: Factor analyses 141
Appendix 3: Data acknowledgment 145
List of References 147

List of tables

Table 2.1. NTS survey items analysed in this study....34

Table 2.2. Standard errors (SE) under SRS and complex design and DEFF values for NTS survey items....40

Table 2.3. Variance partition coefficient (VPC) for the interviewer and area....41

Table 2.4. Variance partition coefficients for null models and models with covariates....42

Table 2.5. The 'railway station' model with the interviewer-related variance partitioned by the 'satisfaction with pay' variable....43

Table 3.1. Null models with complex variance component for 'transport' and 'amenities' variables....80

Table 3.2a. Transport models 3-5....82

Table 3.2b. Transport models 6 and 7....84

Table 3.3a. Amenities models 8-10....86

Table 3.3b. Amenities models 11 and 12....87

Table 4.1a. Multilevel models for the 'nearest bus stop' item....107

Table 4.1b. Statistical models for the 'nearest bus stop' item that include controls....108

Table 4.2. The effect of weak and strong satisficing on the response time for the remaining eight items....109

List of figures

Figure 1.1 Survey lifecycle from a quality perspective (Adapted from Groves *et al.*, 2009, p. 48) 5

Figure 2.1. A path model showing how a negative correlation between interviewer-induced nonresponse bias and measurement error might arise (source: Brunton-Smith *et al.*, 2012) 29

Figure 2.2. Diagrammatic representation of the data structure for this study 32

Figure 2.3. Crossed structure of respondents nested within interviewers and MSOAs (adapted from Rashbash and Browne, 2008, p. 303) 36

Figure 2.4. Proportion of the total variance introduced by both interviewer groups (satisfaction with pay) – duration items: 45

Figure 2.5. Proportion of the total variance introduced by both interviewer groups (satisfaction with pay) – satisfaction items: 46

Figure 2.6. Proportion of the total variance introduced by both interviewer groups (determination) – duration items: 47

Figure 2.7. Proportion of the total variance introduced by both interviewer groups (determination) – satisfaction items: 48

Figure 2.8 Proportion of the total variance introduced by four interviewer groups (NTS experience) – duration items: 49

Figure 2.9 Proportion of the total variance introduced by four interviewer groups (NTS experience) – satisfaction items: 50

Figure 2.10. Mean proportion of the total variance introduced by four interviewer groups (NTS experience): 51

Figure 2.11 Proportion of the total variance introduced by four interviewer groups (communication skills) – duration items: 52

Figure 2.12 Proportion of the total variance introduced by four interviewer groups (communication skills) – satisfaction items: 52

Figure 2.13. The mean proportion of the total variance introduced by four interviewer groups with different levels of communication skills: 53

Figure 2.14 Proportion of the total variance introduced by four interviewer groups (conscientiousness) – duration items: 54

Figure 2.15 Proportion of the total variance introduced by four interviewer groups (conscientiousness) – satisfaction items: 54

Figure 2.16. The mean proportion of the total variance introduced by four interviewer groups with different levels of conscientiousness: 55

Figure 2.17 Proportion of the total variance introduced by four interviewer groups (extroversion) – duration items: 56

Figure 2.18 Proportion of the total variance introduced by four interviewer groups (extroversion) – satisfaction items: 57

Figure 2.19. The mean proportion of the total variance introduced by four interviewer groups with different levels of extroversion: 58

Figure 3.1. Time taken to pick up and check the diaries (in minutes) 73

Figure 3.2. Distribution of the responses to the 'How long would it take to get to the nearest grocery store' item 75

Figure 3.3. Distribution of the responses to the 'How long would it take to get to the nearest hospital' item 75

Figure 3.4. Distribution of the dependent variables: transport and amenities 76

Figure 3.5. Predicted probabilities of a respondent being in the highest heaping group, for four different interviewer's heaping values (for female interviewer interviewing male respondent, who is employed full-time, white and married) 88

Figure 3.6. Predicted probabilities of a respondent being in the highest heaping group, for different levels of interviewer's agreement with the 'refusal should be accepted' item (for female interviewer interviewing male respondent, who is employed full-time, white and married) 89

Figure 4.1. Type of answer given to each survey item 101

Figure 4.2. Audit trail: how long did it take to place and explain the diary and other documents? Include any time spent preparing the diaries beforehand. Record to the nearest minute (N = 5422) 103

Figure 4.3. Reported by interviewers: how long did it take to place and explain the diary and other documents? Include any time spent preparing the diaries beforehand. Record to the nearest minute (N = 5422) 103

Figure 4.4. Median response times for each item and response type in seconds..... 105

Figure 4.5. Median times for the 'nearest bus stop' item for various truncated samples.. 110

DECLARATION OF AUTHORSHIP

I, Malgorzata Turner, declare that the thesis entitled' *New Perspectives on Interviewer-Related Error in Surveys: Application of Survey Paradata'* 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:……..

Acknowledgements

This study was made possible by funding from Economic and Social Research Council.

First and foremost, I am indebted to my fantastic supervisors – Professor Patrick Sturgis, Professor David Martin and Professor Chris Skinner – for their continued patience, support and invaluable guidance throughout these three-and-a-half years. I am grateful to them for keeping me motivated and moving in the right direction.

I would also like to thank NatCen who collected the National Travel Survey data, especially to Mrs Gerry Nicolaas for her help with the complicated datasets. I am grateful to NCRM for their support to attend conferences, especially to Ms Nicky Jackson, for her extraordinary organisational skills and willingness to help. I would also like to thank Mrs Chris Baker for helping me with administrative issues. I am especially grateful to Dr Ian Brunton-Smith for teaching me all about MLwiN.

Thanks are also due to my colleagues at the Student Data Management and Analysis team at the University of Oxford for their comforting words and especially to my line manager Dr Katy Holmes and to Ms Jackie Hoyle for their understanding and flexibility in the months leading to submission. I would like to acknowledge Ms Elizabeth Hufton for her help with proofreading of the final draft.

I am extremely grateful to my husband Jon, who was not aware what he signed up for when I told him I wanted to go back to university. Thank you so much for your enormous reserves of patience and for keeping me going in the moments of doubt. Finally I would like to thank Dr Aleksandra Porankiewicz, my first survey methods teacher, without whom I would not be here.

Definitions and abbreviations

ANOVA	Analysis of variance
BHPS	British Household Panel Survey
CAPI	Computer Assisted Personal Interview
CATI	Computer Assisted Telephone Interview
DEFF	Design effect
DEFT	Square root of design effect (also design factor)
DIC	Deviance Information Criterion
DP	Digit Preference
EPSEM	Equal Probability of Selection Method
GB	Great Britain
HHB	Hansen, Hurwitz and Bershad model
ILO	International Labour Organization
IGLS	Iterative Generalised Least Squares
IMC	Instructional Manipulation Check
MCMC	Monte Carlo Markov Chain
MLwiN	A software package for fitting multilevel models
MSOA	Middle Super Output Area
NatCen	National Centre for Social Research
NHIS	National Health Interview Survey (US)
NTS	National Travel Survey
NUTS2	Nomenclature of Units for Territorial Statistics
OA	Output Area
ONS	Office for National Statistics
PAF	Postcode Address File
PAPI	Pen and Paper Interview
PCA	Principal Component Analysis
PEDAKSI	Pre-Emptive Doorstep Administration of Key Survey Items
PSU	Primary Sampling Unit
RTE	Response Time Effort
SRS	Simple Random Sampling
TSE	Total Survey Error
UK	United Kingdom
VPC	Variance Partition Coefficient

1. Introduction

From their inception in the late 18th century to the present day, social surveys have almost certainly been the most important means of collecting information on populations in almost every country in the world. Survey research has been central to understanding populations and for gathering information for development and policy evaluation purposes. In the UK, the Labour Force Survey is the source of information on the unemployment rate and other key labour statistics, the Family Resources Survey provides information on living conditions of UK households, the Crime Survey for England and Wales is used to estimate levels of crime[1]. The decennial Census of England and Wales provides a snapshot of the entire population at very low levels of geography[2], while the 'Understanding Society' study is an example of a longitudinal survey that affords a consideration of how the lives of the UK population change over time and provides data that can help understand the impact of recession or evaluate specific government policies[3].

These surveys, and many others like them, are conducted using carefully designed methodologies that, if basic assumptions are met, provide accurate estimates of population parameters of interest, such as unemployment, prevalence of cancer, experience of crime, and so on. These estimates are, however, prone to various kinds of error. Understanding and quantifying this error is crucial to ensure that academic, policy and private sector research is based on accurate evidence. This three-paper doctoral thesis contributes to methodological research on survey errors and is primarily concerned with errors of measurement, as opposed to sampling and nonresponse error. Measurement error can originate from a number of different sources, including the respondent, the measurement instrument, and the interviewer. This thesis is primarily concerned with how interviewers can introduce measurement error into survey estimates.

Since most government and academic surveys in the UK and many other parts of the world continue to use a face-to-face data collection mode, understanding interviewer effects on survey estimates is an important area of investigation. While the interviewer effect has been extensively examined in existing research, it has rarely been explained using detailed information on interviewers such as personality traits, interviewing skills, attitudes towards

[1] http://www.ons.gov.uk/ons/guide-method/surveys/respondents/household/index.html
[2] http://www.ons.gov.uk/ons/guide-method/census/2011/index.html
[3] https://www.understandingsociety.ac.uk

interviewing and interviewers' own response styles when completing questionnaires themselves. To address this gap, the research presented across the three papers in this thesis makes use of 'paradata', a recent and increasingly popular concept in the field of survey methodology. Paradata are data that arise, usually, as by-products of the process of data collection and which may provide useful insight into how survey data have been collected, for how long, by whom and so on. The paradata used in this thesis include information on interviewers and electronically generated time stamps used to calculate response latencies.

The main objective of this thesis is to integrate the new methodological developments, such as paradata, with innovative statistical approaches to estimating interviewer effects; in order to improve our understanding of interviewer-related measurement error in surveys. Additionally, this thesis demonstrates the value of paradata in research on measurement error, whereas the majority of existing applications have tended to focus on nonresponse error. Interviewer-related measurement error is presented here in two different ways. In Paper One (Chapter Two) it is expressed as the interviewer-related variance component in survey estimates; that is to say, the focus is on the extent to which interviewers contribute to variable survey errors. In Papers Two and Three, the focus is more generally on quality indicators of survey response, with the theory of survey satisficing playing a central role (Chapters Three and Four). Here, the interest is in estimating whether interviewers affect the quality of survey responses, defined as the propensity to use various response styles which indicate a less than optimal cognitive strategy has been employed by the respondent. This is based on the theoretical proposition that some response strategies used by respondents, such as heaping or rounding their answer, can be a source of measurement error and potentially could be encouraged by interviewers.

This is a three-paper thesis, which means that each of the next three chapters constitutes a stand-alone paper. However, all three papers use closely related datasets and similar modelling approaches, which inevitably leads to a degree of unavoidable overlap between them. To aid the reading of this thesis as both a set of stand-alone papers and a coherent overall document, some trade-offs had to be made. Primary amongst these is that the main datasets are described in this first introductory section, rather than repeated in each chapter. The conclusion sections in each paper are limited to the discussion of the main results, with the principal messages coming from all three papers as well as the main implications of these studies, limitations and further research opportunities being compiled in Chapter Five.

This introductory chapter sets the scene and provides background for the three papers. The remainder of this chapter is structured as follows. The next section provides an overview of

the different sources of survey error using the Total Survey Error (TSE) framework, focusing in particular on measurement error. This is followed by a description of the survey design effect and how it can be calculated. Section 1.3 presents the substantive findings from existing research on the interviewer effect. Section 1.4 introduces the concept of survey paradata and describes the main research findings in this area. This is followed by a detailed description of the datasets to be used in the thesis, primarily the National Travel Survey (NTS), which forms the primary analytical focus of the three papers.

1.1 Survey errors

There are two key components of valid survey inference: first, that a respondent's answer to the survey question is the true representation of the measured phenomenon, such as the respondent's weight or opinion; and, second, that the survey sample is the correct representation of the population. If both assumptions hold, it is possible to make accurate inferences about the population parameter on the basis of the sample data (Groves *et al.*, 2009). If either of the two conditions is not met then the survey estimate is subject to error. In conceptual terms, survey error can be defined as a deviation between "what is desired in the survey process from what is attained" (Groves *et al.*, 2009, p. 40). The discussion on survey errors originated in the 1930s and early 1940s with the initial emphasis on sampling as the source of error (Bancroft and Welch, 1946; Cornfield, 1942; Neyman, 1934). Research on survey errors began at the United States Census Bureau. Deming distinguished 13 different factors affecting survey usefulness (Deming, 1944). He argued that survey designers should address them all simultaneously and that priority should not be given to any single one of them. Amongst those 13 factors were variability in response, bias and variation arising from the interviewer, sampling errors and processing errors. Deming's article can be seen as an early antecedent of the Total Survey Error (TSE) framework, which has come to dominate modern approaches to understanding and dealing with survey errors (Groves and Lyberg, 2010).

Survey statistics, historically at least, have tended to treat sampling variability as the main source of survey error. The notion of sampling variance is based on the theoretical replicability of samples, survey trials, interviews, and so on. A sample statistic will vary slightly as a result of the same survey design being implemented on different samples. This is because each sample drawn randomly from a population will yield a different realisation of a given sampling approach. The TSE framework is an extension of the sampling statistics approach to survey error (Groves, 2004). In the TSE framework, the mode of data collection

is specified as an additional potential source of variability of estimates (Kish, 1965). That means that the sample value is expected to vary over repeated trials of the same survey design even on *the same* sample. The TSE framework is particularly relevant to this thesis as it gives special attention to the interviewer and it recognises that the same interviewer may perform differently over repeated applications of the survey, resulting in the correlated response variance. From a TSE perspective, it is proposed that a respondent may also give different answers over repeated implementations of the survey, this is referred to as simple response variance (Groves, 2004). The TSE framework is the most comprehensive one existing in the literature, with a complex approach to many different survey errors originating from a numerous sources (Groves, 2004). The TSE framework adds a quality dimension to the survey process by assigning a specific type of error to each stage that is conducted in order to obtain the survey statistic. The survey error that is of particular interest in this thesis is measurement error.

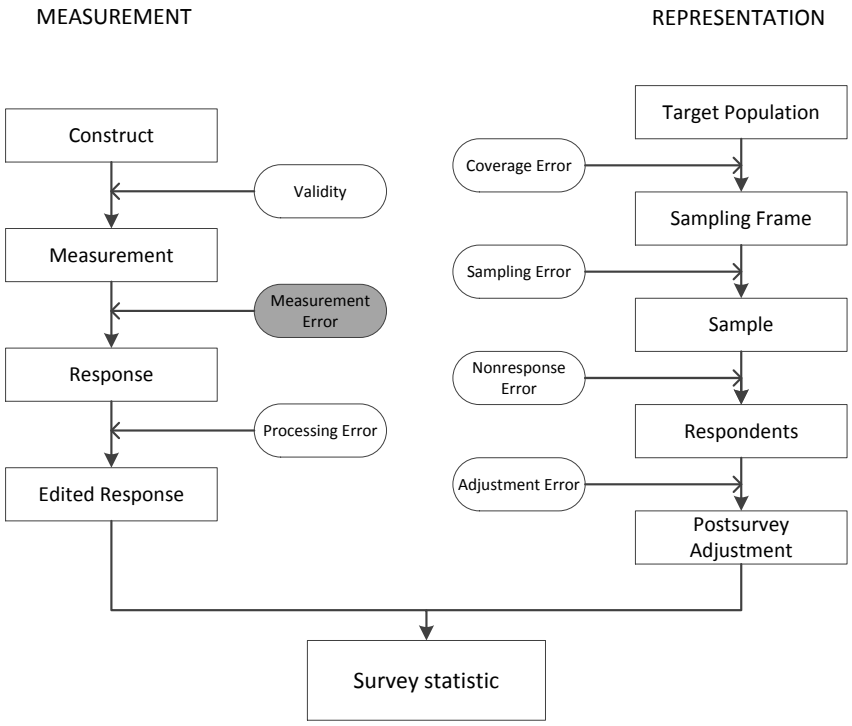
1.1.1 Measurement error

Measurement error is, in simple terms, the difference between the true value of a concept and the measured value of this concept. The concept may be relatively concrete (although this does not mean it is not prone to measurement error) such as age, number of children or earnings, or it can be more ambiguous and difficult to measure such as values, attitudes or opinions. In case of the latter one may even question whether the 'true value' really exists or not. The 'true value' for these types of psychological states is instead specified as being equal to the mean over the repeated trials, assuming that the respondent will have no memory of previous surveys and these can be repeated infinitely (Hansen *et al.*, 1951; O'Muircheartaigh, 1997).

Measurement error can result in both bias and increased variability of estimators. Bias occurs when there are systematic departures from the true value, for example respondents deliberately understating their weight, age or episodes of undesirable social behaviour such as drug use (DeMaio, 1984). Biases can be difficult to detect without (as is usually the case) some sort of external criterion or gold standard, such as administrative or hospital records. Variability of survey estimate, also known as 'response variance' relies on the concept of infinite theoretical trials or replications of the survey process. From the respondents' perspective, that means their responses to the same question will vary slightly every time they are asked the same question, again assuming that they are not influenced by their previous response. This type of error is random as the mean of the individual errors over trials is equal to zero (Groves, 2004).

In face-to-face interview surveys, measurement error can originate from three primary sources: interviewer, respondent and survey instrument (Seymour *et al.*, 1974). Moreover, the setting of the interview can be an additional component of the measurement error (Biemer and Lyberg, 2003). Interviewers as the source of measurement error are the main concern of this thesis, both in terms of bias and variability.

Figure 1.1 Survey lifecycle from a quality perspective (Adapted from Groves *et al.*, 2009, p. 48)



As shown in Figure 1.1, a social survey can be thought of as a process that consists of a number of inter-linked elements or stages: construct (information to be obtained by a survey), measurement (information about the construct), response, edited response, target population, sampling frame, sample, respondents and postsurvey adjustment (Groves *et al.*, 2009). The way these elements are specified and combined, for example what sort of sampling method or mode of data collection has been used, is referred to as the survey design. The way the survey is designed will have an effect on the quality of the final survey statistic. This is called the survey design effect.

1.2 Survey design effect

In most high quality surveys, probability sampling is seen as the gold standard. This is because probability samples, under certain assumptions, allow for statistical inference from a sample to the whole population, whereas non-random samples do not. There are four main criteria for sample selection: goal orientation, measurability, practicality and economy (Kish,1965). All these considered, simple random sampling (SRS) – although being 'the best' available in terms of measurability and yielding unbiased estimates of population values – is generally not used as the sole sampling strategy due to high cost and low practicality. An SRS single-stage sample of the whole UK population would require a UK-wide sampling frame and extensive (and expensive) travel between addresses for interviewers. One of the main features of SRS is that it is an Equal Probability of Selection Method (EPSEM) because every population unit has an equal probably of selection. Random samples that are not EPSEM are more commonly used in social surveys due to their lower cost and higher practicality. Any sampling design that is not EPSEM is referred to as a complex sample. The primary features of survey samples which divert from EPSEM are stratification, clustering and weighting (Heeringa and Liu, 1998).

Stratification is a sampling strategy whereby sampling is undertaken separately within different sub-groups for which the population distribution is known. This ensures that the population sub-group is represented in the sample according to a pre-specified ratio. When stratification is proportionate to the population sub-group size, it can increase efficiency by decreasing standard errors, while disproportionate stratification can ensure a sufficiently large sample size for small population sub-groups (Heeringa and Liu, 1998). Clustering is when a sample is drawn by first, dividing the studied population into sub-units; and then, drawing a random sample of these first-stage units. This can be done in more than one stage, such as drawing a random sample of UK schools and then a random sample of classes within each school. This is why clustered sampling is also called a multi-stage sampling. Clustering is commonly used in high quality government and academic surveys, such as those mentioned in the introduction to this thesis. Clustering is used because it does not require a sampling frame for the whole population of individuals, just for the population of clusters. It is also cost-effective because interviewing within a group of small areas decreases the cost of travel compared to an un-clustered sample. Weighting is a post-survey adjustment technique; it is generally used because in a household survey where only one household member is interviewed, individuals in larger households have a lower selection probability and this is corrected via design weights.

While these complex sampling design features are generally more practical and cost-efficient compared to a simple random sample of the general population, they result in different (usually larger) standard errors of survey estimates[4]. This difference in variances between an SRS and a complex sample is referred to as the 'design effect' (Groves *et al.*, 2009) and can be expressed as the ratio of complex sampling variance to the SRS sampling variance:

$$DEFF = \frac{v(\bar{y})}{v_{SRS}(\bar{y})} \tag{1.1}$$

where $\bar{y}$ is a linear outcome; $v(\bar{y})$ is the variance of the sample mean $\bar{y}$ under the complex survey design; and $v_{SRS}(\bar{y})$ is the variance of the same sample mean $\bar{y}$ under simple random sampling (SRS). If DEFF is different from 1 (usually greater than 1) then a design effect is present.

The largest contribution to the survey design effect derives from the effect of clustering (Kish, 1962). In practice, using geographical clusters as an example, individuals living in the same geographical area (cluster) are likely to be more similar to each other than individuals in the general population in certain characteristics such as property size or income (Tobler, 1970). This internal homogeneity of clusters is what causes the design effect due to geographical clustering, because it tends to underestimate the magnitude of population variance on the outcome in question, thus decreasing the precision of survey estimates compared to an SRS design of the same sample size. To measure the clustering effect Kish (1962) introduced the 'intra-class correlation coefficient', referred to as *rho*, which is a measure of the within-cluster variance component of a measured characteristic. For a binary outcome, y, *rho* can be interpreted as the probability that two units randomly drawn from the same cluster share the same value of y. The design effect due to clustering can be expressed as:

$$\mathrm{DEFF}_{\mathrm{cluster}} = 1 + \rho_{\mathrm{cluster}}(\mathrm{m} - 1) \tag{1.2}$$

[4]See Groves *et al.*, 2009 p. 106-109 for a detailed description of this process.

where *rho* ($\rho_{cluster}$) is the intra-class correlation coefficient and *m* is the average number of lower level units across the primary sampling units. This implies that the size of the design effect due to clustering is a function of the average size of the cluster (*m*) and the internal homogeneity of the outcome within clusters. The design effect will, then, be smaller if only a small number of units within clusters is surveyed. Hence, interviewing only a small number of second-stage units within each cluster is one way of decreasing the design effect due to clustering. This is because the similarity between units within the cluster, for instance respondents, means that at some point each additional respondent does not provide much new information. Moreover, values of DEFF need to be computed separately for each survey item because *rho* is variable-specific. For example, access to local amenities will be similar for respondents living in the same area, while other variables will be very different, for example political attitudes. Values of *rho* have been calculated for various types of surveys and survey items over the years and they tend to be rather small in magnitude, varying between 0.01 to 0.12 (Groves, 2004; West and Olson, 2010).

However, because DEFF also depends on the size of the cluster, even small values of *rho* can increase the variance of survey estimates substantially. For example, if the mean cluster size is 40 (*m* = 40) and the *rho* value is just 0.01, then DEFF = 1.39, which implies that the variance of the sample mean is 39% greater than an SRS design of the same size would yield, as a result of the clustered design. Another measure that can be derived from DEFF is its square root, referred to as the design factor, or DEFT. DEFT is generally more readily interpretable because it is in standard error rather than variance units (Sturgis, 2004). Although the design effect due to clustering usually relates to geographical clusters, it relates equally to any situation in which observations across lower level units are not independent as a result of being made within the same higher-level unit. A common such situation is when individual respondents are clustered with the same interviewer.

1.3 Interviewer effect

Kish (1962) observed that respondents interviewed by the same interviewer gave more similar answers than respondents across the sample as a whole. This effect is thought to arise primarily because interviewers ask all their respondents the same question in the same, specific way, because they make the same mistake in reading a particular question, or because they 'help' respondents to understand unclear or ambiguous questions. Equation 1.1 can be easily adapted to isolate the interviewer-related DEFF as:

$$DEFF_{int} = 1 + \rho_{int}(b - 1) \quad (1.3)$$

where b denotes mean interviewer workload and ρ_{int} is the rate of homogeneity or intra-interviewer correlation on the survey outcome. As with geographical clustering, the ρ_{int} values reported in the literature tend to be small but, because $DEFF_{int}$ also depends on the size of workload, it can lead to significant increases in the variance of estimates if interviewers conduct many interviews.

1.3.1 Separating area and interviewer design effects

Research into the interviewer effect in complex survey designs needs to overcome one major obstacle. Although driven by different mechanisms, both clustering (from this point on referred to as area effect) and interviewer are responsible for the same effect on variance of the survey estimates. This is not problematic in itself. However, it is standard in most surveys for a single interviewer to be allocated to a particular area. Where this is the case, although it is possible to calculate the design effect, it is not possible to separately identify how much of the design effect is due to the area and how much due to the interviewer. In other words, it is not possible to differentiate how much of the interviewer-related variance is due to the interviewers themselves and how much due to the specific cases they were assigned; the two effects are completely confounded (Brunton-Smith *et al.*, 2012).

Three main approaches to estimating the interviewer effect can be distinguished: the Hansen, Hurwitz and Bershad (HBB) model, the ANOVA model and the multilevel approach. The HHB model (Hansen *et al.*, 1961), also known as the US Bureau of the Census model or correlational approach, dominated the way the interviewer effect was estimated in early studies by using variance decomposition techniques. They assumed that interviewers introduce a positive correlation between responses given by 'their' respondents – the correlated response error. The ANOVA model was proposed by Kish (1962). He defined measurement error as the combination of fixed errors, those controlled by the 'essential survey conditions', and variable errors. He further categorised these errors into totally random errors, hence not measurable without replication, and correlated effects that are essentially interviewers' bias imposed on their respondents (Kish, 1962). Kish sees the interviewer variance as a component of the total variance per respondent. These two early approaches to separating the area and interviewer effects used an interpenetrated design, in which respondents are randomly assigned to interviewers (Mahalanobis, 1946). The

interpenetrated design is expensive and difficult to implement and for those reasons it is rarely used in practice. Moreover, any conclusions drawn from research using the interpenetrated design are of limited generality because research institutes do not commonly use it in everyday survey practice.

More recently, it has become popular to employ multilevel models and Bayesian estimation techniques to derive the interviewer and the area variance components simultaneously (Davis and Scott, 1995; Hox, 1994; Pickery *et al.*, 2001; Schnell and Kreuter, 2005). This approach takes advantage of sampling designs in which there is a 'natural' overlap between interviewers and areas (O'Muircheartaigh and Campanelli, 1998). This crossing is then a key to separate interviewer and area effects. However, as it is not done at random, this approach relies on statistical control to be confident that variance components are partitioned accurately (Hox, 1994). The key advantage of these models is that explanatory variables associated with the area, respondent and interviewer can be added at the same time into the model. A statistical estimator of interviewer effect in the multilevel approach is presented in Chapter Two (Paper One).

There is a large body of research concerned with the interviewer effect as the interviewer-related proportion of the total variance of survey estimates. This interviewer-related variance is first estimated and then usually used to assess what types of survey questions are more prone to the interviewer effect, i.e. they exhibit a larger interviewer-related variance component. Dependent variables used in these studies are the actual survey items. Another body of research is mostly concerned with various indicators of survey data quality. In these studies, the dependent variables are indicators such us interview speed, item nonresponse or non-differentiation on survey items. This research is concerned with many factors that can potentially affect those quality indicators, such as respondent, question type, interviewer assignment and finally interviewers themselves and their various characteristics. Both types of studies are described below.

1.3.2 Interviewer-related variance

One example of research in which the interviewer effect is operationalised as the interviewer-related variance component is the study by Davis and Scott (1995). They divided survey questions into four categories: sociodemographic, attitudinal, reports of recent behaviour, and recall of distant behaviour. First they partitioned interviewer and area effects. It turned out that the area effect was larger than the interviewer effect but, at the same time, the proportion of the overall design effect that was due to the interviewer was largest for attitudinal questions and smallest for sociodemographic variables. Second they compared the

effects between two domains: gender (male/female) and ethnicity (European/Non European) of interviewer. Their results suggested that the interviewer effect changed slightly for different domains and on selected items. Moreover it was larger for ethnicity than for gender (Davis and Scott, 1995).

O'Muircheartaigh and Campanelli (1998) used a combination of interpenetrated design (experimental control) and cross-classified model (statistical control) to separate the interviewer and area effects on survey estimates using the British Household Panel Survey (BHPS). In their interpenetrated design they assigned interviewers to areas at random within geographical pools. Each geographical pool consisted of two to three neighbouring primary sampling units (PSUs) and there were two to three interviewers assigned to work in each pool. That means the PSUs and interviewers were cross-classified within larger geographical pools.

They examined the following characteristics of interviewers: age, gender, length of service and whether the same interviewer conducted both interviews, in both waves of the panel. Similarly to Davis and Scott (1995) they divided survey items into categories: interviewer check item, factual items, and attitudes. They found that only interviewer gender was a statistically significant explanatory factor of the interviewer effect on interviewer check items – which in that case was presence of children during the interview. For the factual item, readership of *The Independent* newspaper, two interviewer characteristics were statistically significant: continuity (over two waves) and status (regular interviewer, supervisor or area manager). Finally, for the attitudinal variable – likelihood of having more children – interviewer experience was a significant factor affecting the interviewer effect.

Across all items examined, O'Muircheartaigh and Campanelli found that the interviewer's age was the most frequent statistically significant predictor of the interviewer effect. It is also worth noting that their hierarchical analysis of variance (as in Davis and Scott, 1995) showed that the interviewer effect was larger (for all types of variables) than the area effect: an opposite conclusion to that of Davis and Scott. This may be because they were using a general household survey and interviewers while Davis and Scott analysed medical survey data that used a small group of highly trained interviewers. Extensive training of interviewers tends to decrease interviewer-related variance (Fowler and Mangione, 1990).

Schnell and Kreuter (2005) also used an experimental, interpenetrated design by assigning three interviewers to work within each PSU independently of one another. They utilised a nested hierarchical model with three levels: respondent, interviewer and PSU. In this model

the interviewer and area effects were both treated as random. Schnell and Kreuter's work focused particularly on different types of items in the questionnaire and the magnitude of the interviewer effect on them, rather than trying to explain the interviewer effect *per se* through certain interviewer characteristics. They argue that the one-dimensional classification of survey items is not sufficient when looking at different types of questions; but rather, a combination of different factors relating to the question should be taken into account. These factors are: difficulty of the question, sensitivity of the question, whether the question is factual or not, format of the question (open/closed). They discovered that questions with more 'risk' factors, such as a difficult question that is also an open question, are in general more prone to the design effect. Interestingly, depending on the type of 'risk' factor, either the interviewer or the area effects form the larger component of the overall design effect (Schnell and Kreuter, 2005).

Their results showed that some items are almost entirely affected by the area (e.g. home ownership indicator) while some are almost entirely affected by the interviewer (e.g. incivility items). They concluded that sensitive questions, nonfactual questions and open questions are more prone to the interviewer effect than difficult questions. Schnell and Kreuter (2005) lacked information on interviewers but they included some area characteristics in their model. The available area characteristics: signs of loitering, visible graffiti, presence of abandoned or empty houses, motion detector installed at home all proved to be significant predictors of the overall design effect. That means that certain area characteristics successfully explained some of the variability of survey estimates caused by the design effect.

A similar approach was adopted by Brunton-Smith *et al*. (2012). They used the rolling design[5] of the British Crime Survey to obtain a 'natural' cross-classification of areas and interviewers. They included a range of area and interviewer characteristics and used a cross-classified multilevel model to separate the interviewer and area effects. The main advantage of their approach is that there is no need for a costly interpenetrated sample as long as the survey has a rolling design and interviewers can be 'followed' across the cases. At the same time, this is also the main disadvantage of this approach, as it is still possible that certain interviewers were assigned truly different cases (as the allocation of respondents to interviewers is not random). However, the pattern of results obtained was in line with those achieved with the interpenetrated design, which suggests that the statistical control strategy was effective. They found that factual questions with little interviewer input (i.e. no probing, no showcards)

[5]Continuous survey with monthly interview quotas.

tended to be the least affected by the interviewer. The interviewer effect increased if showcards or probing were employed. Behavioural and attitudinal items were much more prone to interviewer effects and this increased even more with showcards and probing (Brunton-Smith *et al.*, 2012).

1.3.3 Interviewer effect on data quality indicators

The studies described in the previous section are mostly concerned with the interviewer effect on measurement error, operationalised as the interviewer-related variance of survey estimates. Another approach to survey measurement error introduced by the interviewer is the effect on data quality indicators such as item nonresponse, socially desirable responses, or acquiescence. Instead of estimating the actual survey response variance, researchers treat 'data quality indicators' as dependent variables. These models may also be concerned with the interviewer-related variance component, as large values of this component would suggest a substantial interviewer effect on data quality indicators. Estimating those over different models for different quality indicators may help identify 'good' and 'bad' interviewers, 'good' being those responsible for lower variance across many items (Pickery and Loosveldt, 2004). However, more often, studies that are concerned with data quality indicators use various interviewer characteristics as independent variables in order to explain the relationship between interviewers and data quality.

Hox (1994) investigated how certain interviewer characteristics can affect the quality of the data and whether this effect is different for different respondents. He did not investigate the area effect, but rather the relationship between respondent and interviewer and the effect on interview speed as a measure of data quality. Factors that explained some of the interviewer effect were: interviewer training, preference for the Computer Assisted Personal Interview (CAPI) method and extroversion.

Surveys concerning sensitive topics are especially prone to data quality issues such as drop-outs, missing data or social desirability where respondents tend to underestimate their reports of undesirable or harmful behaviour, such as smoking or drug use. These problems could be solved if interviewing conditions were carefully monitored or customised to some extent. One example may be matching the interviewer's and the respondent's gender when questions concern sexual behaviour (Catania *et al.*, 1996) or controlling for the interviewer's race or religion if these are likely to resonate with the survey topic, such as views on racial discrimination (Anderson *et al.*, 1988).

Many measures of data quality concern respondents rather than interviewers. These are usually linked with respondents' cognitive skills, their motivation or levels of fatigue (Burton and Blair, 1991). Krosnick (1991) coined the term 'satisficing' to describe a range of strategies employed by respondents when replying to survey questions. Providing responses to questions can be a burden, especially if a questionnaire is long and questions are difficult. Some respondents may avoid putting in sufficient effort to provide a response that is accurate and well-thought-through and instead take a 'cognitive shortcut' by providing only a 'good enough' response (Krosnick, 1991). Indicators that are used to detect survey satisficing are also considered measures of data quality (Barge and Gehlbach, 2011). Examples are the prevalence of 'Don't know' responses, acquiescence or 'yeah-saying', non-differentiation on survey items containing scales, refraining from stating an opinion by choosing a 'middle' option or choosing extreme response options from either end of the scale. Choosing the first or last viable response option from the available list or estimating/rounding a response are also indicators of satisficing (Krosnick *et al.*, 1996).

Satisficing is thought to be one of the main sources of measurement error in surveys and it is a quite well researched topic. However, studies on satisficing focus mostly either on respondents and their characteristics (Kaminska *et al.*, 2011; Oppenheimer *et al.*, 2009)or on the survey instrument (Chang and Krosnick, 2010; Krosnick *et al.*, 2002). Much less attention in research literature is given to interviewers (Olson and Peytchev, 2007) and interviewers' own satisficing behaviour has not been used in studies on this type of measurement error. As in the majority of studies on measurement error, this may be because of a lack of available measures concerned with interviewers. If researchers decide to include some interviewer level variables in their models, these are mainly basic demographic characteristics. This thesis addresses this gap by using a range of variables describing interviewers, such as their personality traits, skills, attitudes towards interviewing and even their own response styles. This has been possible thanks to the development of survey process data – paradata – and their availability for use in methodological research.

1.4 Paradata

This section presents the concept of paradata and its origins. It also briefly explains current developments in survey methodology using paradata and their practical application. The term paradata stems from a much broader term of metadata. A somewhat vague definition was offered by Berg *et al.* (1992) for whom metadata are 'data about data'. A more specific and more useful expanded definition was presented by Bethlehem (1997) for metadata to be a "formal statement of the information required for collecting, processing and publishing

survey data" (p. 381). A similar view of metadata is shared by Dippo and Sundgren (2000), who argue that metadata play a crucial role in understanding statistical information. For example, the variable label and value labels for a variable are basic information needed to actually use the variable. Some further information about the variable, if and how it was computed or when in the questionnaire it was asked, add to researchers' understanding of the variable. Another purpose of metadata is to provide information about the process of data collection and quality measures. Process data started to be of importance once the quality measures of the survey process became a more pressing issue for research agencies (Lyberg, 2009).

Paradata is a much narrower concept than metadata; simply, paradata are a certain type of metadata that describe the process of data collection. This term was first coined by Couper (1998), who mainly saw paradata as potentially useful by-products of CAPI. A clear and agreed upon definition of paradata does not exist and the term has been quite freely used in the literature since its inception. In narrow terms, paradata are technical process data automatically collected by CAPI systems such as time stamps, keystrokes and error messages. In broader terms, paradata are more directly linked to the interviewer, such as interviewer's observations; for instance, notes on the condition of the household compared to other buildings on the same street, interviewer's comments on the interview process, administrative information entered by interviewers and so on. Perceived sometimes as the 'borderline case' of paradata is information about interviewers such as administrative data or even additional data collected from interviewers for methodological purposes[6] (Kreuter and Casas-Cordero, 2010).

Paradata can be collected at the macro and micro level (Scheuren, 2001). Macro paradata are summary process data for the survey as a whole, for example the response rate at a specific point in time during the fieldwork. Micro paradata refer to a single case in the dataset, for example the specific time at which each call to the sampled household has been made. While macro paradata are already widely used, social researchers are less familiar with micro paradata due to them being novel and more difficult to use in practice (Scheuren, 2001). The increased importance of paradata is the effect of, amongst others, increased use of computing in survey research, the growing importance of accessing micro data and the growing need to quantify all survey impacts on inference (Scheuren, 2005). It is important to look at paradata

[6]One example could be the 'Who is behind the NatCen ID card' survey of the interviewer network working for the National Centre for Social Research.

in conjunction with the quality of the survey process. This is where paradata can be usefully integrated with the total survey error framework reviewed earlier. The majority of examples of paradata applications are to do with nonresponse error.

As already mentioned, macro paradata refer to the survey process as a whole and are used to make adjustments in terms of cost reduction while maintaining the required quality of survey estimates. One example is the responsive design methodology (Groves and Heeringa, 2006). In this approach, the data collection process is split into sequential design phases. Phase one usually uses a few different methods, which are then evaluated using paradata. The most suitable method is then applied to the subsequent design phases. Quite often, phase one uses a random experiment approach to control for design phase sample characteristics.

In the responsive design approach, the paradata used are well defined and built into the process from the very beginning. They include observations on each interviewer, observations on segments (groups of households, neighbourhood), observations on households, observations on each call and each call yielding a contact and, finally, observations on the interviewer's day. Responsive design rests on the assumption that paradata are a good proxy for measuring a respondent's propensity to participate in the survey. Statistical methods are used to estimate this propensity for each call and then concentrated efforts to gain cooperation from the respondents with the highest propensity scores are invested in the subsequent design phases (Groves and Heeringa, 2006).

A similar approach is used by 'active management' as described by Laflamme *et al.* (2008) at Statistics Canada. Active management works as the in-built quality control tool for the survey process in order to keep the costs down while keeping the survey estimates as accurate as possible. The main elements of this tool are planning, monitoring, identifying collection problems, taking appropriate action, communicating and evaluating. All this happens through extensive, real time reporting to every level of management. This live reporting is possible thanks to the collection and analysis of paradata.

As Statistics Canada mostly concentrates on the effectiveness of the data collection process (in terms of cost), much less is said about how the quality of survey statistics is measured. One of the reports includes a 'progress at sub-population level', which shows sample realisation for a particular domain or population of interest. For example, if the survey estimate of interest was the number of sexual partners for a woman and older women were underrepresented at stage one of the data collection process, then estimates will be affected (as older women are more likely to have had more sexual partners due to being sexually active for longer). Using the 'progress at sub-population' report, this issue can be detected

while still in the fieldwork phase and the data collection process can be adjusted appropriately (Laflamme *et al.,* 2008).

The same approach was followed by Olson (2009) and Olson and Groves (2009), who used paradata to estimate the response propensity of respondents. They showed that switching between data collection methods – i.e. introducing an incentive, using a mail questionnaire instead of a phone call – during the data collection process increases the response rate. They also demonstrated that this effect wears off over time and a new 'stimulus' – another switch – is necessary after a while. Similar research was conducted by Lee (2009). She concluded that tailoring incentives may increase participation in certain groups of respondents.

The biggest criticism of these approaches is that each design phase has its own nonresponse and measurement errors, as well as different modes of data collection, and for that reason the estimation of survey statistics, especially in the interim, becomes extremely complicated. Another criticism concerns paradata themselves. They are new, still under development and more work is needed, especially in the area of their measurement error.

Both responsive design and active management are most concerned with macro level paradata, used first and foremost to manage the data collection process effectively based on cost and quality criteria. However, there is also a rapidly growing body of research that involves paradata at micro level, especially call records. Paradata can be used to predict the best times to contact households in order to obtain interviews. The major advantage of call record data is that they are usually available for both respondents and nonrespondents (Durrant, 2009). Moreover, linking nonresponding cases with census information allows a more in-depth analysis of nonresponse issues taking into account the demographic characteristics of nonrespondents. This research was later extended to include interviewer characteristics and their effect on the success rates of obtaining interviews (Durrant *at al.*, 2010).

Paradata have become very popular for investigating the research issue of survey nonresponse. Paradata about interviewer characteristics – such as experience on the job, satisfaction with the job, personality traits, skills and commitments outside work – were used to predict success in obtaining interviews on the doorstep (Jäckle *et al.*, 2011; Sinibaldi, 2009). Other research on nonresponse in surveys that utilise paradata are Lynn's PEDAKSI methodology (2003), in which a special short questionnaire on non-responding households is completed, or Wang *et al's*. (2005) investigation into interviewer's effort and nonresponse bias. Interviewer observations, another kind of paradata, are mainly collected during the call

and include examples such as: presence of litter in the area, condition of the building compared to others in the same street, presence of an entry phone system and presence of a car on the drive. This information is mainly used to get more insight into doorstep interaction between the respondent and the interviewer (Groves and Couper, 1998; Sturgis and Campanelli, 1998).

The most recent development in the area of paradata is the measurement of paradata themselves. This strand of research focuses mostly on the variability of interviewers' observations (McCulloch *et al.,* 2010; Sinibaldi, 2010; Sinibaldi *et al.*, 2010; West, 2010). All these studies focused on the extent to which replies given by different interviewers actually vary, or on the discrepancies between the objective source of information, for instance gender stated by the respondent in the questionnaire and the interviewer's observation. For example, McCulloh *et al.* (2010) discovered that around 10% of interviewers' observations on respondents' gender were incorrect, with female respondents more often categorised as males.

Although, as outlined above, some work has been done towards estimating the measurement error of interviewers' observations, the general error associated with paradata is still a major issue for the researchers who wish to use them. The paradata collected automatically by computer systems are more reliable by definition but, in turn, their use is difficult due to their messiness. In general it can be said that paradata lack sufficient metadata in order to make them easily usable. Another issue is the availability of paradata. They are not routinely deposited in data banks along with micro data, hence access needs to be gained via the survey agencies, usually following a complicated and lengthy process. Some change in that respect can be observed however. For example, data from the US National Health Interview Survey (NHIS) in 2006 were released with an accompanying paradata file (Taylor, 2008). Nevertheless, the increasing use of paradata is driven mainly by academic research and, despite positive signs, more dialogue is needed between academia and survey agencies in order to make the most of paradata. This thesis is a good example of such cooperation taking place with mutual benefits, for instance Paper Three (Chapter Four) illustrates the methodological benefits of audit trail data if they are extracted in a usable format.

The overview of different applications of paradata presented in this section shows that paradata have only rarely been used to understand and mitigate measurement error. Couper (2009) initially proposed this application of paradata, when he suggested that inexpensive automatically generated paradata can, at least partially, replace the more expensive methods such as digital recording and behavioural coding, cognitive testing or eye-tracking studies

(Couper, 2009). Since then some progress has been noticed, especially in studies that evaluate response latencies as the indicator of data quality or the indicator of a problematic question (Couper and Kreuter, 2010; Yan and Tourangeau, 2008). Even though some studies use paradata to address the issue of measurement error, they mostly focus on self-administered questionnaires, such as web surveys and mostly use client-side paradata such as response change, mouse movement and item-level times. In other words, these studies focus on the instrument or respondent. Even studies that use data from the face-to-face survey focus more on question types, presumably due to lack of information on interviewers (Couper and Kreuter, 2013). These studies are presented in more detail in Paper Three (Chapter Four).

1.5 Data

This section provides an overview of the main datasets used in all three papers in this thesis. The descriptions of data and measures developed that are specific to the different analyses are included separately in each paper.

1.5.1 The National Travel Survey

All three papers in this thesis use the National Travel Survey (NTS) collected by the National Centre for Social Research (NatCen)[7]. Paper One and Paper Two use seven years' worth of NTS data collected between 2002 and 2008. In Paper Three only NTS data collected in 2011 were used. The NTS 2011 is the first NTS dataset produced by NatCen that includes a formatted Blaise[8] audit trail – the source of electronically collected paradata such as time stamps and keystrokes. Until recently this type of paradata remained largely an untapped resource due to lack of easy extraction methods and lack of useful formatting. Researchers wishing to make use of such paradata had to go through the laborious data cleaning and preparing process themselves.

[7]NatCen is an independent centre for social research. They have a status of non-profit organisation and conduct research in the main social areas: health, education, income and welfare, transport, social inclusion etc. They have been conducting surveys commissioned by government bodies e.g. Department of Transport (NTS), as well as local authorities. The NatCen interviewer's network also works on the 'Understanding Society' panel survey. www.natcen.ac.uk

[8] Software preloaded on interviewers' laptops

The NTS covers the whole of Great Britain[9] and is the main source of information on the travel patterns of its citizens. Topics covered in the questionnaire include: distance to amenities, deliveries, shopping, children's travel, transport options, household vehicles, transport barriers, transport-related benefits and long-distance journeys. Data are collected using CAPI with all household members (children under 11 are interviewed by proxy) and a seven-day self-completion travel diary. The NTS data used in this thesis are available at two levels: household (CAPI interview with the household reference person; usually the person with the highest income) and individual (CAPI interview with the household reference person and all the remaining household members including children). To reduce the burden on respondents, some NTS questions are asked of a random subsample of respondents or are asked every other year. Moreover, some questions have been dropped over the years. The implication for this study is that analyses for each dependent variable are based on different numbers of cases.

The NTS employs a multi-stage stratified clustered sample in which, each year, over 15,000 addresses are drawn from the Postcode Address File (PAF). In stage one 683 (684 in some years) postcode sectors (PSUs) are selected using SRS. In the second stage, 22 addresses are selected within each PSU, also using SRS. Variables used for stratification are: region (NUTS2[10]), car ownership (based on census results) and population density. The NTS also uses sample boosters for Inner and Outer London areas. Since 2002, the NTS has used a quasi-panel design in which half of the PSUs (but never the addresses) are re-used in the following year. The main purpose of this design is to decrease the variability of survey estimates. As each NTS cohort used for analyses in this thesis is subject to the same quasi-panel design, it is largely irrelevant to the analyses conducted in this thesis and hence it is ignored.

The NTS uses a rolling design. This is because travel patterns depend on seasons. In each survey year, interviews are conducted in monthly quotas of 57 PSUs. They are allocated in such a way that every quarterly sample, 171 PSUs, is representative of the whole survey population. During the survey year each interviewer works in more than one PSU and each PSU is 'served' by a number of interviewers. As an effect, a 'natural' cross-classification of

---

[9] Although the NTS population are the private households in Great Britain, for practical reasons the Scottish islands and Isles of Scilly, which account for 0.2% of addresses in GB, are removed from the sampling frame.

[10] Nomenclature of Units for Territorial Statistics (NUTS) as developed by Eurostat. For more details please see the NTS technical report 2003-2004; p 9-10 (http://www.esds.ac.uk/doc/5340/mrdoc/pdf/5340technical_report_2003_2004.pd).

interviewers and PSUs occurs. Because the data analysed here span seven years (2002 – 2008), cross-classification also occurs across the years and due to the quasi-panel design some interviewers work again in the same PSUs in the subsequent survey years. NTS yearly response rate remains at the 60% level. More detailed information on the NTS methodology can be found in the technical reports available on the NTS website (for example Hayllar *et al.*, 2005).

1.5.2 The survey data on interviewers

Information about interviewers comes from two different sources: basic administrative information such as age (in 2010), gender, grade and indicator which NTS administrations worked on, and the results from the methodological survey of all NatCen interviewers conducted in 2008 and titled 'Who is behind the NatCen ID card?'. The survey results are the main source of interviewer-related paradata used in this thesis. It used the pen and pencil (PAPI) collection method and was sent by post to 1,478 interviewers who worked for NatCen between January 2006 and March 2008. Both former and current employees were included. The overall response rate was 81% and it was higher amongst current interviewers. The results from the survey were supplied as a separate file for those interviewers who worked on NTS between 2002 and 2008 only. It contains 450 interviewers of whom 47% were males and 53% females. The mean age was 60 years old. The mean tenure at NatCen was eight years. Only one interviewer worked on all seven NTS administrations, while 23% of the sample worked on the NTS only once (modal value). The sample consisted of quite experienced interviewers as almost half of them (46%) were C grade and 17% of them were supervisors[11]. Half of them had some other interviewing experience apart from at NatCen.

The main section of the questionnaire asked interviewers about their personality traits and transferable skills. Interviewers used a seven-point Likert scale (1 – Does not apply to me at all; 7 – Applies to me perfectly) to reply to 67 statements including the main personality traits – 'Big Five' – agreeableness, conscientiousness, extraversion, neuroticism and openness (Goldberg ,1990) as well as others such as ability to pick up cues, ability to read other people and empathy. Questions about skills included time management, organisation and attention to detail, problem solving, presentation of self, adaptability, resilience, verbal and nonverbal communication, listening, persuasion and assertiveness. Other items in the 'Who is behind

[11] NatCen interviewer grades range from A to C then supervisors and finally a team leader. A stands for the lowest grade.

the NatCen ID card?' questionnaire of interest were: what aspects of the interviewer's job are the most important; satisfaction with the main aspects of the job such as pay as well as the overall satisfaction with the job at NatCen (this item was only asked of current interviewers); and interviewers' attitudes towards persuading respondents. The 'Who is behind the NatCen ID card?' questionnaire can be found in Appendix 1.

1.5.3 Area-level variables

The NTS data were also linked with area-level characteristics. Postcode sectors are the PSUs in the NTS. These are not particularly meaningful as spatial units because their design was mainly driven by the requirements of the mail delivery system (Raper *et al.*, 1992). The area characteristics used in this thesis are at the Middle Super Output Area (MSOA) level rather than the postcode sector level.

MSOAs are based on the UK census geography. New geographical units – Output Areas (OA) were designed for the 2001 census, and onward censuses, in such a way as to be more meaningful geographical areas. That means they are stable over time, similar in size and account for natural boundaries such as rivers (Martin, 2001). The OAs were later merged into the Super Output Areas (SOAs) at two distinct levels: lower and middle, taking into account social homogeneity of the area and its size. The creation of MSOAs followed extensive consultation with local authorities resulting in 7,193 MSOAs in England and Wales. In Scotland the equivalent of MSOAs are called Intermediate Geographies[12]. MSOAs have on average a population of 7,200.

The area characteristics at the MSOA level used in this thesis include: socio-economic disadvantage, urbanisation, population mobility, age profile, housing structure, crime rate and ethnic diversity. These indicators were computed using various neighbourhood measures taken from the 2001 census data as well as administrative data from ONS and were turned into uncorrelated factors using Principal Component Analysis (PCA) (Brunton-Smith, 2008). Because the administrative data sources vary for the constituent countries these indicators were computed for England only, subsequently narrowing the analyses in this thesis to respondents residing in England. Originally these indicators were designed to work as the measure for social disorganisation and disorder within neighbourhoods for the study on the fear of crime (Brunton-Smith and Sturgis, 2011) but they can be successfully used as

[12] More information on the SOAs and MSOAs is available from the ONS website: http://www.ons.gov.uk/about-statistics/geography/products/geog-products-area/names-codes/soa

controls for any differences between interviewers' assignments due to the different geographical areas they work in.

1.6 Contribution of this thesis

The research presented in the next three chapters of this thesis addresses important gaps in the literature on measurement error, interviewer effects and satisficing. Paper One in the next chapter uses interviewer-related paradata to examine the interviewer effect expressed as the interviewer-related variance of survey estimates. Although similar paradata have been used to investigate interviewers' cooperation and contact rates, they have never been used to understand the generation of measurement errors. The results of this study show that certain interviewer characteristics such as satisfaction with pay, determination, experience with the survey, communication skills, conscientiousness and extroversion can affect how much variance interviewers introduce into the survey estimates. Interviewers who introduce more variance and hence negatively affect survey data quality are those less satisfied with their pay, less determined, with reasonable survey experience and with lower level communication skills. Moreover, somewhat counter-intuitively, those with high levels of conscientiousness and low levels of extroversion introduce the most variance into survey estimates.

The following two chapters are mainly concerned with survey satisficing. Paper Two (Chapter Three) presents current research into satisficing, its indicators and sources, with the emphasis on the interviewer. A new indicator of satisficing, heaping, is introduced and computed using survey responses (for respondents) and the NTS Administrative Block (for interviewers). A cross-classified multinomial ordered logistic model is then used to estimate respondents' propensity to satisfice. The results show that interviewers' own satisficing behaviour can positively affect satisficing amongst their respondents. Those who were interviewed by highly satisficing interviewers were more likely to be in the high satisficing group of respondents.

Satisficing respondents should take less time to answer the survey (Callegaro *et al.*, 2009; Oppenheimer *et al.*, 2009). The logical implication of satisficing as saving cognitive effort when responding to a survey is saving time as well. The study presented in Paper Three (Chapter Four) uses automatically generated response-time data to better understand respondent satisficing behaviour, while controlling for the interviewer effect. It shows that providing a 'Don't know' response – a commonly used satisficing indicator – actually results

in longer response times than providing any other type of response such as rounded response or optimised (as opposed to satisficed) response. If cognitive effort invested by interviewers is measured as time needed to formulate the answer, then a presumed satisficed response actually requires *more* cognitive effort. The results in Paper Three suggest that some popular measures of satisficing may also be measures of optimising, depending on the difficulty of the question. Throughout this thesis, the value of paradata in social survey research and their application in research on measurement error in surveys is being demonstrated.

2. Using paradata to explain the interviewer contribution to the design effect in face-to-face surveys

2.1 Introduction

Every survey is subject to error. Historically, sampling error was the first type of error to be considered by methodologists (Neyman, 1934). With time and the development of the field of survey statistics, more sources of survey errors were progressively recognised (Deming, 1944). Biemer and Lyberg (2003) distinguish two broad types of survey error: sampling error, which emerges when only a sample of a population is interviewed; and nonsampling error, which emerges due to mistakes in the survey process. Groves (2004) offers a similar typology of errors in two separate categories: errors of representation, such as coverage error and nonresponse error, and errors of measurement. He also states that it is important to look at survey errors from the 'process quality' perspective (Dippo, 1997) which means that survey errors are identified and addressed at every stage of the survey process in order to obtain a high quality survey statistic.

The interviewer was recognised as a source of error in survey estimates as early as the 1930s (Rice, 1929) and has been investigated in a large number of subsequent studies. The interviewer can introduce errors in a number of different ways: through nonresponse, that is through failing to obtain an interview; through non-coverage error, that is failing to identify or select the correct household; and through measurement error, failing to record information correctly, or influencing the respondent in a way that leads him or her to provide an inaccurate answer (Groves, 2004). Measurement error can also, of course, be introduced by a respondent, when failing to give the right answer, or a questionnaire, when it is too long or includes badly designed questions (Biemer and Lyberg, 2003).

The interviewer can introduce measurement error in many different ways: different interviewers may vary in the way they assist respondents with the questionnaire such as probing or using show cards; different interviewers may read out questions in a different way; and, finally, certain interviewer characteristics such as gender or ethnicity may affect the relationship with the respondent in a way that affects the response he or she provides (Groves, 2004). For example, the presence of a male interviewer may cause a social

desirability effect for female respondents (DeMaio, 1984), which results in the underreporting of socially undesirable behaviours. Respondents may misrepresent their drug use or number of cigarettes smoked and this may be exacerbated or weakened by interviewer characteristics. The measurement error introduced by the interviewer is referred to as the interviewer effect. It comprises errors of both bias and variance.

Interviewer characteristics seem likely to play an important role during the survey interview. Some interviewers may simply be better suited to the job in terms of their personality and particular skill set. Interviewer characteristics such as gender, physical attractiveness or dress may affect the interview if some of these features are particularly liked by respondents (Groves *et al.*, 1992). For example, there may be a different dynamic between a male interviewer and male respondent than between a female interviewer and male respondent (Catania *et al.*, 1996). Finally, an interviewer's ethnicity or religion can change the whole context of the interview when, for example, questions about attitudes towards black people or the Jewish faith are asked (Robinson and Rohde, 1946; Schuman and Converse, 1981).

The objective of this paper is to assess whether interviewer personality, attitudes, skills and demographic characteristics are predictive of the magnitude of interviewer variance. This is the first time that interviewer paradata – extensive information on interviewers such as their attitudes towards interviewing, personality traits and skills – have been used to address this question. The remainder of the paper is structured as follows. The next section describes the ways in which interviewers contribute to survey error and the methods, which can be used to estimate the interviewer effect. This is followed by an overview of survey paradata and a description of recent developments in this area of research. The data and measures sections are followed by the model specification and analysis strategy. The results are presented in Section 2.7. The paper concludes with a consideration of the conclusions to be drawn about how interviewer characteristics affect the magnitude of the design effect.

2.2 The interviewer design effect

Lesley Kish (1965) first coined the term 'design effect' (DEFF) initially to denote the effect of complex survey design features on survey estimates. Complex survey designs usually arise when features of the sample design, such as stratification and clustering, mean that sample units have unequal selection probabilities. Complex sample designs generally result in standard errors which are different from (usually larger than) what would be obtained using a simple random sampling (SRS) of the same size (Heeringa and Liu, 1998; Zelin and Stubbs, 2005). DEFF is item-specific and it is calculated as the ratio of the variance of the estimator

under the complex design to the variance using a simple random sample with the same sample size (Gabler *at al.*, 2006; Heeringa and Liu, 1998; Lynn and Gabler, 2005; Skinner and Vieira, 2005).

In multi-stage (or clustered) sample designs, the primary sampling units (PSUs) are usually geographical areas, variously defined. These geographical clusters are internally more homogenous than the population as a whole because people who live in close proximity to one another may be more alike and share similar socio-economic characteristics than is the case for the population as a whole. This phenomenon is called Tobler's first law of geography (Tobler, 1970). The degree of within-cluster homogeneity was referred to by Kish (1962) as the intra-class correlation. This concept was soon extended to include interviewers as well as areas as they manifest in the same way. The intra-interviewer correlation is based on the assumption that responses obtained by one interviewer are, on average, more similar to one another than are responses obtained by two different interviewers (Kish, 1962).

In order to minimise travel costs, each interviewer usually works in one primary sampling unit/cluster only. That means that the interviewer adds even more homogeneity to 'their' cluster, potentially increasing the design effect (Schnell and Kreuter, 2005). Moreover, such set-up means that the cluster and interviewer effects are completely confounded and thus are difficult to estimate. The initial approaches to estimating the interviewer effect such as Hansen Hurwitz and Bershad (HHB) and ANOVA models required a number of restrictive assumptions, including the requirement of an interpenetrated sample design, where interviewers are assigned to respondents at random (Bailar and Bailey, 1977; Bailar, 1976; Biemer and Stokes, 1985; Fellegi, 1964, 1974; Hansen *et al.*, 1961; Hartley and Rao, 1978; Kish, 1962; Koch, 1973; Sukhatme and Sukhatme, 1970). More recently, multilevel modelling has been used to deal with this problem of confounded effects (O'Muircheartaigh and Campanelli, 1998; Pickery, Loosveldt, and Carton, 2001; Schnell and Kreuter, 2005). The multilevel approach allows estimating the area and interviewer effects separately and uses a number of covariates at the respondent, area and interviewer level to control for the non-random interviewer assignment (Hox, 1994).

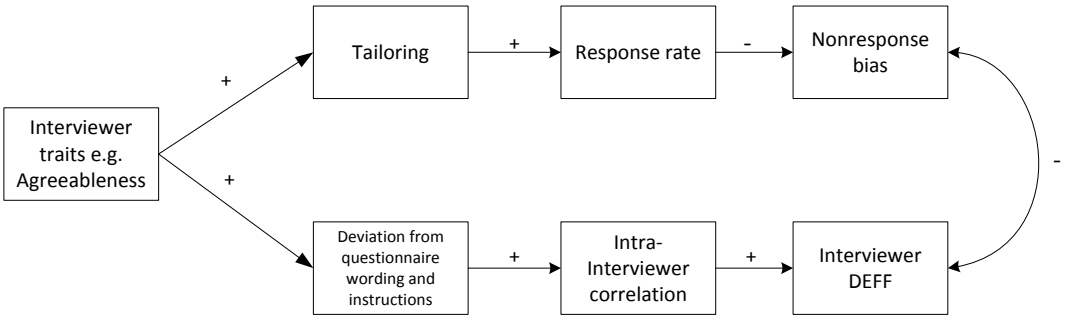
Research into the interviewer design effect has focused primarily on how different types of question may be differentially prone to interviewer effects rather than on characteristics of an interviewer (Couper and Kreuter, 2010; Davis and Scott, 1995). Research found that questions that are relatively more difficult or sensitive in nature are more prone to the interviewer effect. Questions about salient facts or closed questions are, in general, less

affected by the interviewer than questions asking about opinions or open questions (Schnell and Kreuter, 2005).

Where interviewer characteristics have been included as predictors, they have mostly been rather limited in scope and focused on demographic characteristics such as gender, age or interviewer status (for example, supervisor or area manager). The interviewer's gender turned out to explain some of the interviewer-related variance (O'Muircheartaigh and Campanelli, 1998), while age or interviewer status did not (Lipps, 2007). Matching the interviewer's and the respondent's gender in a survey concerning sexual behaviour yielded better data quality by decreasing the nonresponse (Catania *et al.*, 1996). The main reason for using only administrative information on interviewers is that, until recently, little if any data was routinely collected about the survey process, including information on interviewers. As paradata have been of growing interest to survey researchers, the availability of more detailed information on interviewers has improved.

Brunton-Smith *et al.* (2012) developed a model to explain how interviewers can influence both nonresponse and measurement error. This is set out in Figure 2.1. The authors suggested that interviewers' personality traits and attitudes could influence both type of errors. For example, interviewers who are high in the agreeableness personality trait may be better able to tailor their doorstep behaviour in order to obtain an interview. However, the very same trait may also lead them to adopt a more conversational interviewing style, relaxing the rules of standardised interviewing and deviating from the questionnaire script (Fowler and Mangione, 1990). Hence, any reduction in nonresponse error could be offset by an increase in measurement error as a result of their dependence on the same underlying common cause. Brunton-Smith *et al.* used cross-classified multilevel models with a complex variance term to separately identify the area and interviewer effects. They found that interviewers with the poorest record of obtaining contact also contributed the most variability in the survey estimates. Additionally, interviewers who were the least and the most successful in obtaining an interview were also responsible for a larger interviewer variance component than those in the middle of this distribution. This pattern suggests that indeed, those interviewers whose work potentially results in lower nonresponse error, due to high cooperation rates, are likely to introduce more measurement error instead.

Figure 2.1. A path model showing how a negative correlation between interviewer-induced nonresponse bias and measurement error might arise (source: Brunton-Smith *et al.*, 2012)



While interviewer personality and beliefs are specified as the common causal factors affecting both nonresponse and measurement errors, a limitation of the Brunton-Smith *et al.* (2012) study is that they do not actually measure these variables. Instead, they are simply assumed to be the common cause. The study presented here addresses this important shortcoming by using interviewer paradata – interviewer characteristics including their personality traits – and investigates the relationship between these variables and the magnitude of the intra-interviewer correlation. In other words, the research presented here includes direct measures of the primary causal variables in the diagram in Figure 2.1.

There is a body of research which uses interviewer attitudes and personality traits but it focuses almost entirely on nonresponse (Campanelli *et al.*, 1997; Morton-Williams, 1993; Summers and Beck, 1973). Although some studies have shown that certain personality traits such as extraversion, openness, agreeableness or persuasion skills can affect the cooperation rate, the direction of these relationships is not always as straightforward as one might expect. For example, while interviewers who are high in extroversion have more success on the doorstep, those who are more open and agreeable have a lower rate of success (Johnson and Price, 1988; Jäckle *et al.,* 2011).

Research concerned with interviewer-related variance in the measurement error context, instead of personality traits or skills, mostly employs respondent demographic characteristics such as gender, age and ethnicity (Cleary *et al.*, 1981; Hox, 1994; Pickery *et al.*, 2001a) or interviewing experience (Olson and Bilgen, 2011; Olson and Peytchev, 2007). There are also studies that have shown that interviewers' prior beliefs and expectations may lead to suggestive interviewing and skew the results (Clarke *et al.*, 2003; Singer *et al.*, 1983).

It has been found that interviewers can affect responses provided by their respondents according to the researchers' expectations if those were made known to interviewers before the interview. This may happen if the interviewer deliberately reads out only the certain response options available to the respondent (Smit *et al.*, 1997). So far, no study concerned with the interviewer effect has tried to explain this phenomenon using interviewer personality traits (Van Meter, 2005).

2.3 Paradata

The additional information on interviewers that is beyond the scope of basic administrative information such as personality traits can be considered as survey paradata. The term 'paradata' has emerged relatively recently and refers to a new form of survey process data. The term itself was first coined by Couper (1998) who defined paradata as data automatically collected by the computer systems supporting survey interviewing: Blaise[13] software in the Computer Assisted Personal Interview (CAPI) and in the Computer Assisted Telephone Interview (CATI), the two most popular methods of data collection in social surveys. Paradata, as initially defined by Couper, include time stamps, questionnaire routing, keystrokes, and how many times error or help windows were displayed. Since then, the term has evolved and broadened to include information supplied by interviewers such as household contact sheets recording each call to the household; interviewer's observations of the area; as well as comments and remarks on the particular interview. Also considered as paradata is information about interviewers themselves, including administrative information such as age, gender, and pay grade, as well as skills and personality traits obtained through separate data collection exercises (Kreuter and Casas-Cordero, 2010).

Paradata have been used for methodological research, especially in the area of survey process quality and survey error. Macro paradata (Scheuren, 2001) at the survey level, such as interim response rate, can be used to monitor and improve the data collection process (Groves and Heeringa, 2006; Laflamme *et al.,* 2008; Olson and Groves, 2009). Micro paradata (at the case level) have been used in research on nonresponse, for example to predict the best times to call in for an interview (Blom, 2009; Durrant, 2009; Durrant *et al.*, 2010; Kreuter *et al.*, 2007; Kreuter *et al.*, 2010; Sinibaldi, 2009). Paradata were also used, although to a lesser extent than for nonresponse, to assess the measurement error in surveys and data quality (Couper, 2009; Couper and Kreuter, 2010). This strand of research used solely computer-generated paradata rather than interviewer paradata. Paradata, which provide very detailed

[13]Blaise is purpose-designed computer software for CAPI and CATI, commonly used by survey agencies

information on interviewers, are rare. An innovation of this paper is the use of rich interviewer paradata to inform research on measurement error in surveys.

The popularity of paradata can be explained by the fact that they are inexpensive and often free, as they arise mainly as a by-product of the computerised survey process. Some paradata, though, require an additional cost such as observations about the address unit or area, or gathering more information on interviewers. Disadvantages of paradata include their usually complex structure, as the audit trail for a single case will include hundreds, if not thousands, of fields. Moreover, they are not always readily available and extracting them can be a lengthy process, although some improvement in this area can be seen (Taylor, 2008). Finally, some types of paradata are themselves prone to measurement error (Matsuo and Loosveldt, 2011; McCulloch *et al.*, 2010; Sinibaldi *et al.*, 2010; West, 2010).

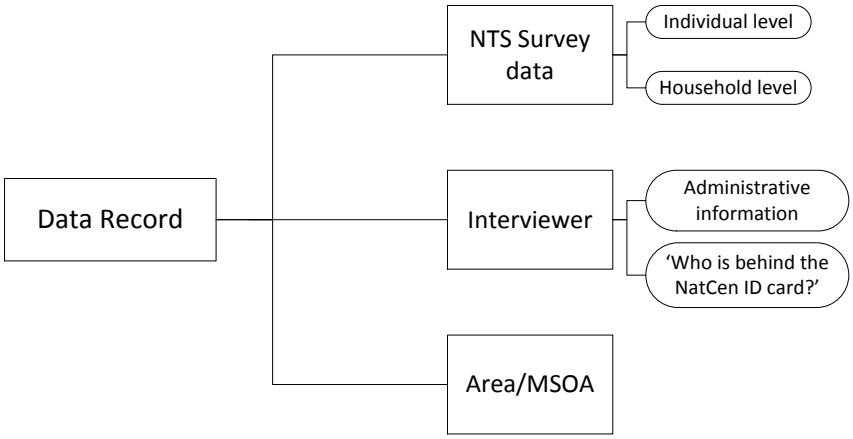
2.4 Data

The study described in this paper uses the National Travel Survey (NTS) linked with area characteristics at the Middle Super Output Area (MSOA) level (Martin, 2001). The NTS covers the whole of Great Britain[14] but the dataset used in this study is limited to cases for England only, because the area level information is not collected consistently across the constituent countries.

The information about interviewers comes from two different sources. The first is administrative information held by the data collection agency such as age, gender, pay grade and an indicator of which NTS administrations the interviewer had worked on. The second is a survey of all NatCen interviewers, which was conducted in 2008. Figure 2.2 shows a diagrammatic representation of the complete data record. NTS data are available at two levels: household (CAPI interview with the household reference person; usually the person with the highest income) and individual (CAPI interview with the household reference person and all the remaining household members including children). The household level dataset contains 40,244 cases and the individual level dataset contains 70,645 cases. The reduced number of cases relative to the complete NTS dataset is mostly due to the fact that not all interviewers who worked on the NTS between 2002 and 2008 took part in the NatCen interviewer survey.

[14]Excluding the Scottish islands and Isles of Scilly (0.2% of GB addresses)

Figure 2.2. Diagrammatic representation of the data structure for this study



2.5 Measures

Thirteen variables – survey items – were selected for analysis from the household (12 items) and individual (one item) questionnaires. The survey items are detailed in Table 2.1. They can be categorised in the following way: 'duration' questions asking about time taken to reach the local means of transport (in minutes) and 'satisfaction' questions asking about satisfaction with certain local amenities linked to transport, as well as reliability and frequency of those measured on a five-point Likert scale.

The area level (MSOA) characteristics used in this study were indicators computed using variables taken from the 2001 census as well as administrative data from the Office for National Statistics (ONS) and were transformed into uncorrelated components using Principal Component Analysis (PCA). This procedure yielded indicators of the following area-level characteristics: socio-economic disadvantage, urbanisation, population mobility, age profile, housing structure, crime rate and ethnic diversity (Brunton-Smith, 2008). The individual level NTS dataset contains 4,010 unique MSOAs.

The main section of the interviewer questionnaire asked interviewers about their personality traits and skills. The choice of variables used in partitioning the interviewer-related variance was informed by some general expectations. For example, interviewers with higher levels of conscientiousness or better attention to detail would potentially introduce less variance into the survey estimates than those scoring lower on those personality traits. Those who score higher on empathy or flexibility could be more likely to engage in a more 'personal' way of interviewing and hence introduce more variability than their counterparts.

PCA with oblique rotation was used to derive a common factor representing each personality trait or skill and individual-level scores were computed using the Anderson-Rubin method (Field, 2005). The full list of questionnaire items belonging to each factor is included in Appendix 2. Factors representing personality traits and skills include four out of the 'Big Five' personality dimensions[15] (Goldberg, 1990): openness, neuroticism, conscientiousness, extraversion, as well as empathy, efficiency, flexibility, attention to detail and communication skills. As these variables have been used to partition interviewer-related variance, they were turned from continuous variables into categorical indicators by splitting the scores into quartile groups, where group one have the lowest personality/skill score and group four have the highest.

[15]The agreeableness personality trait could not be successfully obtained in the PCA analysis and the items did not form a reliable scale as the Cronbach's Alpha coefficient was lower than 0.7

Table 2.1. NTS survey items analysed in this study

Variable name	Question wording	Special instructions
bus stop (*nearbus*)	About how long would it take (me) to walk from here to the nearest bus stop or place where I could get on a bus? I am interested in the nearest one even if it isn't the main one you use.	
railway station (*nearsta*)	Now thinking of your local train service, how long would it take (me) to walk to your nearest railway (that is, National Rail) station? (Again I am interested in the NEAREST one, even if it is not the main one or the one you use).	
station by bus (*bussta*)	How long would it take (me) to get to the railway station by bus? Please include any time spent walking but not waiting time.	
tube station (*neartube*)	How long would it take (me) to walk to your nearest [tube/metro/light rail/tram] stop?	
tube station by bus (*bustube*)	And how long would it take (me) to get to the nearest [tube/metro/light rail/tram] stop BY BUS? (Including any time spent walking but not waiting time?)	
walk (*i295*)	Thinking about your last walk of 20 minutes or more, how long did the walk take you, one way?	
cycle lanes provision (*xcyclane*)	How would you rate the provision of cycle lanes/paths locally?	showcard
pavement provision (*xpavement*)	How would you rate the condition of pavements locally where you live?	showcard
bus services provision (*xsatserv*)	Now I would like to ask some questions about your local bus services. By local, I mean services that operate near your home. How satisfied are you with your local bus services?	showcard
bus reliability (*xreliabus*)	First of all, how would you rate the reliability of local buses?	showcard
bus frequency (*xfrqbus*)	How would you rate the frequency of local buses?	showcard
tube reliability (*xrelmetro*)	How would you rate the reliability of the [tube/metro/light rail/tram]?	showcard
tube frequency (*xfrqmetro*)	How would you rate the frequency of the metro [tube/metro/light rail/tram]?	showcard

In addition to personality traits, Section Six of the questionnaire – persuading respondents – was used to provide information on interviewers' attitudes towards interviewing. The four-point scale items were transformed into binary variables where interviewers agreed or disagreed with seven statements on persuading respondents (Jäckle *et al.,* 2011). One of those was labelled 'determination'. Interviewers who agree with the statement – 'If a respondent is reluctant, a refusal should be accepted' – were labelled as less determined while those who disagreed with the statement were labelled as more determined. The interviewer questionnaire also included questions about satisfaction with various aspects of the job (questions four and five). The full questionnaire is attached in Appendix 1.

Finally, a measure of interviewer experience was derived. There were several variables available in the dataset that could be used as a proxy measure of interviewer experience. Time working for NatCen (in months), pay grade and the number of NTS assignments worked, which was eventually used. That is because the number of NTS administrations is a proxy of interviewer's experience directly related to this particular survey. Interviewers were divided into four groups:

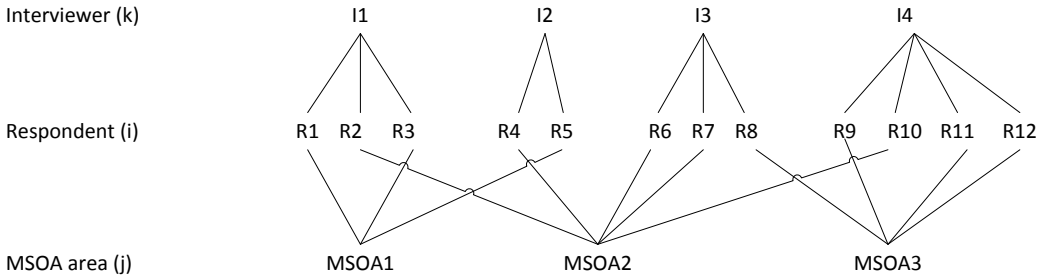
- Group 1: One NTS administration – 23.3%
- Group 2: Two to three NTS administrations – 31.1%
- Group 3: Four to five NTS administrations – 23.1%
- Group 4: Six or more NTS administrations– 22.4%.

2.6 Model

Following Brunton-Smith *et al.* (2012), a cross-classified multilevel model with complex Level 2 variance was used to estimate the interviewer effect. This model was introduced and described by Rashbash and Goldstein (1994). In general, multilevel models are used when the data have a hierarchical structure and lower level units, e.g. students, are nested within higher level units such as classes and schools. The multilevel approach allows estimation of the variability associated with each level in the data. For example, pupils' academic performance may be affected not only by the class in which they study but also the school they attend. Both class and school may have a different effect on their final exam marks (Snijders and Bosker, 2004). The cross-classified model is appropriate to use when the lower level units are nested within more than one classification at the same, higher level. An

example used by Rashbash and Browne (2008) is that of patients nested within hospitals and, at the same level, nested within neighbourhoods. Figure 2.3 shows the diagram for crossed structure of the data used in this paper.

Figure 2.3. Crossed structure of respondents nested within interviewers and MSOAs (adapted from Rashbash and Browne, 2008, p. 303)



The cross-classified model can be extended to incorporate a Level 2 random effect. This approach allows estimation of the area and interviewer effects first, and then partitioning the interviewer effect as a function of measured interviewer characteristics. The model has the form:

$$y_{i(j,k)} = \alpha + u_j + v_k + e_{i(j,k)} \quad (2.1)$$

where $y_{i(j,k)}$ is a response on item y of the *i*th respondent in the j/kth interviewer combination expressed as the sample mean value α; two random error terms express the area (u_j) and interviewer effect (v_k) and the respondent residual error term, $e_{i(j,k)}$. These error terms are assumed to be independent and identically distributed (IID) with zero mean and variances σ_j^2, σ_k^2 and $\sigma_{i(j,k)}^2$. The interviewer variance can be expressed as:

$$\rho_{int} = \frac{\sigma_k^2}{\sigma_j^2 + \sigma_k^2 + \sigma_{i(j,k)}^2} \quad (2.2)$$

and similarly the area variance is:

$$\rho_{MSOA} = \frac{\sigma_j^2}{\sigma_j^2 + \sigma_k^2 + \sigma_{i(j,k)}^2} \quad (2.3)$$

This is equivalent to Kish's intra-class correlation coefficient and is called a variance partition coefficient (VPC) in the multilevel modelling literature (Rasbash *et al.*, 2009). Each random effect has the following assumptions:

$$v_{0k} \sim N(0, \Omega_v) : \Omega_v = \sigma_{0k}^2$$
$$u_{0j} \sim N(0, \Omega_u) : \Omega_u = \sigma_{0j}^2 \quad (2.4)$$
$$e_{0ijk} \sim N(0, \Omega_e) : \Omega_e = \sigma_{0i(j,k)}^2$$

Brunton-Smith *et al.* (2012) used the model in (2.1) to include a complex variance term at the interviewer level. This allows the interviewer random effects to vary between groups of interviewers. The null model, with a complex Level 2 variance term, partitions σ_k^2 (interviewer variance) between two (or more) mutually exclusive interviewer groups. For example, if we are interested in the difference in interviewer variance across groups, defined by their level of determination to obtain an interview, and this determination variable is a binary variable, the model takes form:

$$y_{i(j,k)} = \alpha + u_j + v_{1k}\, less\ determined_k + v_{2k} more\ determined_k + e_{i(j,k)} \quad (2.5)$$

Then the interviewer-related variance for the two groups can be expressed as the variance-covariance diagonal matrix Ω_v:

$$\begin{bmatrix} v_{1k} \\ v_{2k} \end{bmatrix} \sim N(0, \Omega_v): \Omega_v = \begin{bmatrix} \sigma_{1k}^2 & 0 \\ 0 & \sigma_{2k}^2 \end{bmatrix} \quad (2.6)$$

where σ_{1k}^2 is the interviewer effect for less determined interviewers and σ_{2k}^2 is the interviewer effect for more determined interviewers. The covariances in Equation 2.6 are set to 0 because the interviewer groups are mutually exclusive.

All models were estimated using MLwiN version 2.13 (Rashbash *et al.*, 2009). The estimation technique used was the Monte Carlo Markov Chain (MCMC). Iterative Generalised Least Squares (IGLS) point estimates were used to provide the starting values for the MCMC chains (Browne, 2009). After extensive testing and the analysis of MCMC diagnostics, it was decided that the default MLwiN MCMC controls (burn-in length of 500, chain length 5000, thinning 1) were an acceptable trade-off between the accuracy of the estimates and the computation time required. Comparing complex multilevel models can be difficult. If models are estimated using MCMC methods, no equivalent of the log-likelihood statistic is available to enable model fit comparison. One solution suggested in the literature is to compare the Deviance Information Criterion (DIC). A decrease in DIC compared to the previous model suggests an improved model fit (Spiegelhalter *et al.*, 2002).

The model is used to estimate the interviewer effect on the respondents' replies to the 13 NTS items listed in Table 2.1. A range of covariates was added to the model to control for uneven sample composition across areas and interviewer assignments. The respondent covariates are: age, gender, employed vs. self-employed, employment status (including ILO unemployed), marital status and ethnicity (white vs. non-white). The interviewer covariates are: gender and age at the time of the interview.

2.7 Results

First, 13 continuous items from the National Travel Survey were identified as potentially prone to an interviewer design effect. Then cross-classified multilevel models were estimated to disaggregate the design effect into the area and interviewer effects. A range of area and respondent level covariates were added to the models. In the final models, the interviewer-related variance was partitioned using six paradata variables.

The values of DEFF were calculated for 13 NTS items using the formula:

$$DEFF = \frac{v\,(\bar{y})}{v_{SRS}\,(\bar{y})} \tag{2.7}$$

where y is the linear outcome in Equation 2.7, $v\,(\bar{y})$ is the variance of the sample mean under the complex sample design and $v_{SRS}\,(\bar{y})$ is the variance of the same sample mean $\bar{y}$ under an SRS design.

Table 2.2 shows that the unconditional DEFF values for sample means for all 13 NTS items are subject to a design effect ranging from DEFF = 1.49 to DEFF = 4.63. These were calculated in STATA using the 'svy' commands and Taylor linearisation[16]. These DEFF values were calculated using the original NTS design and NTS PSUs rather than MSOAs. The mean DEFF based on 13 items is 2.47. The duration items in general showed larger DEFF values (mean DEFF = 3.36) than the satisfaction items (mean DEFF = 1.71), which is partially due to the different measurement scales used. In order to reduce the burden of a long questionnaire on the respondent, certain NTS questions were administered every other year. Other items analysed in this study were dropped along the years (e.g. condition of the pavement). This is the reason why the number of cases for each item in Table 2.2 varies.

[16] http://www.stata.com/capabilities/survey-commands/

Table 2.2. Standard errors (SE) under SRS and complex design and DEFF values for NTS survey items

Variable	N	Mean	SE(SRS)	SE (NTS design)	DEFF
bus stop	39616	4.400318	0.0245318	0.0326941	1.78
railway station	32223	31.92313	0.1240233	0.2667602	4.63
station by bus	26689	22.35951	0.0929896	0.1881383	4.09
tube station	3509	15.92334	0.2675676	0.5517641	4.25
tube station by bus	1975	29.08354	0.5193673	0.9738508	3.52
walk	51723	33.07718	0.1149091	0.1595565	1.92
cycle lanes provision	3618	2.994748	0.0222506	0.0304859	1.88
pavement provision	5988	2.85855	0.0163104	0.0213476	1.71
bus services provision	12868	3.689773	0.010689	0.013333	1.56
bus reliability	13467	3.849584	0.0090212	0.113232	1.58
bus frequency	13698	3.822164	0.0093061	0.0132394	2.02
tube reliability	11815	4.006602	0.0086672	0.0105814	1.49
tube frequency	11884	4.070178	0.0077526	0.0103425	1.78

Having established that all 13 items are subject to a design effect, the second stage was to partition this overall design effect into area and interviewer components. A cross-classified multilevel model (Equation 2.1) was used to separate the two effects and their VPCs were calculated. The results are summarised in Table 2.3.

Table 2.3. Variance partition coefficient (VPC) for the interviewer and area

Variable	Total variance	VPC (Interviewer)*	VPC MSOA (Area)*	VPC Residual*	N
bus stop	214.005	4.8%	46.1%	49.1%	39616
railway station	523.149	14.4%	46.3%	39.2%	32223
station by bus	23.983	0.6%	11.5%	87.9%	26689
tube station	335.62	2.0%	79.7%	18.3%	3509
tube station by bus	557.83	1.6%	67.9%	30.5%	1975
walk	638.705	1.9%	4.5%	93.6%	51723
cycle lanes provision	1.196	4.6%	17.4%	78.0%	3618
pavement provision	1.091	3.4%	8.7%	87.9%	5988
bus services provision	1.806	7.5%	11.6%	80.8%	12868
bus reliability	1.591	3.4%	7.4%	89.3%	13467
bus frequency	1.475	3.7%	10.5%	85.8%	13698
tube reliability	0.895	5.0%	6.8%	88.2%	11815
tube frequency	0.725	6.8%	12.6%	80.7%	11884

*All variances are significantly different from 0.

It is clear from Table 2.3 that for all 13 items, the interviewer effect is considerably smaller than the area component. This is likely to be because the topic of the NTS is travel and all questions are strongly area-related. Obviously, the frequency of buses or the condition of the pavement will be strongly correlated within local areas. However, the results also show that a significant minority of the total variance can be attributed to interviewers. On average, across all 13 items, interviewers are responsible for 4.6% of the total design effect. This percentage is lower for the duration items – on average 4.2% can be attributed to interviewers – and it is higher for the satisfaction items – on average 4.9% of those can be attributed to interviewers. Similar values for the proportion of interviewer-related variance were reported in Lipps (2007) and Hox (2002). Schnell and Kreuter (2005) reported interviewer contribution to the overall variance to be below 2%; however, the mean DEFF for their study was only 1.48. The smallest proportion of interviewer-related variance, at 0.3%-0.6%, was reported for a health survey that used a small network of highly skilled interviewers (Davis and Scott, 1995).

When the area and respondent covariates were added, both variance components decreased, as summarised in Table 2.4. In almost all cases, both variance components were reduced in magnitude. As expected, interviewer-related variance was less affected than area-related variance. This is because area-related variance is likely to be explained by area

characteristics to a greater extent than interviewer-related variance; conversely, interviewer-related variance is likely be explained to a greater extent by interviewer characteristics.

Table 2.4. Variance partition coefficients for null models and models with covariates

	Null models		Models with covariates		Difference	
Variable	**Interviewer**	**Area**	**Interviewer**	**Area**	**Interviewer**	**Area**
bus stop	4.8%	46.1%	4.5%	40.8%	-0.3%	-5.3%
railway station	14.4%	46.3%	10.1%	43.2%	-4.3%	-3.1%
station by bus	0.6%	11.5%	0.4%	8.1%	-0.2%	-3.4%
tube station	2.0%	79.7%	2.4%	74.3%	0.4%	-5.4%
tube station by bus	1.6%	67.9%	1.8%	63.2%	0.2%	-4.7%
walk	1.9%	4.5%	1.9%	3.7%	0.0%	-0.8%
cycle lanes provision	4.6%	17.4%	4.0%	13.6%	-0.6%	-3.8%
pavement provision	3.4%	8.7%	3.3%	8.2%	-0.1%	-0.5%
bus services provision	7.5%	11.6%	6.9%	11.4%	-0.6%	-0.2%
bus reliability	3.4%	7.4%	3.4%	6.2%	0.0%	-1.2%
bus frequency	3.7%	10.5%	2.8%	9.2%	-0.9%	-1.3%
tube reliability	5.0%	6.8%	5.0%	6.5%	0.0%	-0.3%
tube frequency	6.8%	12.6%	6.7%	12.3%	-0.1%	-0.3%

The next step was to include the interviewer paradata variables, as described in Section 2.5 and decide which are the most suitable to be used to partition the interviewer-related variance. Interviewers' satisfaction with pay can potentially be an important indicator of interviewer performance and it could be somehow manipulated by the survey agency. Interviewers' determination based on the their attitudes towards interviewing, has already turned out to be statistically significant when predicting interviewers' cooperation rates and could also affect interviewer-related variance (Jäckle *et al.,* 2011). Communication skills are one of the most important skills required from interviewers hence this variable was also used to partition the interviewer-related variance. Olson and Peytchev (2007) have shown that interviewer experience can affect interview pace and thus affect data quality while Lipps (2007) noticed larger proportions of item nonresponses amongst more experienced interviewers. For that reason, interviewers' NTS experience was also considered. Finally, two personality variables were chosen: extroversion and conscientiousness. Extroversion is a variable already used in research by Hox (1994) and turned out to be a statistically significant predictor of interviewer-related variance. Moreover, Jäckle *et al.* (2011) found that extroversion was positively correlated with higher cooperation rates amongst interviewers.

Conscientiousness is one of interviewers' characteristics that Brunton-Smith *et al.* (2012) considered to potentially affect both interviewer-related variance and cooperation rates. These six variables were used to partition the interviewer-related variance across all 13 models and the results are presented below.

2.7.1 Satisfaction with pay

The 'satisfaction with pay' binary variable had an approximately even distribution of those who were satisfied with pay (53.7%) and those who were not satisfied with pay (46.3%). For all 13 NTS items, the interviewer variance component was estimated using Equation 2.6. Due to limited space, the full set of parameter estimates are presented by way of illustration for only a single variable, 'railway station' (time taken to walk to the nearest station) in Table 2.5. The results for the remaining models are presented using summary statistics of the key variance component parameters in Figures 2.3 and 2.4. The results in Table 2.5 show that interviewers who are not satisfied with their pay introduce more variance into the estimate for the 'railway station' variable than interviewers who are satisfied with their pay. The DIC value in the random component model has decreased from 259939 to 259927, suggesting that introduction of the Level 2 variance term improves the fit of the model to the data and that the interviewer variance components are significantly different between the two groups.

Table 2.5. The 'railway station' model with the interviewer-related variance partitioned by the 'satisfaction with pay' variable

	Model with covariates	**Model with complex Level 2 variance**
Interviewer variance (σ_k^2)	8.651 (1.209)	
satisfied with pay (σ_{1k}^2)		4.955 (1.123)
not satisfied with pay (σ_{2k}^2)		13.098 (2.144)
Area variance (σ_j^2)	303.756 (8.489)	303.806 (8.269)
Residual variance ($\sigma_{i(j,k)}^2$)	167.965 (1.435)	167.955 (1.433)
DIC	259939.48	259927.547
Classification:		
Interviewers	398	398
MSOA	3603	3603
Respondents	32223	32223

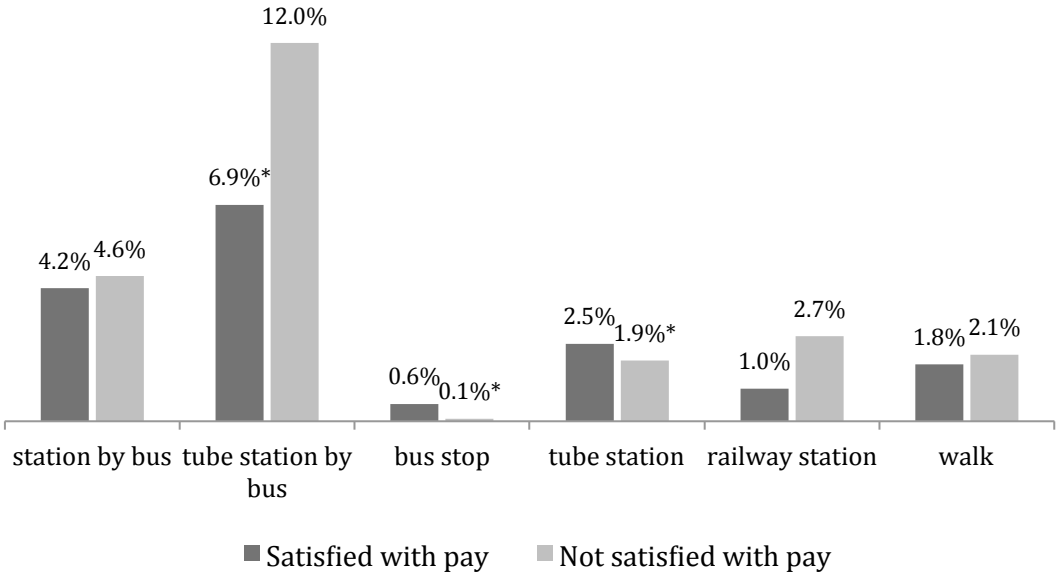
The VPCs were calculated for each interviewer group as follows:

$$VPC_{satisfied} = \left(\frac{\sigma^2_{1k}}{\sigma^2_{1k} + \sigma^2_j + \sigma^2_{i(j,k)}}\right) \quad (2.8)$$

$$VPC_{not\ satisfied} = \left(\frac{\sigma^2_{2k}}{\sigma^2_{2k} + \sigma^2_j + \sigma^2_{i(j,k)}}\right) \quad (2.9)$$

where σ^2_{1k} is the variance introduced by the interviewers satisfied with pay and σ^2_{2k} is the variance introduced by the interviewers not satisfied with pay; the σ^2_j and $\sigma^2_{i(j,k)}$ are the area and residual variance components. The use of VPCs means that the group-specific variance components can be easily compared for each model but also across all models as the VPC means. Continuing with the 'railway station' variable example, the interviewers who were satisfied with their pay were responsible for only 1% of the total variance for this item while their colleagues, who were not satisfied with pay, introduced 2.7% of the total variance. Overall, it is a small effect; however, those not satisfied with their pay introduced three times more variance than their counterparts. Across all six duration items, interviewers who were satisfied with pay contributed, on average, 2.8% of the total variance, while those were not satisfied introduced, on average, 3.9%. As can be seen in Figure 2.4, for four items out of six, interviewers who were not satisfied with pay introduced more variance than their counterparts. There is a particularly large difference between the two interviewer groups for the 'tube station by bus' variable. Those who were not satisfied with pay introduced twice as much variance (12%) than interviewers who were satisfied with pay (6.9%). However, the variance component associated with interviewers satisfied with pay is not statistically significant, suggesting that only interviewers not satisfied with pay significantly contributed to the variance of this estimate.

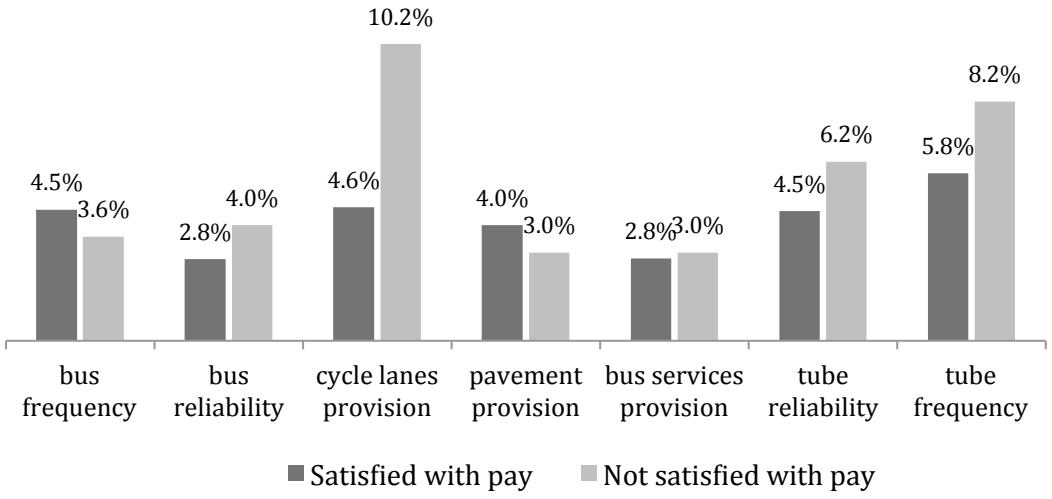
Figure 2.4. Proportion of the total variance introduced by both interviewer groups (satisfaction with pay) – duration items:



* =the variance component is not statistically significant for this interviewer group

The same relationship can be seen for seven satisfaction items in Figure 2.5. Interviewers who were satisfied with pay introduced, on average, 4.1%, while those who were not satisfied introduced 5.5% of the total variance. In general, satisfaction items appear to be more prone to the interviewer effect: on average, interviewers were responsible for 4.8% of the total variance across the satisfaction items, while across the duration items, interviewers introduced only 3.4% of the total variance. As Figure 2.5 shows, for five items out of seven, interviewers who were not satisfied with pay introduced more variance than their counterparts. A particularly large difference can be seen for the 'cycle lanes provision' item where interviewers not satisfied with pay introduce twice as much variance (10.2%) as their counterparts (4.6%).

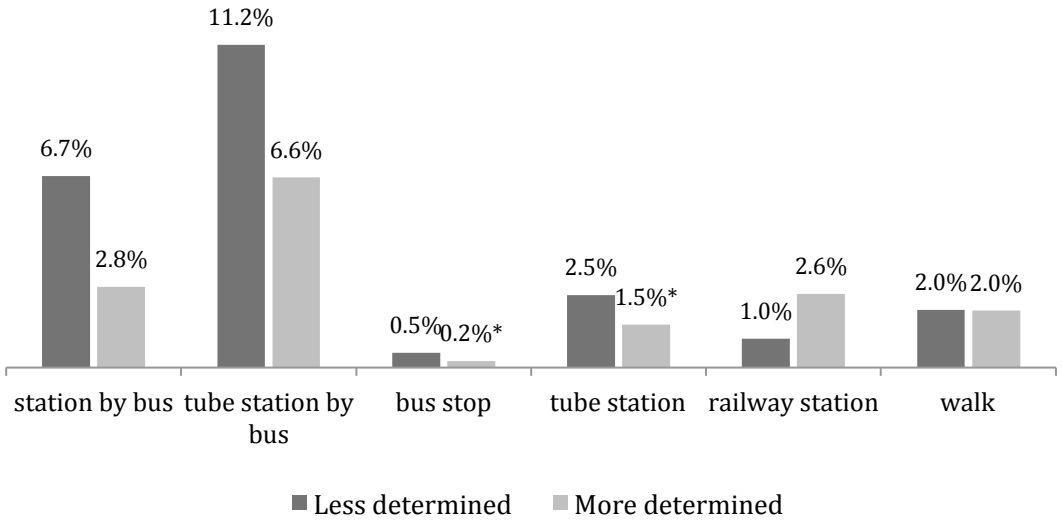
Figure 2.5. Proportion of the total variance introduced by both interviewer groups (satisfaction with pay) – satisfaction items:



2.7.2 Determination to obtain an interview

The second paradata variable used to partition the interviewer variance component was the binary 'determination' variable that split interviewers into more and less determined groups. Both groups were approximately equal in size as 46% were less determined and 54% were more determined. In five models out of 13, the addition of the complex variance term resulted in improved fit. Figure 2.6 shows the interviewer variance components estimated for the two groups across the six duration items. Interviewers in the more determined group introduced less variance, on average 3.5%, than interviewers in the less determined group, who on average introduced 4.8% of the total variance. For four items out of five, the less determined interviewers introduced more variance, while for the 'walk' item, there was no difference between the two groups. Across all duration items, the less determined interviewers introduced on average 4.0% of the total variance while their counterparts introduced only 2.6% of the total variance. Amongst the duration items in Figure 2.6, two show rather large differences between the two interviewer groups. These are the 'station by bus' and 'tube station by bus' items; in both cases the less determined interviewers introduced twice as much variance as their counterparts.

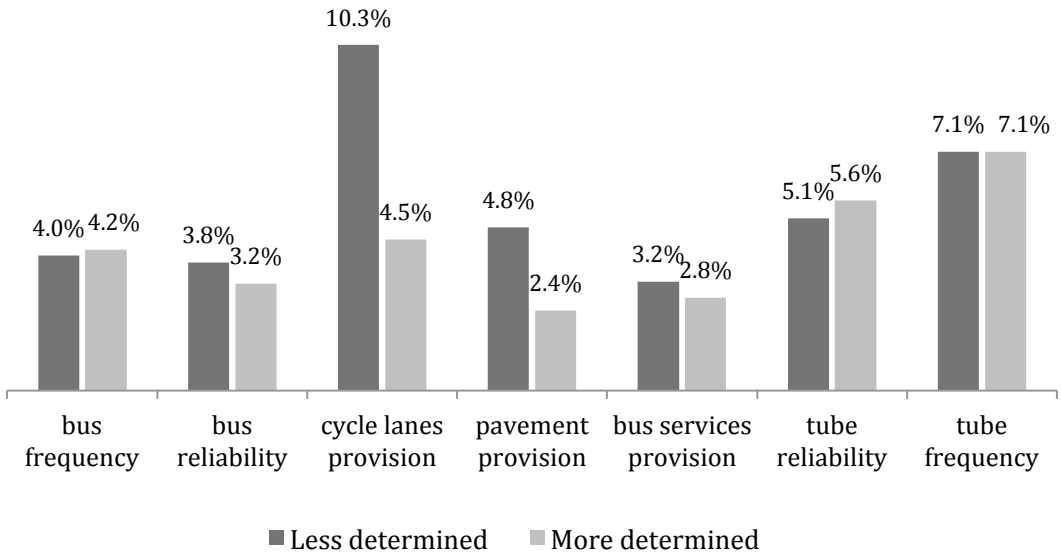
Figure 2.6. Proportion of the total variance introduced by both interviewer groups (determination) – duration items:



* = the variance component is not statistically significant for this interviewer group

As shown in Figure 2.7, for four items out of seven, the less determined interviewers introduced more variance than their counterparts. For one item there was no difference. On average, the less determined interviewers introduced 5.5% of the total variance, while the more determined interviewers introduced on average 4.2% of the total variance – 1.3 percentage points less than their less determined colleagues. For the 'cycle lanes provision' the less determined interviewers introduced twice as much variance (10.3%) as their more determined colleagues (4.5%).

Figure 2.7. Proportion of the total variance introduced by both interviewer groups (determination) – satisfaction items:

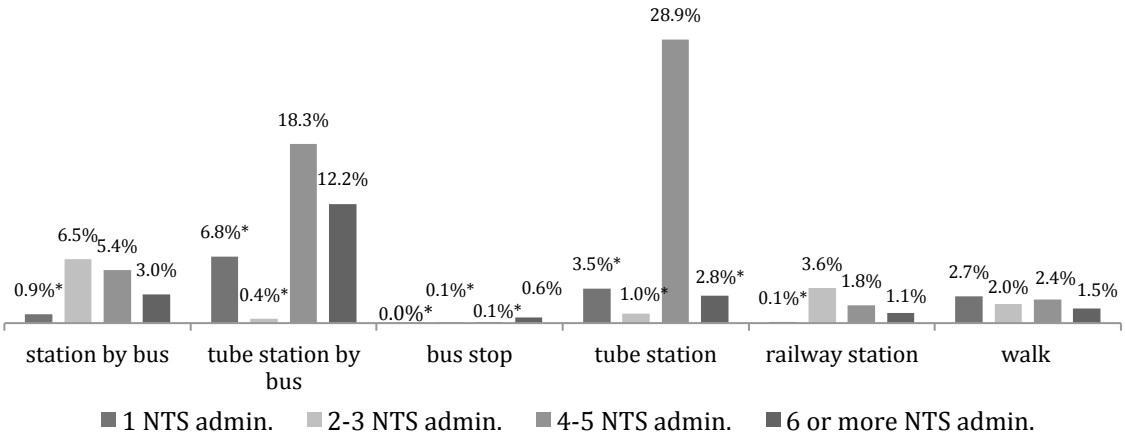


2.7.3 Interviewer experience

The specific NTS interviewer experience, expressed as the number of NTS administrations they had worked on, was used to partition the interviewer-related variance into four groups. Out of the 13 models, eight resulted in lower DIC values suggesting a better model fit if the complex variance component was included.

The variance introduced by interviewers working on the NTS for the first time was non-significant in all models except for the 'walk' variable. That means these interviewers did not contribute to the increased variance of the survey estimates. The interviewers who introduced the most variance amongst the duration items were those with substantial experience (four to five NTS administrations). On average, they introduced 9.5% of the total variance. However this is mainly due to the 'tube station' item, for which they introduced 28.9% of the total variance, a suspiciously high figure (Figure 2.8).

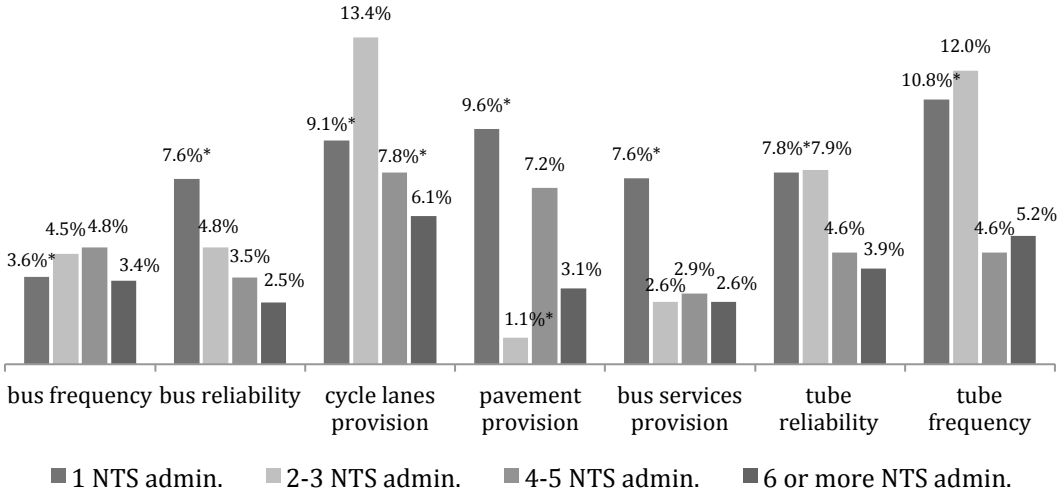
Figure 2.8 Proportion of the total variance introduced by four interviewer groups (NTS experience) – duration items:



* = the variance component is not statistically significant for this interviewer group

In general, satisfaction items are more prone to the interviewer effect, as can be seen in Figure 2.9. Amongst satisfaction items, 'cycle lanes provision' and 'tube frequency' items exhibit large differences between interviewers with some experience, two to three NTS administrations, and those with the most experience, six NTS administrations or more. Those in the former group introduced more than twice the variance of the latter group. The 'pavement provision' item shows a slightly different pattern, where those with substantial NTS experience, four to five NTS administrations, introduced twice as much variance (7.2%) than those with the most NTS experience (3.1%).

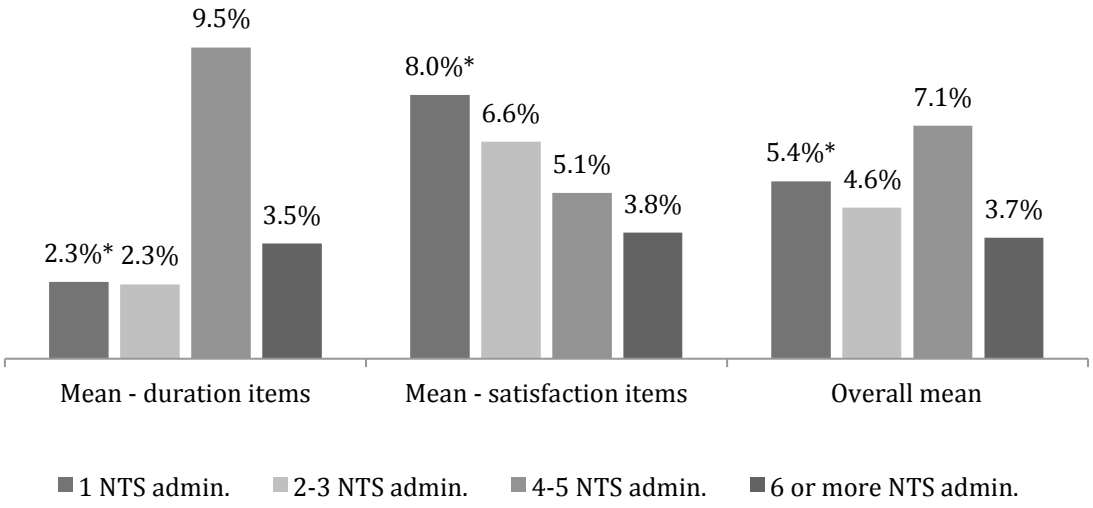
Figure 2.9 Proportion of the total variance introduced by four interviewer groups (NTS experience) – satisfaction items:



* = the variance component is not statistically significant for this interviewer group

Figure 2.10 shows the mean proportion of the total variance introduced by four interviewer groups across all duration and satisfaction items. A clear pattern can be observed for the satisfaction items where the mean variance introduced by each group decreases monotonically as the interviewers' experience increases (Figure 2.10). However, this is not the case for duration items where the mean variance introduced by those with substantial experience (four to five NTS administrations) is much higher than for any other group.

Figure 2.10. Mean proportion of the total variance introduced by four interviewer groups (NTS experience):

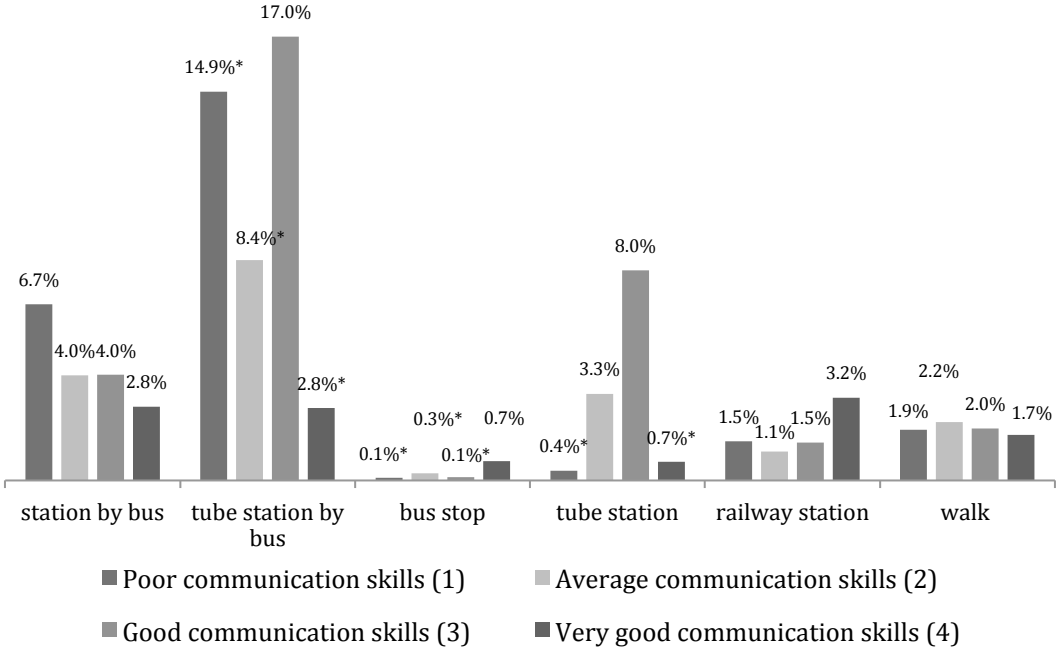


* = the variance component is not statistically significant for this interviewer group

2.7.4 Communication skills

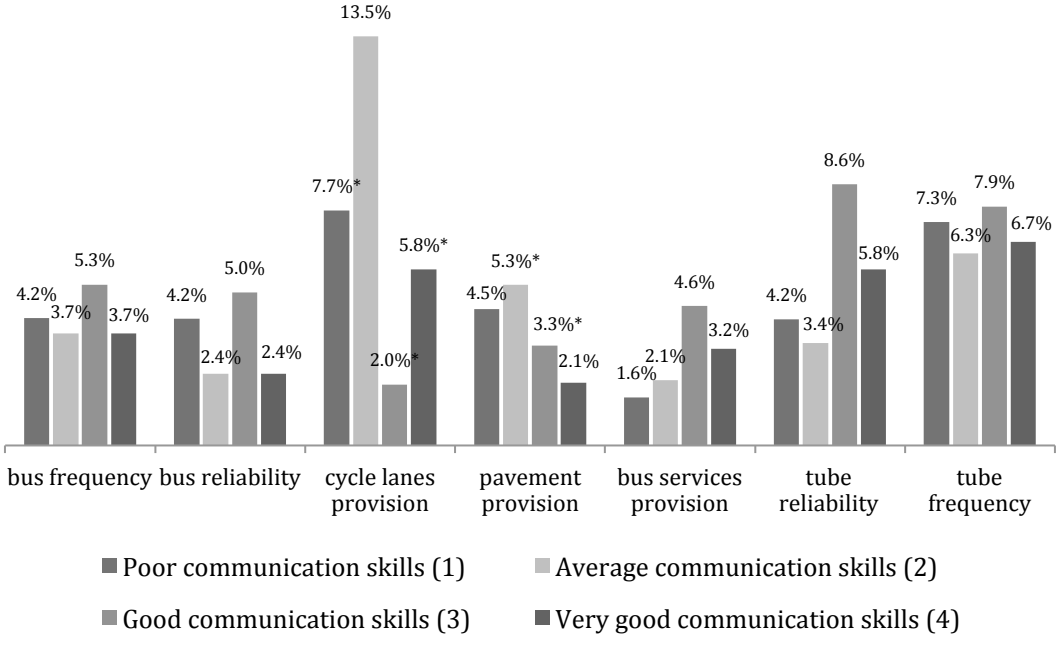
The 'communication skills' variable is a factor score representing the level of communication skills, including persuasion, amongst interviewers. Amongst the 13 models, seven resulted in lower DIC values, suggesting an improved model fit after partitioning the interviewer-related variance component. Amongst the six duration items, interviewers with good communication skills introduce substantially more variance for the 'tube station by bus' and 'tube station' items (Figure 2.11). Amongst the seven satisfaction items shown in Figure 2.12, the 'cycle lanes provision' item exhibits a similar pattern to previous models with one group of interviewers substantially different from the others. Those with average communication skills introduced twice as much variance (13.5%) than any other interviewer group (7.7% for the group with poor communication skills) and only 'their' variance component is statistically significant (Figure 2.12).

Figure 2.11 Proportion of the total variance introduced by four interviewer groups (communication skills) – duration items:



* = the variance component is not statistically significant for this interviewer group

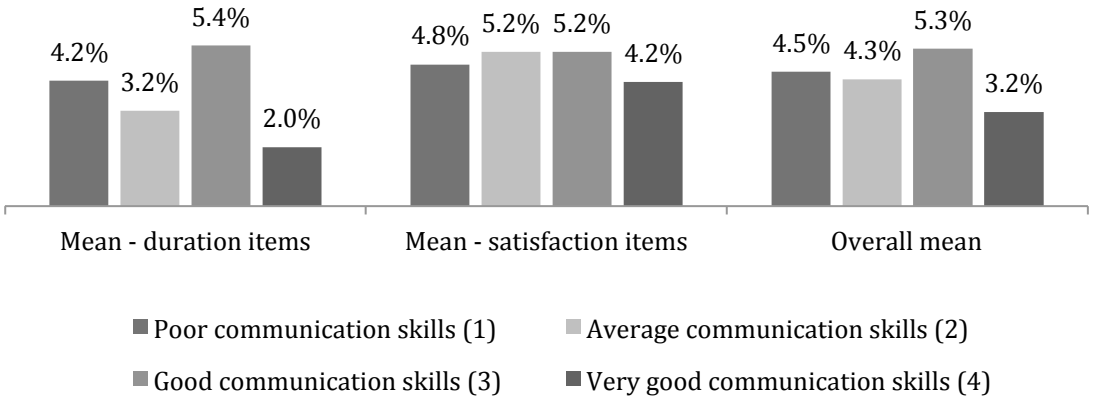
Figure 2.12 Proportion of the total variance introduced by four interviewer groups (communication skills) – satisfaction items:



* = the variance component is not statistically significant for this interviewer group

The mean proportion of the total variance across all duration and satisfaction items for each interviewer group is presented in Figure 2.13. No obvious patterns can be noticed apart from the fact that interviewers with very good communication skills introduced, on average, the least variance into the survey estimates.

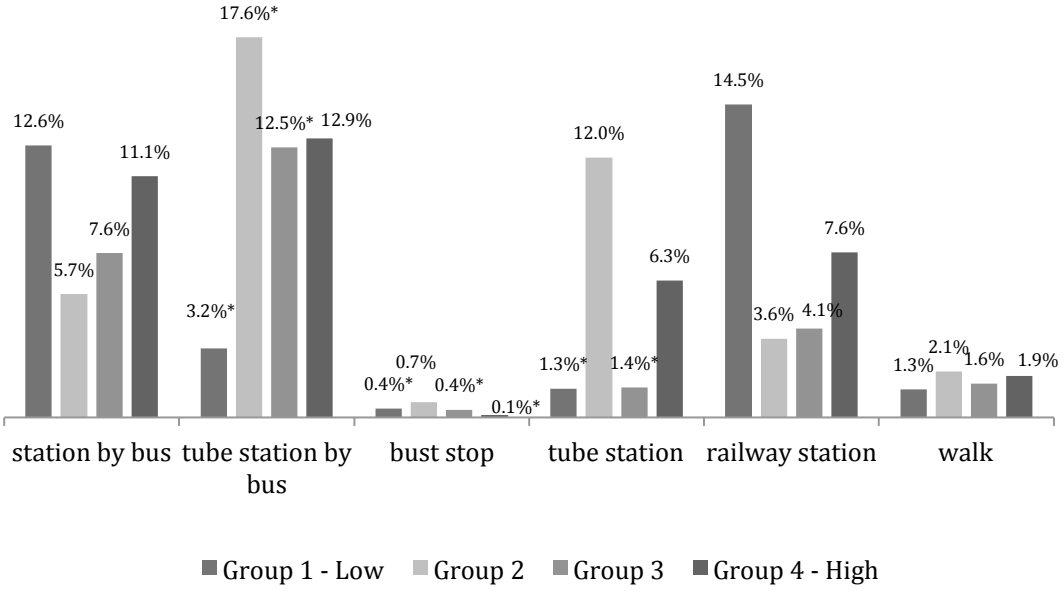
Figure 2.13. The mean proportion of the total variance introduced by four interviewer groups with different levels of communication skills:



2.7.5 Conscientiousness

Interviewers were split into quartile groups according to the level of self-reported conscientiousness. Interviewers with the lowest levels of conscientiousness were in group one and those with the highest levels were in group four. Partitioning the interviewer-related variance component resulted in an improved model fit in five models out of 13. Interviewers in group one introduced the most variance for the 'railway station' and 'station by bus' items (Figure 2.14). It is also worth noticing that the 'bus stop' and 'walk' items attracted very little interviewer-related variance.

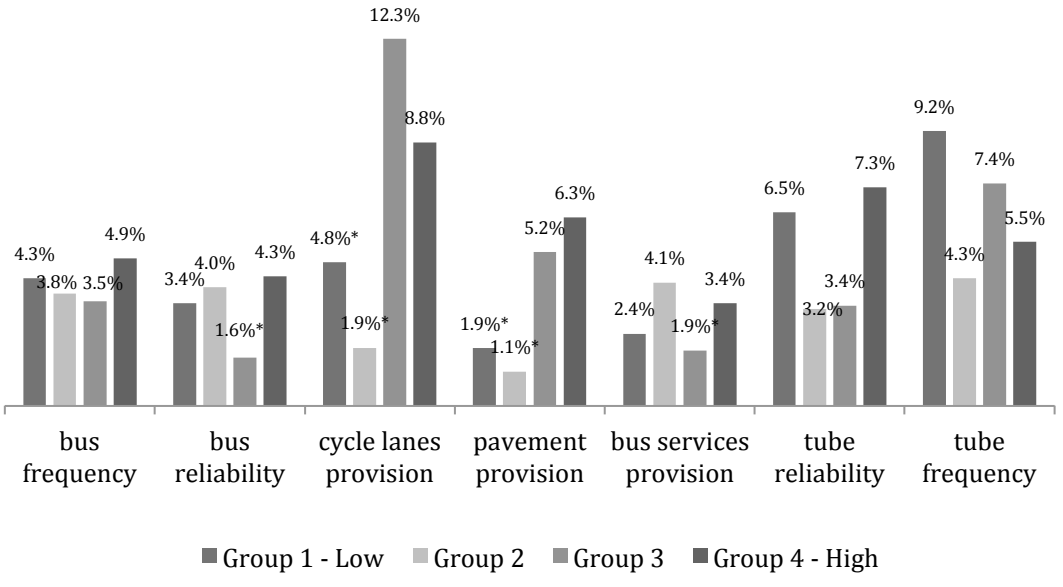
Figure 2.14 Proportion of the total variance introduced by four interviewer groups (conscientiousness) – duration items:



* = the variance component is not statistically significant for this interviewer group

As can be seen in Figure 2.15, the satisfaction items were more uniformly affected by the interviewer, comparing to the duration items. Substantial differences between different interviewer groups can be noticed for the 'cycle lanes provision' items and items concerning tube.

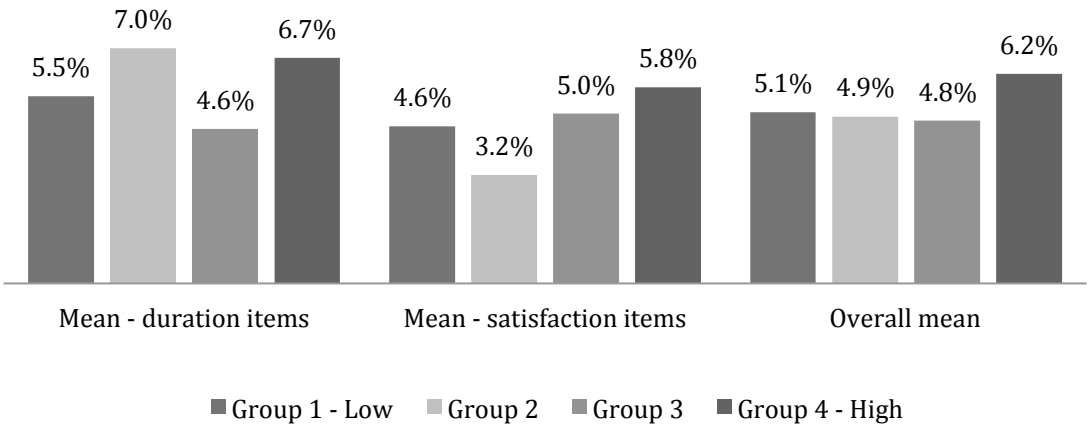
Figure 2.15 Proportion of the total variance introduced by four interviewer groups (conscientiousness) – satisfaction items:



* = the variance component is not statistically significant for this interviewer group

As Figure 2.16 shows, interviewers in group two and group four introduced the most variance (7% and 6.7% respectively) into the survey estimates across all duration items. Interviewers in group four introduced, on average, the most variance amongst all satisfaction items (on average 6.7%) and also across all items (6.2%).

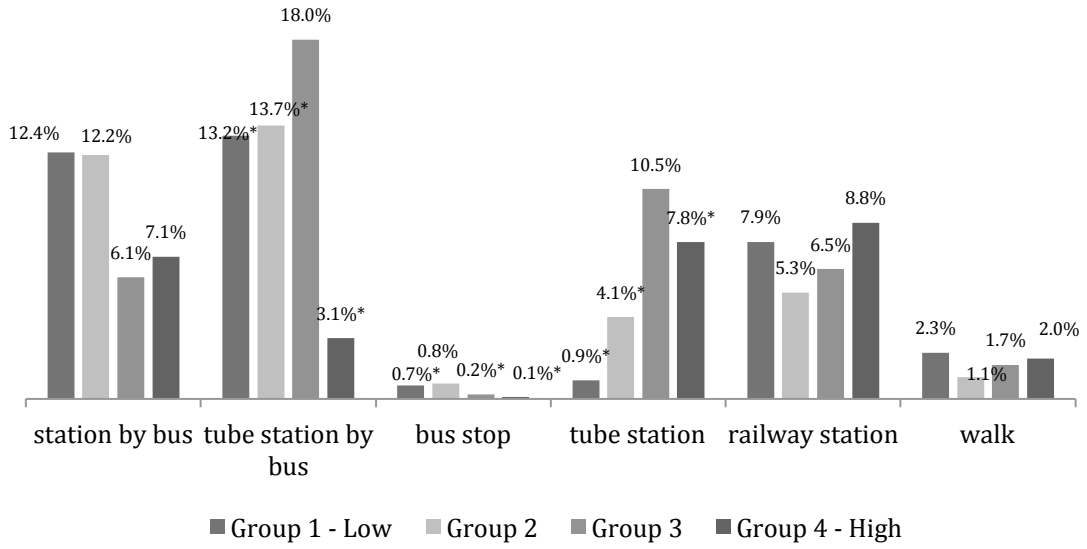
Figure 2.16. The mean proportion of the total variance introduced by four interviewer groups with different levels of conscientiousness:



2.7.6 Extroversion

The final paradata variable used to partition the interviewer-related variance was self-reported extroversion. As with the conscientiousness variable, interviewers were split into quartile groups where group one exhibited the lowest levels of extroversion and group four the highest. Extroversion turned out to be the 'worst performing' variable as only four models out of 13 yielded lower DIC values, suggesting an improved model fit once the interviewer-related variance component was partitioned. Interviewers in group two introduced the most variance into the survey estimates of 'tube station by bus' and 'tube station' items. The 'bus stop' item and 'walk' item were again the least prone to the interviewer effect (Figure 2.17).

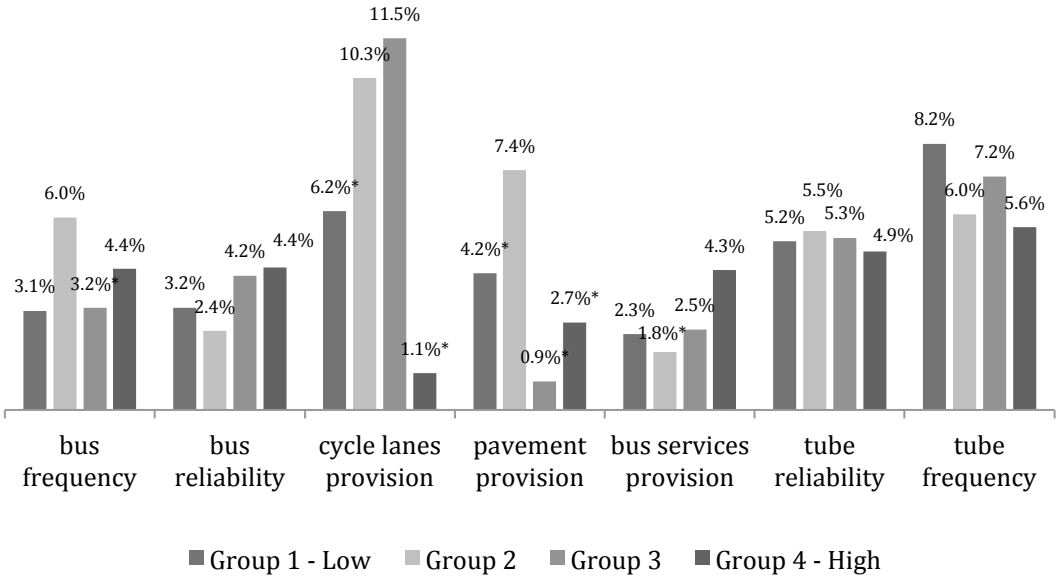
Figure 2.17 Proportion of the total variance introduced by four interviewer groups (extroversion) – duration items:



* = the variance component is not statistically significant for this interviewer group

As shown in Figure 2.18, the differences in the proportion of total variance introduced by interviewers are less apparent between the satisfaction items. The 'cycle lanes provision' and 'pavement provision' items exhibit the most varied pattern, with large differences between interviewer groups. Interviewers in group two introduced the most variance into survey estimates of these two items (11.5% and 7.4% respectively).

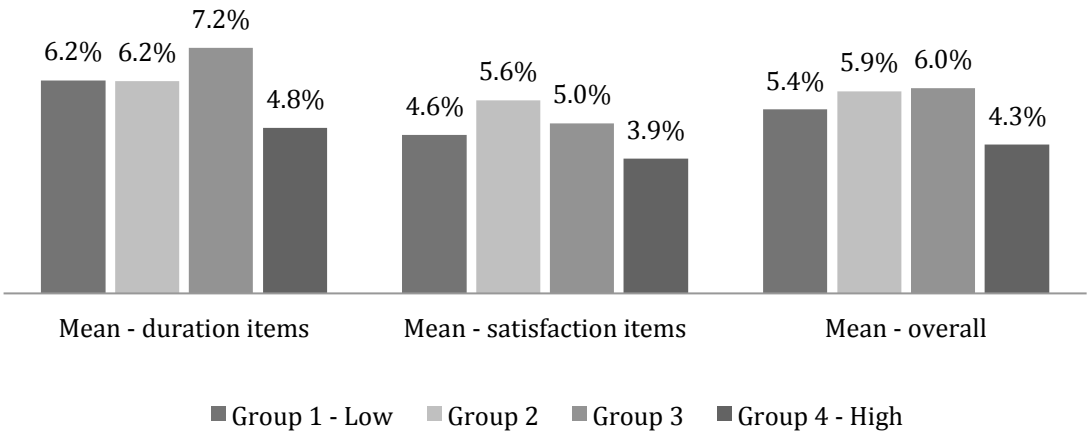
Figure 2.18 Proportion of the total variance introduced by four interviewer groups (extroversion) – satisfaction items:



* = the variance component is not statistically significant for this interviewer group

Figure 2.19 shows the mean proportion of total variance introduced by four interviewer extroversion groups. In general, interviewers with the highest extroversion levels introduced the least variance into the survey estimates across both duration and satisfaction items. Interviewers in the two middle groups introduced on average the same amount of variance across all items.

Figure 2.19. The mean proportion of the total variance introduced by four interviewer groups with different levels of extroversion:



2.8 Discussion

It has long been known that interviewers are an important source of measurement error in social surveys but the research literature in this area fails to explain the interviewer effect in terms of interviewers' personality traits and skills, mostly due to the lack of availability of such information. Another obstacle to estimating an interviewer effect are the complex sampling designs commonly used in contemporary survey research, in which the area and interviewer effects are completely confounded. Research presented in this paper took advantage of the availability of extensive interviewer-related paradata associated with the NTS. Availability of such paradata allowed explanation of the identified interviewer-related variance of the survey estimates with various interviewers' characteristics. Moreover, the study took advantage of the NTS rolling sample design in which a natural cross-classification of interviewers and areas occurs. This set-up allowed the interviewer and area effects to be disentangled in the multilevel cross-classified models. Additionally, a range of covariates at the area and respondent level were added to the models to control for lack of random interviewers' assignment.

Thirteen NTS continuous items were identified for analysis and all were subject to a survey design effect. Considering the topic of the NTS, unsurprisingly this effect was mostly attributed to the area but the interviewer effect, although small, was also found and ranged

from 0.6% to 14.4% of the total variance, depending on the item. The mean interviewer variance component across all 13 items is 4.6% of the total variance. Controlling for area and respondent characteristics decreased some of the interviewer-related variance but not for all items. The 'satisfaction' items were more prone to an interviewer effect; that is, proportionally more variance was attributed to interviewers than for 'duration' items, which were more factual in nature and less open to interpretation. Another reason why satisfaction items were more affected by the interviewer is that showcards were used with the Likert Scales. The use of showcards means more interviewer involvement, hence increased potential for the interviewer effect. This is in line with other research in this area (O'Muircheartaigh and Campanelli, 1998; Schnell and Kreuter, 2005).

Six interviewer attribute paradata variables were used to partition the variance they introduced: satisfaction with pay, determination, NTS experience, level of communication skills, conscientiousness and extroversion. In total, 78 models were fitted and just under half (37) of them showed lower DIC values, suggesting that the complex variance component results in an improvement to model fit. In other words, the differences between interviewers introducing different levels of variance are statistically significant. Some paradata variables performed better than others in this respect. For example, the differences between different levels of interviewers' NTS experience were statistically significant in nine models out of 13. That would suggest that NTS experience is an important factor affecting interviewers' performance and results in better data quality, as more experienced interviewers introduced less variance into the survey estimates. Conversely, differences between interviewer groups with different levels of extroversion were statistically significant in only four models out of 13. This result suggests that there may not be a clear association between extroversion and interviewer effect.

Interviewers not satisfied with their pay, those who were less determined, those with substantial NTS experience (four to five administrations) and those with good communication skills were found to introduce on average more variance than their counterparts. For example, the less determined interviewers introduced more variance because they may have been 'less bothered' to secure the interview and thus may have been 'less bothered' to complete the task diligently. Similar reasoning may be behind the fact that interviewers not satisfied with their pay introduced, on average, more variance than their satisfied counterparts. The finding that those with the most NTS experience (rather than survey experience as such) introduce on average less variance is somewhat contradictory to the one presented by Olson and Peytchev (2007) who discovered that interviewers with

more experience take less time to conduct interviews, potentially negatively affecting the data quality.

Interviewers with the highest conscientiousness levels introduce the most variance out of all four interviewer groups. This is rather surprising as one would expect those who are the most conscientious to perform the task in the most diligent way and introduce less variance than others. Such a result begs the question whether the self-reported conscientiousness measure can be considered valid. Another possibility is that due to their factual and salient nature, the NTS questions are not the most suitable to investigate interviewer effect by using personality traits. It is plausible that personal and sensitive questions may be more prone to the interviewer effect as the result of interviewers' personality traits like openness or empathy.

Those who were the most extrovert introduced, on average, the least variance into the survey estimates. Since Jäckle *et al.* (2011) found a positive correlation between extroversion levels and cooperation rates, then the hypothesis posed by Brunton-Smith *et al.* (2012) is not confirmed empirically. It is possible that those scoring highly on the extroversion scale not only achieved higher cooperation rates and decreased the nonresponse error but also decreased measurement error as they were responsible for a lower proportion of interviewer-related variance. This result should however be treated with caution as extroversion turned out to be statistically significant in only four models.

There were certain survey items that exhibited especially large differences between interviewer groups. Among the duration items these were 'tube station by bus' and 'tube station' and among satisfaction items it was 'cycle lanes provision' item. It is rather puzzling why those exhibited such a big differences regardless of the paradata variable used to partition the interviewer effect. None of them were subject to a particularly large or small interviewer effect. However, fewer cases were available for these items, as the tube-related questions were only asked in the Greater London area, while the question about cycle lanes was dropped in the later NTS administrations. That leads to one possible explanation: estimation problems. These items may require further investigation and possibly should be ignored when the overall patterns are considered.

The main message to be taken from the study presented in this paper is that evidence suggests that certain interviewer characteristics may affect the variability of survey estimates. Depending on their skills, attitudes and experience, interviewers introduce different amounts of variance into survey estimates. This is an important finding that can be

incorporated by survey research agencies into their interviewer recruitment and training procedures in an attempt to decrease the variance of survey estimates.

3. Survey satisficing: how interviewer personality traits, attitudes and satisficing behaviour influence respondents' propensity to satisfice

3.1 Introduction

Between the early 1980s and the late 1990s, the science of cognitive psychology became increasingly influential within survey methodology. A key insight from the introduction of a cognitive perspective was that, after a question has been administered, a respondent goes through a cognitive process in order to formulate a response (Sirken and Schechter, 1999). This process consists of four elements: comprehension of the question, retrieval of the information, judgment, and response (Tourangeau *et al.*, 2000). It is widely accepted that interviewers can facilitate formulating a response by motivating their respondents and thus obtain a better quality of data compared to other modes of data collection such as self-completion (Holbrook *et al.*, 2003).

In his seminal article, Krosnick (1991), drawing on the rational choice framework of decision making and behaviour, identified two different strategies that can be used by a respondent in the interview situation: optimising and satisficing. While optimising is defined as providing the best possible answer to a survey question given the available time and cognitive resources, satisficing is more frequently adopted by a respondent. Krosnick (1991) defined satisficing as replying to survey questions superficially, expending a degree of mental effort sufficient only to provide a reasonable or minimally acceptable answer. Satisficing therefore has a negative effect on survey data quality and is seen as one of the primary sources of survey measurement error (Green *et al.*, 2001; Jäckle *et al.*, 2010).

There are three main sources of satisficing identified in the existing literature: the respondent, the task/instrument and the interviewer. While the majority of research attention has focused on the first two, this paper focuses on interviewers as a source of satisficing response behaviour, which may arise through their encouraging (or at least tolerating) satisficing strategies by their respondents and thus negatively affect the survey data quality. Four hypotheses are proposed. First, that interviewers' attitudes towards interviewing can affect respondents' propensity to satisfice. Second, that interviewers' personality traits and skills can affect respondents' satisficing. Third, that interviewers' demographic characteristics can affect respondents' satisficing. And finally, that interviewers' own response strategies, namely satisficing, can affect their respondents' probability of

satisficing. The general expectation is that interviewers with particular characteristics, for example those who pay less attention to detail or employ satisficing strategies themselves, will increase the prevalence of satisficing strategies among the respondents they interview.

By testing these four hypotheses on a large, nationally representative survey, this paper seeks to address the main research question of which interviewer characteristics affect survey satisficing and, consequently, survey data quality. A data quality measure referred to as 'heaping', which is providing a numerical response that is a rounded rather than an exact number – is used as the key outcome. The findings from these analyses contribute to a better understanding of the interaction between respondents and interviewers and could be used to reduce interviewer-related measurement error in face-to-face surveys.

This paper consists of eight sections, of which this is the first. The second provides definitions of satisficing and heaping, including a review of the existing research in this area. The third section discusses the three sources of satisficing in surveys with the emphasis on the interviewer as the main subject of investigation. This is followed by a description of the data employed and an explanation of how measures of satisficing used in the analysis were developed. The sixth section describes the statistical model and analytical strategy used and is followed by the presentation of the results in section seven. The final section contains discussion of the main results.

3.2 Satisficing: theory and measurement

Krosnick (1991) uses satisficing as the umbrella term for various strategies that may be employed by respondents in order to save mental effort when answering questions. In the survey literature, satisficing is often used as a data quality indicator (Barge and Gehlbach, 2011) and the prevalence of satisficing behaviour among respondents is used to compare modes of data collection (Roberts, 2007). What Krosnick refers to as 'weak' forms of satisficing occur when the four cognitive stages of response formulation are executed but not in an accurate way (Krosnick *et al.*, 1996). This may involve 'acquiescence', which is when a respondent agrees with the statements provided by an interviewer regardless of their actual beliefs. It can be defined as the percentage of 'yes' and 'strongly agree/agree' replies over the total number of questions asked that provided this option. While the reasons why individuals have a tendency to agree with statements posed to them remain unclear, this acquiescence bias, or 'yeah-saying', is an important source of error in survey research as it can inflate the positive attitudes of respondents towards the issues raised in the survey (Knowles and Condon, 1999; Olson and Bilgen, 2011). Another example of weak satisficing is selecting the

first available, or 'good enough', response option from a list rather than reading to the end and selecting the most appropriate answer – a form of primacy effect (Krosnick and Alwin, 1987). In telephone interviews, the opposite to this phenomenon can occur when respondents choose the most recently heard option which is read out to them from a list – a recency effect (Holbrook *et al.*, 2003).

Strong satisficing, on the other hand, is when respondents omit the retrieval and judgment steps from the response process entirely, for example, choosing the 'Don't know' or 'No opinion' option when they hold a genuine attitude on an issue. Another example of strong satisficing is 'non-differentiation' which is when exactly the same point on a scale is used to rate a number of different items (Krosnick *et al.*, 1996). Here, respondents give no thought to the answer they are choosing but merely take the quickest and easiest route through the questionnaire. It is particularly common in web surveys that tend to take less time than other data collection modes of the same questionnaire (Chang and Krosnick, 2010; Fricker *et al.*, 2005; Heerwegh and Loosveldt, 2008). Kaminska *et al.* (2011) used a range of indicators for strong satisficing including the 'Don't know' indicator – if a respondent provided a 'Don't know' answer at least once; reversed straight-lining indicator – if the respondent provided the same answer to all questions in one block where at least one question was coded in reverse; and inconsistent responses – if inconsistency is observed in at least one out of four logical question pairs. A very similar approach was used by Lipps (2007) who operationalised satisficing as the number of 'Don't know' or 'No answer' responses among 22 satisfaction and attitude questions complemented with respondents' propensity to choose extreme answers from the 0-10 scale on the same items.

One implication of satisficing is that respondents try to complete the questionnaire as quickly as possible (Olson and Peytchev, 2007). This is more common in self-administered data collection modes (Lelkes *at al.*, 2012). Oppenheimer *et al.* (2009) used what they called Instructional Manipulation Checks (IMC). If someone failed the IMC – did not read the instructions and answered the question rather than skip the question by clicking on the title – they were considered to be those prone to satisficing. Those were found to take less time to complete the questionnaire.

Another example of a satisficing strategy may be 'heaping'. Heaping occurs when respondents are asked a question which requires a numerical estimate for the answer. Rather than expending the cognitive effort required to provide an accurate answer, a heaped response is rounded to the nearest multiple of a common integer such as 5, 10, or 100 (Burton and Blair,

1991; Wang and Heitjan, 2008). Heaping is a particular distortion of data collected in the survey, which is only visible at the aggregated level because spikes or 'heaps' are apparent at these common integer values when the individual responses are aggregated and plotted on a histogram (Pudney, 2008).

A common example of heaping is the number of cigarettes smoked per day, with heaps found at values 10 and 20 which are equivalent to small and large packets of cigarettes (Bar and Lillard, 2010; Wang and Heitjan, 2008). These reports could easily lead to overestimation of the number of cigarettes smoked if respondents actually smoke less than a full standard pack, e.g. eight to nine or 18 to 19 cigarettes. Heaping may also be caused by the phenomenon of Digit Preference (DP) as individuals seem to prefer, for reasons which are not entirely clear, numbers ending with 0 or 5. Heaping may be caused by rounding techniques used to estimate income (Moore *et al.,* 1999). It was also noticed that when reporting on the frequency or number of events, e.g. hospital visits, the longer the span of time between the actual event and reporting it, the more heaping is present in the data (Huttenlocher *et al.*, 1990). Similarly, the longer the distance reported for questions about distance, the more the tendency to heaping is observed, an effect which is apparent in the dataset investigated in this paper.

At the individual level, there is usually no certain way to determine which responses are genuine and which are the effect of heaping. For example, it is impossible to distinguish between respondents who smoke exactly 20 cigarettes a day and report this accurately, and others who smoke 18 cigarettes a day but, by employing a satisficing strategy, round their response to 20. Heaping usually becomes a data quality issue when approximately a dozen of the most frequent sample values account for 54%-69% of the total population (Pudney, 2008). Beaman *et al.* (2005) developed the concept of a 'neighbourhood' that relies on observed values that are 'neighbours' to the heaping value. If 10 is potentially a heaping value, as a multiple of 5, its 'neighbours' in the data are 8, 9, 11 and 12. One would expect that each of these values has an equal chance of being selected by a respondent as a legitimate answer, hence each of them should account for 20% of the sample. If the heaping value 10 is selected substantially more often (significantly more than 20% of times) then it is 'truly' a heaping value (Beaman *et al.*, 2005).

It is no surprise that the occurrence of heaps in the data can seriously affect survey estimates such as the mean number of cigarettes smoked daily, the mean number of visits to a GP in the past 12 months, the mean time taken to walk to the nearest grocery store, etc. There have been some efforts to devise special estimation methods that account for this type of error (Heitjan and Rubin, 1991; Pudney, 2008; Roberts and Brewer, 2001; Wang and Heitjan,

2008). The main concern of this paper, however, is with the causes of heaping rather than statistical remedies which can be applied after the error has already occurred. Heaping is a very common phenomenon in fields such as medicine, economics, sociology and agriculture, and applies to various aspects of human life: gestational age, breastfeeding duration, blood pressure, drug use, number of sexual partners, cigarettes smoked, GP/hospital visits, unemployment duration, household income and expenditure and even harvest estimates (Beaman *et al.*, 2005; Browning *et al.*, 2003; Denic *et al.*, 2004; Locker and Mason, 2006; Pickering, 1992; Pudney, 2008).

3.3 Sources of satisficing

So far, the definition of satisficing, and empirical means of measuring it, have been discussed, but the existing research in this area is also concerned with the origins of the phenomenon. Poor data quality can be caused by many factors such as badly designed questionnaires, poor standards of fieldwork, respondent fatigue or unwillingness (Blasius and Thiessen, 2012). These factors can be grouped into three major sources of measurement error: the respondent, the task and the interviewer (Sudman and Bradburn, 1974) and satisficing can originate from all three. Each of these is considered in turn below.

3.3.1 The respondent

The measurement error introduced by a respondent depends on many factors, such as interest in the survey topic and motivation to participate, ability to interpret the questions correctly, memory, ability to estimate the right answer or willingness to share true opinions about survey topic. A primary determinant of respondent-driven measurement error is the level of cognitive effort applied in answering questions (Groves *et al.*, 2004; Narayan and Krosnick, 1996). The emphasis on achieving a high response rate means that sometimes reluctant and less motivated respondents are persuaded to reply to a survey (Kaminska *et al.,* 2011). Due to low motivation, these respondents may employ a range of satisficing techniques. However, Kaminska *et al.* (2011) showed that the decrease in quality of answers is mainly attributable to the lower cognitive skills of 'satisficers'. Oppenheimer *et al.* (2009) compared 'satisficers', those who ignored the instructions in the questionnaire to skip the question, with 'non-satisficers' and found that the latter group provided more consistent and better quality responses across the questionnaire. In addition, 'satisficers' took significantly less time to complete the whole task which, according to the authors, suggested that they

were less motivated to do the task and rushed through it. This particular effect of satisficing is explored in more detail in Paper 3 (Chapter Four).

3.3.2 The task/instrument

A number of studies have focused on mode of data collection to investigate whether different modes yield different levels of satisficing. The comparisons of a Computer Assisted Telephone Interviews (CATI) and web-based surveys yielded contradictory results and revealed the complex nature of interaction between mode of data collection and satisficing. Some research noted less satisficing, expressed as non-differentiation, in web surveys than in CATI. It was discovered that respondents with lower cognitive skills, judged by their school tests, responded better to the visual features of a web survey (Chang and Krosnick, 2010). However, an earlier study by Fricker *et al.* (2005) showed that that web survey respondents differentiated less on a range of items comparing to CATI respondents, possibly because they were presented to them in a grid format. The broad effects of question presentation and questionnaire visualisation on data quality have also been investigated by others (Couper, 2009; Couper and Kreuter, 2010; Fuchs, 2009).

Web surveys may also attract more satisficing, higher prevalence of 'Don't know', non-differentiation and more item nonresponse than face-to-face interviews (Heerwegh and Loosveldt, 2008). Holbrook *et al.* (2003) compared telephone interviews with face-to-face interviews and discovered that CATI methods generated more satisficing measured as the prevalence of choosing the 'no opinion' response and non-differentiation on a battery of questions. The mode of data collection is closely related to respondents' behaviour, as self-administered instruments generate less motivation in respondents than methods using an interviewer. Low motivation can decrease the reliability of respondents' answers by decreasing accountability and encourage satisficing (Fricker *et al.*, 2005; Lelkes *et al.*, 2012). However, the anonymity of self-administered methods is thought to reduce errors which arise due to social desirability bias (Kreuter *et al.*, 2009).

There are also aspects of questionnaire design that may encourage satisficing behaviour; for example, the explicit availability of a 'No opinion' response option may encourage respondents to select it (Krosnick *et al.,* 2002). Ambiguity in question wording may also seriously impact survey data quality and encourage satisficing by requiring more cognitive effort to formulate an answer (Belson, 1981; Fowler, 1992). Increased task difficulty, such as detailed questions about various income sources, can trigger satisficing behaviour strategies (Krosnick *et al.*, 1996). Long questionnaires may result in fatigue for the respondent and decrease in motivation, leading to satisficing (Yan and Tourangeau, 2008).

3.3.3 The interviewer

The interviewer can introduce measurement error in a number of ways: through the level of assistance given to the respondent, through the way questions are read out and through the effect of interviewer characteristics on the interaction with the respondent (Groves *et al.*, 2009). There are certain types of interviewer behaviour that can affect the quality of the data such as adaptation of questions for the respondent, inattentiveness or hinting to clarify an appropriate response (Van der Zouwen and Dijkstra, 1988). Even the mere presence of the interviewer may trigger a social desirability effect, whereby the respondent tries to appear in the best light given the characteristics of the interviewer (DeMaio, 1984). Fowler and Mangione (1990) contend that some of the measurement error introduced by the interviewer can be avoided by standardisation of the interviewer's work.

Although, generally, the interviewer effect is a well researched topic (Catania *et al.*, 1996; Olson and Bilgen, 2011; Olson and Peytchev, 2007; Reinecke and Schmidt, 1993; Smit *et al.*, 1997), the interviewer as the source of satisficing is less well researched compared to the respondent and the instrument. Moreover, studies on interviewer effect are mostly concerned with the extent to which variance of survey estimates can be attributed to interviewers and then explained by their characteristics such as prior expectations of the response (Clarke *et al.*, 2003; Singer and Kohnke-Aguirre, 1977; Sudman *et al.*, 1977); experience (Olson and Peytchev, 2007); and demographic characteristics (Pickery *et al.*, 2001; Singer *et al.*, 1983). While these studies make use of rich information on interviewers, they are not specifically concerned with satisficing. Hypotheses posed in this paper employ various interviewers' characteristics and use them specifically to better understand this particular cause for measurement error.

Interviewers' attitudes towards interviewing have been used in research on nonresponse and cooperation rates achieved by interviewers (Jäckle *et al.,* 2011). It has been found that interviewers who believe that the refusal should not be accepted were more successful in obtaining an interview. In the context of measurement error Lipps (2007) used a multilevel model to assess the effect of a range of interviewer and respondent variables on data quality. He operationalised data quality using a range of measures, including a social desirability index and nonresponse to an income question as well as indicators of satisficing such as the prevalence of 'Don't know' or 'No opinion' answers, extreme category responses and non-differentiation. Among many interviewers' characteristics, he also included a measure of interviewers' willingness to disclose their own income and their willingness to participate in a similar surveys in the future. Both variables turned out to be insignificant in explaining the

interviewers' effect on data quality indicators. However, the results presented in Paper One (Chapter Two) in this thesis have shown that the interviewer's determination to secure an interview plays a significant role in explaining some of the interviewer-related variance introduced into the survey estimates. Hence, the first hypothesis posed and tested in this paper is that interviewers' attitudes towards interviewing, such as accepting a refusal or respecting the privacy of a respondent, are correlated with their respondents' propensity to satisfice.

Dijkstra (1987) found that if a more 'personal' or conversational style of interviewing is employed by the interviewer then the respondent is more motivated and makes more effort to answer the questions than if a more 'formal' style is used. However, those interviewed using a personal style were more prone to providing socially desirable responses and could be more prone to satisficing as well. This more personal style of interviewing may stem from interviewers' particular personality traits such as openness, extraversion or flexibility. For that reason, the second hypothesis posed in this paper is that interviewers' personality traits and skills are correlated with their respondents' propensity to satisfice.

The research presented in Chapter Two (Paper One) showed that the degree of variability introduced into survey estimates by interviewers depends on their characteristics. For example, interviewers with more specific survey experience introduce less variability than interviewers with very little experience with the same survey. In the Lipps' study (2007), basic interviewer characteristics such as gender, mother tongue, education status and age were used to predict the data quality indicators. Lipps found that small proportions of the total variance in the data quality indicators were attributable to interviewers – 5% of the total variance in extreme category usage and 10% for income nonresponse. However, of the interviewer characteristics he examined, only interviewers' experience was statistically significant in explaining this variability. Those who conducted more interviews had a higher rate of income nonresponse (Lipps, 2007). Similar results were observed in a study that investigated the role of interviewers on satisficing and focused on acquiescence rates and how these vary between very inexperienced interviewers and interviewers with survey-specific experience (Olson and Bilgen, 2011). Respondents' education turned out to be a significant predictor of their acquiescence propensity, but interviewers' survey-specific experience explained more acquiescence variance than respondents' education. Results from this research inform the third hypothesis that interviewers' demographic characteristics are correlated with their respondents' propensity to satisfice.

The fact that interviewers can affect a respondent's motivation suggests by implication that interviewers can affect the satisficing behaviour of respondents. The majority of research on satisficing focuses either on respondents or on the interview task where both questionnaire and mode of data collection are considered. Studies that consider the interviewer's role in satisficing use only basic information on interviewers and none of them use measures of interviewers' own satisficing behaviour (Lipps, 2007; Olson and Bilgen, 2011; Olson and Peytchev, 2007). Some studies on suggestive interviewing style show that it can skew the survey results by affecting the response in the chosen direction, even when this is not intended (Smit *et al.*, 1997; Clarke *et al.*, 2003) but they do not address the question of how the interviewers' own response styles might play a role in this process. The final hypothesis posed in this study is that interviewers' own propensity to satisfice is positively correlated with their respondents' propensity to satisfice.

3.4 Data

The data used in this study came from three different sources: the National Travel Survey (NTS) data collected between 2002 and 2008, geographical area characteristics at the Middle Layer Super Output Area (MSOA) level and administrative information on interviewers supplemented by more detailed information coming from the 'Who is behind the NatCen ID card?' survey. These sources are described in detail in Chapter 1 in Section 1.5. The main dataset used in this study contains 29,569 respondents from England, who are nested within 3,594 MSOA areas and were interviewed by 400 interviewers.

3.5 Measures

The first hypothesis states that interviewers' attitudes towards interviewing are correlated with their respondents' propensity to satisfice. The interviewers' attitudes towards interviewing were measured using the 'Persuading respondents' section from the 'Who is behind the NatCen ID card?' interviewer survey. This section consisted of seven four-point Likert scale questions (1 – Strongly Agree; 4 – Strongly Disagree) as follows:

- Reluctant respondents should always be persuaded to participate.
- With enough effort, even the most reluctant respondent can be persuaded to participate.
- An interviewer should respect the privacy of the respondent.
- If a respondent is reluctant, a refusal should be accepted.
- One should always emphasize the voluntary nature of participation.
- It does not make sense to contact reluctant target persons repeatedly.
- If you catch them at the right time, most people will agree to participate.
- Respondents persuaded after great effort do not provide reliable answers.

These items do not form a reliable scale or scales as judged by their Cronbach's Alpha coefficient(s), hence they were treated as separate continuous measures of interviewers' attitudes towards interviewing (Jäckle *et al.*, 2011).

The second hypothesis is that interviewer personality traits and skills are correlated with their respondents' propensity to satisfice. The interviewer survey covered a range of personality traits and specific skills: agreeableness, conscientiousness, extraversion, neuroticism, openness, ability to pick up cues, empathy, reading other people, time management, problem solving/reasoning, persuasion, nonverbal communication and many others. Dimension reduction methods were employed to compute personality traits measures. Principal Component Analysis (PCA) with Direct Oblimin rotation was used, as personality traits are rarely independent. Factor scores were saved using the Anderson-Rubin method and each scale had a Cronbach's Alpha coefficient larger than 0.7. The interviewers' personality traits and skills used to test the second hypothesis were: openness, neuroticism, conscientiousness, extraversion, empathy, efficiency, flexibility, attention to detail and communication skills including persuasion (see also Appendix 2).

The third hypothesis is that interviewers' demographic characteristics are correlated with their respondents' propensity to satisfice. This information comes from the administrative file with interviewers' personal details and includes age and gender. Their NTS specific

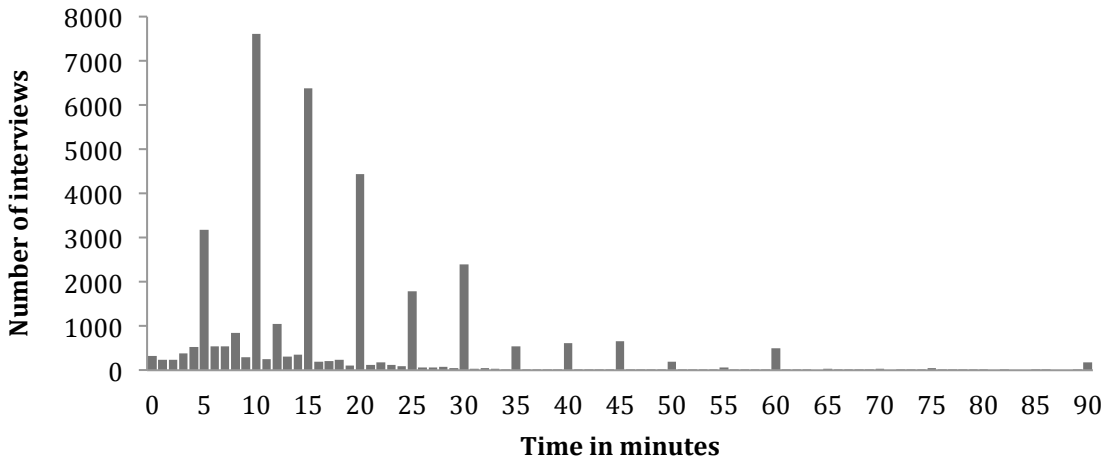
experience expressed as the number of NTS administrations they worked on was also included in this category.

The fourth hypothesis is that interviewers' own propensity to satisfice is positively correlated with their respondents' propensity to satisfice. Satisficing behaviour in this paper is operationalised as heaping, for both interviewers and respondents . The NTS questionnaire contains an Administration Block section that is completed by the interviewer for every household in their workload. There are two items in that block that can be used as a measure of heaping for interviewers:

- How long did it take to place and explain the diary and other documents? Include any time spent preparing the diaries beforehand. Record to the nearest minute.
- How long did it take to pick up and check the diary(ies)? Record to the nearest minute.

Figure 3.1 below shows the histogram for the second item – time to pick up the diaries. The multiples of five minutes are the most common answers by some margin.

Figure 3.1. Time taken to pick up and check the diaries (in minutes)



The interviewers' heaping indicator was calculated as follows. If the interviewer's answer was a multiple of 5 then they were given a score of 1 and a score if 0 otherwise. The heaping indicator was computed as the mean over the two items resulting in three possible scores: 0 (no heaping), 0.5 (heaping on one item) and 1 (heaping on both items). As each interviewer conducted a number of interviews, the mean value of the heaping indicator was taken over all interviews conducted by each interviewer. It is a continuous variable that ranges from 0.16 to 1, with a mean of 0.831.

There are six variables in the NTS household questionnaire which ask about the time it takes to get to the nearest transport links and local amenities. On all six items the unit of measurement is minutes and the survey asks how long would it take to get to:

1. Transport
    - the nearest bus stop
    - the nearest railway station
    - the nearest railway station using bus
2. Amenities
    - the nearest grocery store
    - the nearest shopping centre
    - the nearest GP.

All six items exhibit a heaped distribution. Two examples are provided in Figures 3.2 and 3.3. There is a distinct difference between the two histograms. For the grocery store item (Figure 3.2) the responses are mostly concentrated at the lower end of the scale while for the shopping centre item (Figure 3.3) the responses are more scattered. Heaps are also more apparent in Figure 3.3. One reason for this may be that it is likely that almost everybody will live close to a store selling groceries but not all respondents will live close to a shopping centre. Indeed, the mean time taken to get to the nearest grocery store reported by respondents was eight minutes, while the mean time taken to get to the shopping centre was almost 20 minutes. This may be taken as evidence that heaping depends on the distance and the greater the distance to transport links and amenities, the more likely respondents are to round their answer. In addition to the visible heaps at every multiple of five minutes there are also the hourly heaps corresponding to expressions of time not in minutes but in hours. These are especially visible in Figure 3.3 at ‘half an hour’ and ‘an hour’.

Figure 3.2. Distribution of the responses to the 'How long would it take to get to the nearest grocery store' item

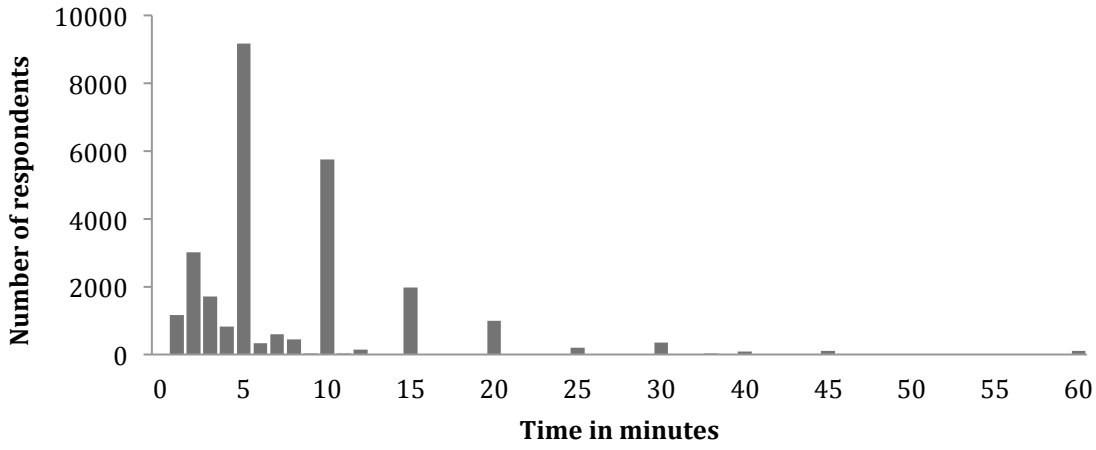
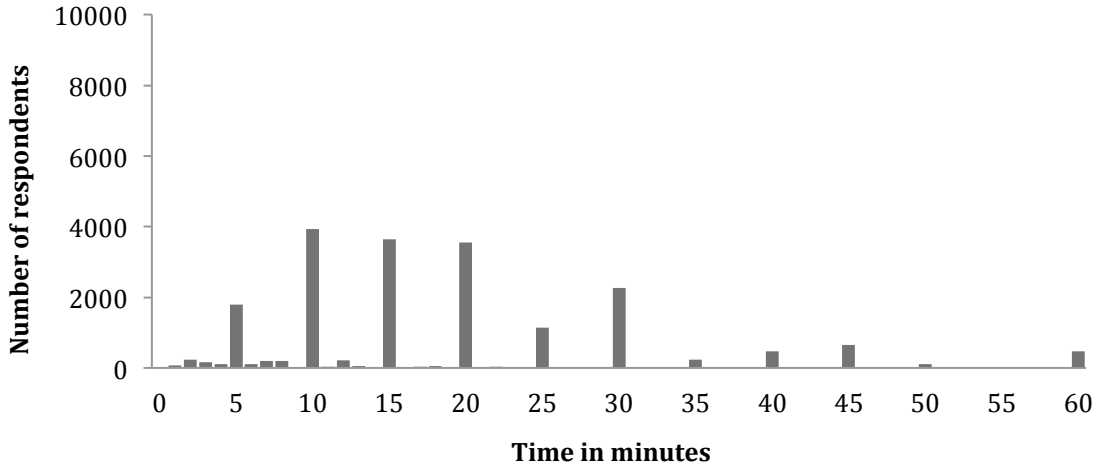


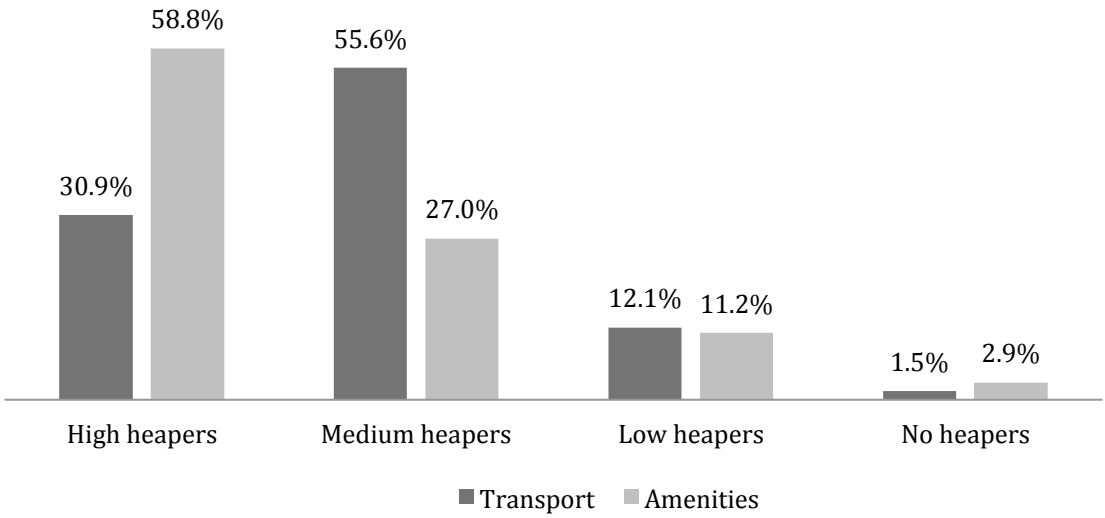
Figure 3.3. Distribution of the responses to the 'How long would it take to get to the nearest hospital' item



Because there seems to be a positive correlation between the frequency distribution of the transport links/amenities variables and the extent of heaping, 'remoteness' indicators were computed to be included as controls in the predictive models. For each respondent, the mean of their distances, expressed as minutes, was taken over three transport and three amenities variables. The mean times taken to the transport links and amenities are 20:15 minutes and 13:53 minutes respectively. The 'remoteness' indicator was used as a control variable to ensure that distance does not affect interviewer assignment – respondents who live in remote areas could potentially give more rounded answers than those living in urban areas.

The NTS questionnaire is designed in such a way that certain blocks of questions are asked only of a random subsample of respondents. Around 50% of respondents answered both the transport and the amenities questions. The amenities block was not introduced until 2005. In order to maximise the number of cases for the analyses, two separate dependent variables were derived, one for the transport items and one for the amenities items. The respondent's heaping indicators were computed in the same way as for interviewers. If the respondent's answer was a multiple of five then they were given a score of 1 and score of 0 otherwise. As the mean was taken over the three items they could then score: 0 (if no heaping at all), 0.33 (if heaping on one item), 0.66 (if heaping on two items) and 1 (if heaping on all three items). The respondents were then categorised into the following mutually exclusive groups: no heapers, low heapers, medium heapers and high heapers. These values were calculated only for respondents who gave an eligible answer. The small number of respondents who did not provide a reply (<1%) were excluded from the analysis. Figure 3.4 shows the distribution of both dependent variables. There are clear differences between the two dependent variables as the respondents answering the amenities questions were much more likely to be in the high heaping group than respondents answering the transport variables. It is also apparent that heaping in general is a preferred response strategy among the respondents as only a very small proportion of them (1.5% for transport and 2.9% for amenities) did not heap at all.

Figure 3.4. Distribution of the dependent variables: transport and amenities



3.6 Model specification

A cross-classified multilevel ordered logit model was used to test the hypotheses. Both dependent variables are discrete ordinal variables that measure a respondent's propensity to heap and have four categories: 3 = High heaping, 2 = Medium heaping, 1 = Low heaping and 0 = No heaping, the last being the reference category. The main outcome of interest is the probability of a respondent being in the high heaping group. The multilevel approach is generally used for this type of research (Hox, 1994; Lipps, 2007; O'Muircheartaigh and Campanelli, 1998; Schnell and Kreuter, 2005) because it is suitable when the data have a hierarchical structure. The multilevel approach allows estimation of the variability associated with each level in the data (Snijders and Bosker, 2004). The cross-classified model is appropriate when the lower level units are nested within more than one classification at the same, higher level.

In the NTS, respondents are not assigned to interviewers at random. As a consequence it is plausible that the interviewer effect may be due different sample composition, rather than a genuine interviewer influence on the respondent by the interviewer. The modelling approach used here allows the usually confounded interviewer and area effects to be disentangled while controlling for area and respondent characteristics at the same time, to account for differences in sample composition between interviewer assignments.

All models were estimated in MLwiN version 2.25 (Rasbash *et al.,* 2009) using the Monte Carlo Markov Chain (MCMC) estimators (Browne, 2009). The Deviance Information Criterion (DIC) was used to assess the improvement in the model fit between the two models, a decrease in DIC suggesting a better fit (Spiegelhalter *et al.,* 2002). To aid understanding of the interviewer effect the variance partition coefficients (VPCs) have been approximated using the linear threshold model method which assumes that the Level 1 variance is approximately equal to 3.29; this method is appropriate for this logistic model (Snijders and Bosker, 2004).

The ordered logit model is an extension of the binary logit model and contains K–1 predictions where K is the number of categories in the dependent variable. In this case the model contains three predictions. Let c denote the heaping category such that $c = \{0,1,2,3\}$ with $c = 3$ being the high heaping category. Then:

$$Pr(Y_i = 1|x_i) = \pi_{1i}$$
$$Pr(Y_i = 2|x_i) = \pi_{2i} \quad (3.1)$$
$$Pr(Y_i = 3|x_i) = \pi_{3i}$$

where Y_i is the heaping category, π_{ci} represents the probability of a respondent being in the relevant heaping category, e.g. π_{3i} is the probability of a respondent to be in the high heaping group, and x_i is the vector of covariates, e.g. respondent and interviewer characteristics for respondent *i*. Using the logit link function the model has the form:

$$logit[\pi_{1i}] = \ln\left(\frac{\pi_{1i}}{1-\pi_{1i}}\right) = \alpha_1 + \beta X_i$$
$$logit[\pi_{2i}] = \ln\left(\frac{\pi_{2i}}{1-\pi_{2i}}\right) = \alpha_2 + \beta X_i \quad (3.2)$$
$$logit[\pi_{3i}] = \ln\left(\frac{\pi_{3i}}{1-\pi_{3i}}\right) = \alpha_3 + \beta X_i$$

where α_c are the intercepts (sometimes called the cut-points or thresholds) at which respondents 'switch' between the groups, and X_i is the vector of independent variables with associated vector of coefficients β; i denotes the *i*th respondent. Solving one of the Equations 3.2 for π_{3i} gives the predicted probability of being in the high heaping group as:

$$\pi_{3i} = \frac{\exp(\alpha_3 + \beta X_i)}{1+\exp(\alpha_3+\beta X_i)} \quad (3.3)$$

Equation 3.3 refers to a one-level ordered logit model and additional subscripts need to be added to account for the cross-classified two-level structure of the data. Each respondent at Level 1 (denoted i) is nested in the cross-classification of the MSOA area (denoted j) and interviewer (denoted k) at Level 2 such that:

$\pi_{3i(j,k)}$ – probability of being in the high heaping group

$\pi_{2i(j,k)}$ – probability of being in the medium heaping group

$\pi_{1i(j,k)}$– probability of being in the low heaping group

The cumulative probabilities are defined as:

$$\gamma_{3i(j,k)} = \pi_{3i(j,k)}$$

$$\gamma_{2i(j,k)} = \pi_{3i(j,k)} + \pi_{2i(j,k)}$$

$$\gamma_{1i(\mathrm{j,k})} = \pi_{3i(j,k)} + \pi_{2i(j,k)} + \pi_{1i(j,k)}$$

$$\gamma_{0i(j,k)} = 1$$

where $\gamma_{2i(j,k)}$ is the probability of being in the medium or high heaping groups and $\gamma_{1i(j,k)}$ is the probability of being in the low, medium or high heaping groups. The no heaping group (0) is the reference category.

The final set of equations is then:

$$Logit\left(\gamma_{3i(j,k)}\right) = \alpha_3 + h_{i(j,k)}$$

$$Logit\left(\gamma_{2i(j,k)}\right) = \alpha_2 + h_{i(j,k)} \qquad (3.4)$$

$$Logit\left(\gamma_{1i(j,k)}\right) = \alpha_1 + h_{i(j,k)}$$

where α_c correspond to Equation 3.2 and $h_{i(j,k)}$ is common to all three equations. The $h_{i(j,k)}$ contains the vector of explanatory variables X_{1i}, X_{2j} and $X_{3k}$with associated coefficients at respondent (i), area (j) and interviewer (k) level, where β_c corresponds to β in Equation 3.2, and two error terms $v_{2j}$and v_{3k}.

$$h_{i(j,k)} = \beta_1 X_{1i} + \beta_2 X_{2j} + \beta_3 X_{3k} + v_{2j} + v_{3k} \quad (3.5)$$

These error terms are assumed to be independent and identically distributed (IID) with zero means and variances σ_j^2 and σ_k^2.

3.7 Results

The 'null' models are presented first as they show whether there is a statistically significant interviewer effect on the respondent's propensity to heap (Table 3.1). Results are presented separately for the transport and amenities variables. The statistical models are presented in the following order: the first model tests hypothesis one – the effect of interviewer attitudes towards interviewing; the second model tests hypothesis two– the effect of interviewer personality traits and skills; the third model test hypothesis three – the effect of interviewer demographic characteristics; the fourth model tests hypothesis four – that interviewers' own response style affects their respondents' propensity to satisfice. The final model includes a range of controls at the respondent and MSOA levels.

Table 3.1. Null models with complex variance component for 'transport' and 'amenities' variables

	Model 1 Transport	**Model 2 Amenities**
Thresholds		
α_3	-0.93 (0.036)	0.319 (0.046)
α_2	1.93 (0.039)	1.874 (0.048)
α_1	4.351 (0.069)	3.637 (0.063)
Covariates	No	No
Random effects		
Interviewer (σ_k^2)	0.198 (0.027)	0.341 (0.043)
Area (σ_j^2)	0.286 (0.041)	0.237 (0.047)

As can be seen in Table 3.1, both interviewer and area-related variance components are statistically significant. The interviewer variance accounts for approximately 5.2% of the total variance[17] for the transport model and 8.8% of the total variance for the amenities model. The difference in the magnitude of the interviewer effect between the two models may be due to the different nature of the questions. The thresholds in Table 3.1 correspond to the frequencies presented in Figure 3.4.

3.7.1 Transport

The models for the transport variable are presented in Table 3.2. Model 3 shows that of all the interviewers' attitudes towards interviewing, only one is statistically significant. Interviewers who disagree with the statement 'If a respondent is reluctant a refusal should be accepted' have a higher probability of satisficing among their respondents. This result provides qualified support for the first hypothesis, that interviewers' attitudes towards interviewing affect satisficing behaviour among their respondents.

[17]The Variance Partition Coefficient (VPC) is calculated using the approximation of the residual variance ≈ 3.29 then in Model 1 $VPC_{interviewer} = \frac{0.198}{0.198+0.286+3.29} = 0.052$ (Snijders and Bosker, 2004).

Table 3.2a. Transport models 3-5

	Model 3	Model 4	Model 5
Thresholds			
α_3	-1.299 (0.275)	-1.162 (0.319)	-1.311 (0.308)
α_2	1.561 (0.273)	1.701 (0.32)	1.553 (0.308)
α_1	3.984 (0.279)	4.125 (0.325)	3.977 (0.313)
Fixed effects			
Interviewers' attitudes[18]			
always to persuade	-0.003 (0.055)	-0.01 (0.056)	-0.018 (0.055)
everyone can be persuaded	0.033 (0.045)	0.019 (0.052)	0.046 (0.054)
respecting privacy	0.04 (0.063)	0.035 (0.067)	0.038 (0.068)
refusal accepted	0.153 (0.053)	0.154 (0.056)	0.15 (0.056)
voluntary nature of survey	-0.036 (0.046)	-0.029 (0.042)	-0.033 (0.044)
no point contacting	-0.056 (0.048)	-0.067 (0.046)	-0.048 (0.05)
catch in the right moment	0.018 (0.047)	0.014 (0.055)	0.02 (0.051)
persuaded not reliable	-0.006 (0.051)	-0.025 (0.051)	-0.021 (0.053)
Interviewers' personality			
openness		-0.047 (0.045)	-0.049 (0.046)
neuroticism		-0.039 (0.038)	-0.048 (0.039)
conscientiousness		-0.013 (0.041)	-0.009 (0.043)
extraversion		0.012 (0.041)	0.012 (0.042)
empathy		-0.044 (0.045)	-0.044 (0.047)
efficiency		0.042 (0.048)	0.033 (0.049)
attention to detail[19]		0.019 (0.04)	0.018 (0.038)
flexibility		0.089 (0.045)	0.097 (0.046)
communication skills		-0.026 (0.052)	-0.021 (0.053)
Interviewers' characteristics			
interviewer's age			0.006 (0.004)
interviewer's gender (female)			-0.119 (0.07)
NTS experience			0.023 (0.017)
Respondent and area covariates	No	No	No
Random effects			
Interviewer (σ_k^2)	0.198 (0.027)	0.201 (0.028)	0.197 (0.028)
Area (σ_j^2)	0.285 (0.040)	0.286 (0.040)	0.287 (0.041)

Note: Estimates and standard errors are the means and standard deviations of 50,000 chains, with burn-in of 500 and orthogonal parameterisation. N = 19,035 respondents nested in cross-classification of 3,113 areas and 358 interviewers.

[18] Treated as continuous variables where 1 – Strongly agree, 4 – Strongly disagree

[19] The factor 'attention to detail' is coded in reverse, the higher the score the poorer the attention to detail.

None of the interviewers' personality trait variables are statistically significant in Model 4. When the interviewers' characteristics are added in Model 5, interviewers' flexibility and interviewers' gender are significant predictors of respondent heaping. Interviewers who are more flexible have a higher rate of satisficing among their respondents, while female interviewers encouraged less satisficing compared to males. However, these two effects disappear when interviewers' heaping score is added to the model (Model 6). These results do not support the second hypotheses that interviewers' personality traits and skills affect respondent satisficing, nor the third hypothesis on the effect of interviewer demographic characteristics.

The interviewer heaping score is statistically significant in Model 6. Interviewers who are prone to satisficing themselves also obtain higher rates of satisficing among their respondents. This effect remains statistically significant after controlling for respondent and area-level characteristics in the final model (Model 7). Moreover, adding the interviewer satisficing indicator to the model significantly reduces the unexplained interviewer variance by 7%. This result supports the fourth hypothesis that interviewers' own satisficing behaviour is positively correlated with their respondents' propensity to satisfice.

Table 3.2b. Transport models 6 and 7

	Model 6	Model 7
Thresholds		
α_3	-2.026 (0.272)	-3.255 (0.367)
α_2	0.839 (0.272)	-0.192 (0.365)
α_1	3.262 (0.279)	2.308 (0.37)
Fixed effects		
Interviewers' attitudes[20]		
always to persuade	-0.005 (0.048)	-0.006 (0.053)
everyone can be persuaded	0.05 (0.052)	0.045 (0.052)
respecting privacy	0.059 (0.063)	0.006 (0.066)
refusal accepted	0.142 (0.053)	0.123 (0.055)
voluntary nature of survey	-0.038 (0.046)	-0.044 (0.043)
no point contacting repeatedly	-0.024 (0.047)	-0.012 (0.048)
catch in the right moment	0.017 (0.047)	0.02 (0.049)
persuaded not reliable	-0.029 (0.054)	-0.032 (0.051)
Interviewers' personality		
openness	-0.043 (0.044)	-0.011 (0.044)
neuroticism	-0.056 (0.039)	-0.043 (0.038)
conscientiousness	0.002 (0.041)	-0.02 (0.041)
extraversion	0.018 (0.04)	-0.009 (0.04)
empathy	-0.038 (0.045)	-0.046 (0.045)
efficiency	0.021 (0.047)	0.009 (0.047)
attention to detail[21]	0.018 (0.039)	0.001 (0.038)
flexibility	0.085 (0.045)	0.049 (0.046)
communication skills	-0.008 (0.05)	0.004 (0.051)
Interviewers' characteristics		
interviewer's age	0.006 (0.004)	0.001 (0.004)
interviewer's gender (female)	-0.087 (0.067)	-0.095 (0.071)
NTS experience	0.026 (0.016)	0.017 (0.017)
Interviewers' own response style		
Interviewer's heaping score	0.754 (0.208)	0.701 (0.213)
Respondent and area covariates	No	Yes
Random effects		
Interviewer (σ_k^2)	0.183 (0.026)	0.169 (0.025)
Area (σ_j^2)	0.288 (0.038)	0.326 (0.044)

[20] Treated as continuous variables where 1 – Strongly agree, 4 – Strongly disagree

[21] The factor 'attention to detail' is coded in reverse, the higher the score the poorer the attention to detail.

3.7.2 Amenities

The models for the amenities variable are presented in Table 3.3. In line with the results for the transport variable, Model 8 shows that interviewers who disagree with the statement 'If a respondent is reluctant a refusal should be accepted' obtain higher rates of satisficing among their respondents. Nevertheless, this effect is mitigated by the addition of the interviewer characteristics in Model 10. This result does not support the first hypothesis that interviewers' attitudes towards interviewing affect the satisficing behaviour oft heir respondents.

Model 9 tests the second hypothesis. None of the interviewer personality traits and skills are statistically significant predictors of respondent heaping. Although in Model 10, the interviewer's gender is statistically significant, females encouraging less satisficing than males, this effect disappears when the interviewer heaping indicator is added in Model 11. This result does not support the second nor the third hypothesis. Interviewers' demographic characteristics do not affect their respondents' propensity to satisfice. As with the transport variables, the interviewer's heaping indicator is statistically significant in Model 11. Interviewers who satisfice in their own questionnaire responses obtain more satisficing from their respondents. This effect remains statistically significant in the final model (Model 12) that includes controls for the respondent and area characteristics. The interviewer heaping indicator in Model 11 reduces the unexplained interviewer variance by 10% comparing to Model 10. The fourth hypothesis is supported by this result.

Table 3.3a. Amenities models 8-10

	Model 8	Model 9	Model 10
Thresholds			
α_3	0.361 (0.254)	0.712 (0.187)	0.657 (0.438)
α_2	1.917 (0.254)	2.269 (0.189)	2.215 (0.438)
α_1	3.679 (0.257)	4.033 (0.194)	3.978 (0.44)
Fixed effects			
Interviewers' attitudes			
always to persuade	0.02 (0.066)	0.014 (0.053)	-0.016 (0.074)
everyone can be persuaded	-0.069 (0.042)	-0.093 (0.059)	-0.075 (0.071)
respecting privacy	-0.032 (0.079)	-0.055 (0.096)	-0.054 (0.092)
refusal accepted	0.14 (0.053)	0.144 (0.063)	0.117 (0.071)
voluntary nature of survey	0.026 (0.056)	0.001 (0.063)	0.021 (0.058)
no point contacting repeatedly	-0.037 (0.062)	-0.101 (0.064)	-0.042 (0.067)
catch in the right moment	-0.041 (0.039)	-0.07 (0.058)	-0.027 (0.069)
persuaded not reliable	-0.031 (0.042)	-0.024 (0.083)	-0.042 (0.07)
Interviewers' personality			
openness		-0.076 (0.056)	-0.066 (0.06)
neuroticism		-0.01 (0.06)	-0.031 (0.051)
conscientiousness		0.015 (0.069)	0.005 (0.057)
extraversion		-0.04 (0.054)	-0.065 (0.054)
empathy		-0.086 (0.064)	-0.096 (0.062)
efficiency		0.063 (0.063)	0.051 (0.065)
attention to detail		0.009 (0.058)	0.009 (0.054)
flexibility		0.02 (0.067)	0.027 (0.061)
communication skills		0.099 (0.071)	0.102 (0.07)
Interviewers' characteristics			
interviewer's age			0.008 (0.006)
interviewer's gender (female)			-0.179 (0.085)
NTS experience			0.007 (0.019)
Respondent and area covariates	No	No	No
Random effects			
Interviewer (σ_k^2)	0.335 (0.044)	0.355 (0.046)	0.335 (0.046)
Area (σ_j^2)	0.229 (0.043)	0.239 (0.041)	0.237 (0.041)

Note: Estimates and standard errors are the means and standard deviations of 50,000 chains, with burn-in of 500 and orthogonal parameterisation. N = 18,173 respondents nested in cross-classification of 2,410 areas and 278 interviewers.

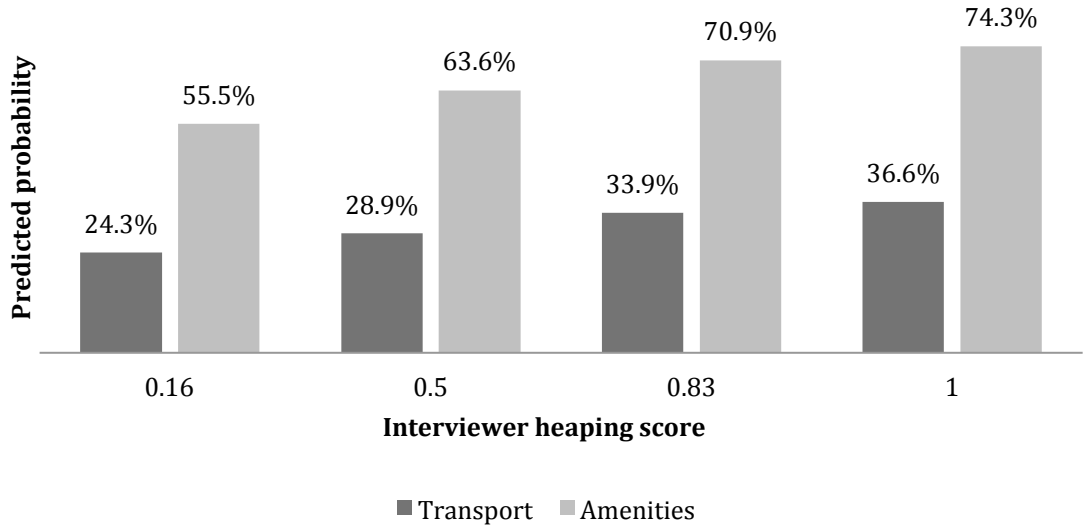
Table 3.3b. Amenities models 11 and 12

	Model 11	Model 12
Thresholds		
α_3	-0.465 (0.181)	-1.361 (0.438)
α_2	1.091 (0.181)	0.344 (0.438)
α_1	2.855 (0.183)	2.196 (0.439)
Fixed effects		
Interviewers' attitudes		
always to persuade	0.055 (0.069)	-0.031 (0.065)
everyone can be persuaded	-0.123 (0.06)	-0.086 (0.062)
respecting privacy	-0.034 (0.077)	-0.059 (0.08)
refusal accepted	0.114 (0.049)	0.111 (0.065)
voluntary nature of survey	0.023 (0.063)	0.002 (0.051)
no point contacting repeatedly	0.003 (0.057)	0.019 (0.058)
catch in the right moment	-0.011 (0.057)	-0.024 (0.059)
persuaded not reliable	-0.076 (0.051)	-0.105 (0.061)
Interviewers' personality		
openness	-0.072 (0.054)	-0.061 (0.053)
neuroticism	-0.036 (0.048)	-0.049 (0.044)
conscientiousness	0.015 (0.06)	-0.014 (0.051)
extraversion	-0.03 (0.054)	-0.031 (0.049)
empathy	-0.083 (0.065)	-0.075 (0.055)
efficiency	0.034 (0.063)	-0.002 (0.058)
attention to detail	0.009 (0.054)	-0.009 (0.047)
flexibility	0.003 (0.052)	-0.004 (0.054)
communication skills	0.13 (0.06)	0.077 (0.061)
Interviewers' characteristics		
interviewer's age	0.007 (0.006)	0.002 (0.005)
interviewer's gender (female)	-0.242 (0.096)	-0.115 (0.083)
NTS experience	0.012 (0.026)	0.018 (0.023)
Interviewers' own response style		
interviewer's heaping score	1.246 (0.2)	1 (0.255)
	No	Yes
Respondent and area covariates		
Random effects		
Interviewer (σ_k^2)	0.303 (0.044)	0.232 (0.034)
Area (σ_j^2)	0.22 (0.04)	0.23 (0.045)

The results from the ordered logit model can be better interpreted by calculating predicted probabilities for the outcome of interest. In this case the outcome of interest is the probability for a respondent to be in the highest heaping group. Separate probabilities were calculated

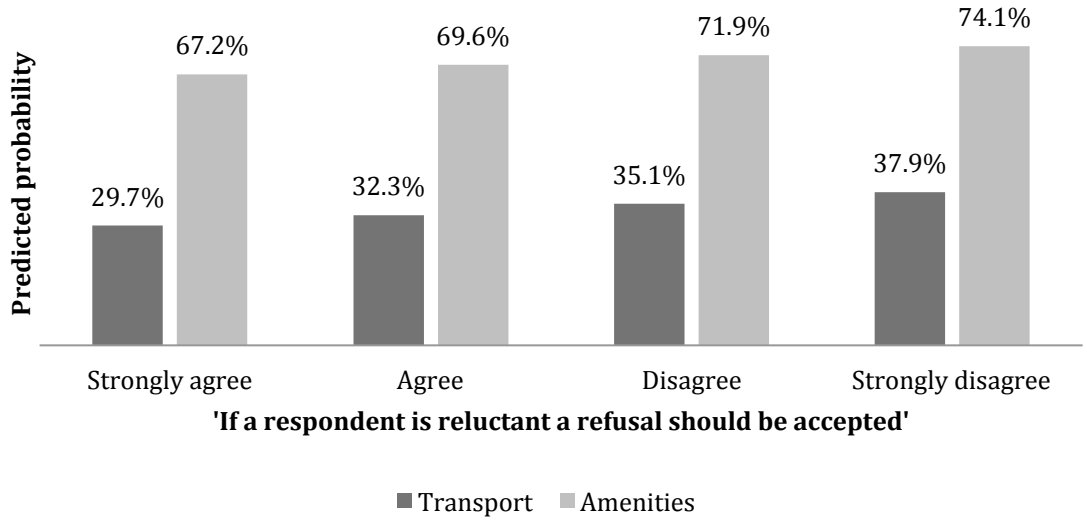
for four different values of the interviewer heaping variable and for different levels of agreement with the 'refusal should be accepted' item. All the remaining variables in the model were set to their sample mean values for continuous variables and modal values for categorical variables.

Figure 3.5. Predicted probabilities of a respondent being in the highest heaping group, for four different interviewer's heaping values (for female interviewer interviewing male respondent, who is employed full-time, white and married)



As shown in Figure 3.5, the probability of a respondent being in the high heaping group increases monotonically as interviewers' heaping score increases for both the amenities and the transport variables. Since respondent heaping is one of the indicators of lower data quality, it is clear that satisficing interviewers can affect the data quality by influencing satisficing behaviour among their respondents.

Figure 3.6. Predicted probabilities of a respondent being in the highest heaping group, for different levels of interviewer's agreement with the 'refusal should be accepted' item (for female interviewer interviewing male respondent, who is employed full-time, white and married)



The interviewers who disagree with the statement that if a respondent is reluctant a refusal should be accepted can be labelled as more determined to secure the interview. As Figure 3.6 shows respondents interviewed by these interviewers have higher probability of being in a high heaping group than those interviewed by less determined interviewers.

3.8 Discussion

Although is it widely accepted that face-to-face interviews produce high quality survey data, specifically due to the role of the interviewer and respondent-interviewer interaction, the interviewer is also a source of measurement error. One out of many different ways in which interviewers can introduce measurement error is by encouraging (or failing to discourage) their respondents to satisfice, i.e. provide a lower quality response by not investing enough cognitive effort into formulating the response. In the context of satisficing, the bulk of existing research is concerned with the respondent and task/instrument as a source but less with the interviewer. Studies of satisficing that consider the interviewer do not focus on interviewers' own response styles and their propensity to satisfice and in most cases use only basic information on interviewers.

The dataset used in this paper allowed four hypotheses to be tested: that (1) interviewers' attitudes towards interviewing, (2) interviewers' personality traits and skills, (3) interviewers' demographic characteristics and (4) interviewers' own satisficing behaviour

may affect respondents' propensity to satisfice. The results suggest that interviewers who have a less accepting attitude to respondent refusals elicit more satisficing behaviour among their respondents. The second hypothesis has not been confirmed empirically; there is no evidence that interviewers' personalities or skills affect the satisficing behaviour among their respondents. The third hypothesis on the effect of interviewers' demographic characteristic has not been confirmed either. However, the fourth hypothesis about interviewers' own response style has been confirmed and this is an important result.

To date, research in this area has identified satisficing strategies among respondents but never the interviewers. It is clear from the results presented above that interviewers can be satisficers and that their own satisficing can influence the satisficing of their respondents; even when controlling for the remoteness and differences in the interviewers' assignments, the interviewer effect on respondents' heaping remains statistically significant. The question remains about how this happens. One possible mechanism is that satisficing interviewers can be more accepting of such behaviour among their respondents. Another explanation may be that satisficing interviewers use probing in such a way as to encourage an estimate when respondents struggle to give exact answer. Alternatively, the rounding could happen while interviewer records the answer. Interviewers who are more prone to heaping may record 'four' as 'five' or 'nine' as 'ten' and so on.

There are many commonly used measures of data quality adopted by the theory of satisficing but heaping, according to the best knowledge of the author of this thesis, has never been used as one of them. Heaping fits Krosnick's (1991) satisficing theory. Even though it is not apparent why respondents round their answers, nor are the sources for the 'digit preference' really known, rounding an answer bears the signs of some sort of 'mental shortcut' when performing the cognitive task of providing a numerical response. Since heaping is such a common problem with data in many research fields, it is hard to think of any numerical or frequency survey questions that are not prone to heaping, which makes this measure of satisficing quite widespread. As with other measures of satisficing, heaping is also a measure that is easily available from the data and is also quite trivial to compute.

The advantage of the satisficing measure used in this study is that it has been computed for respondents and interviewers using the same method and the same survey instrument. Moreover the main questionnaire and the Administrative Block are completed within a short time of each other. The results presented in this paper also illustrate how strongly heaping can manifest itself among interviewers. Even though the questions in the Administrative Block ask explicitly to 'record time to the nearest minute', the heaps in the data are apparent.

This may actually be a cause for concern to survey agencies, as interviewers are those specifically trained to follow all the instructions in the questionnaire.

One issue with the measure of heaping is that, without the external measures that can validate the response, it is impossible to state which round number is a genuine measurement and which is an effect of satisficing. A second issue is the sensitivity of this measure to distance and time. Some respondents potentially heap more because of the distance to amenities or time that has passed since the reported event. For that reason heaping should always be considered together with a measure of distance or time. The 'remoteness' variable helped to correct this problem. Unsurprisingly, it was highly statistically significant in all models considered here. Area characteristics such as urbanisation may also be a good proxy measure of the distance to amenities or transport links.

The importance of heaping and its damaging effect on data should be particularly stressed during the interviewer training process, especially that heaping does not manifest itself at the time of interview. One way of dealing with it would be to alert interviewers to the problem. If a respondent provides a rounded response, interviewers should be instructed to probe further in order to make sure that the given answer is not an estimate. The existing literature tackles the issue of modelling heaped data but not much is being done in order to understand it and prevent it. Similarly, no research on satisficing considered interviewers' own satisficing behaviour that can potentially affect their respondents. Even though the results presented in this paper do not provide an insight on how it happens, they indicate that there is a positive relationship between respondent and interviewer satisficing, resulting in increased heaping in the data, a universal problem for many research fields.

4. False economy? Testing the hypothesis that satisficing response styles reduce cognitive effort in survey respondents

4.1 Introduction

The theory of survey satisficing developed by Krosnick (1991) suggests that respondents take 'cognitive shortcuts' in replying to survey questions and as a result they provide a 'good enough' answer rather than an optimal one (Krosnick, 1991). The concept of satisficing has rapidly become the dominant framework for understanding data quality from a measurement perspective, with many indicators of satisficing such as prevalence of 'Don't know' response or inconsistency between a logical pair of questions (Kaminska *et al.*, 2011; Oppenheimer *et al.*, 2009; Roberts, 2007). A logical implication of satisficing theory is that, in saving the cognitive effort required to answer the question accurately, respondents who satisfice also complete the question more rapidly than they would had they used an optimising strategy. Therefore, the time it takes for a respondent to complete a question, referred to as a response latency, will vary depending on the extent to which they expend cognitive resources in answering (Bassili and Scott, 1996).

Response latencies have been extensively researched in cognitive psychology and have generally been used as an indirect measure of attitude strength or crystallisation (Bassili and Fletcher, 1991; Fazio, 1990). In the survey methodology context, response latencies are mainly used to evaluate survey questions during the design phase, where long response time is perceived as an indicator of a potential problem with a question (Yan and Tourangeau, 2008). The use of response latencies for methodological purposes has become more popular due to the availability of paradata, especially as a result of the automated recording of 'time stamps' in computer-assisted surveys (Couper, 1998).

The aim of this paper is to assess whether satisficing response styles actually result in reduced 'cognitive costs' among satisficing respondents. Although this is a central axiom of the theory of survey satisficing, it has rarely been evaluated empirically. The expectation is that respondents who use satisficing strategies will take less time to answer a question than 'optimisers' – those who do not use a satisficing strategy but rather expend additional cognitive resources to produce an 'optimal' answer. To date, only one published study has employed response latencies in research on satisficing (Callegaro *et al.,* 2009). Response latencies were used as a measure of the cognitive effort invested by respondents and

compared between satisficers and optimisers with an expectation that the latter group would take longer to answer the same questionnaire. However, this study used an indirect measure of satisficing, which was based on the assumed levels of motivation between sub-groups of respondents. The sample used was not drawn from the general population which also raises an issue of how general these findings might be in other contexts. The research presented in this paper extends the work of Callegaro *et al.* (2009) by comparing response latencies across respondent groups defined by their observed satisficing behaviour. Three groups are defined: 'weak' satisficers defined as respondents who provide 'heaped' responses to questions requiring a numerical estimate; 'strong' satisficers defined as respondents who provide a 'Don't know' answer; and 'optimisers', who provide a non-heaped numerical estimate. The cognitive effort required to provide each type of response is operationalised as the response latency.

The remainder of this paper is structured as follows. The next section contains an overview of research on satisficing, including the different sources of satisficing and measures used to detect it. This is followed by a section describing the concepts of paradata and response latency. The relationship between satisficing and response latencies is then considered. These sections are followed by a description of the data and the measures developed and used in this study. The model specification section is followed by the results, including a presentation of models for one item and a summary of the models for the remaining items. The paper concludes with a discussion of the main results.

4.2 Survey satisficing

Survey satisficing theory originates from economics and borrows the core notions of satisficing and optimising from Herbert Simon (1957) who observed that humans often violate the principle of rational choice; instead of pursuing 'the best' option they often settle for a 'good enough' alternative (Schwartz *et al.*, 2002). Krosnick (1991; 1996) adapts Simon's general satisficing concept to describe a particular way in which some respondents reply to survey questions. Those who satisfice do not perform all four stages of the cognitive process of answering a question – comprehension, retrieval of information, judgment and response (Tourangeau *et al.*, 2000) – and as a result provide a response that is minimally acceptable or 'good enough' but not necessarily as accurate as it might be with greater cognitive effort. In contrast to satisficing respondents, optimisers expend the necessary cognitive effort to provide the most accurate answer they can, given their resources such as knowledge and ability to retrieve the necessary information from memory (Krosnick, 1991).

Krosnick (1991) also distinguishes between weak and strong satisficing and, following his work, survey researchers have developed many measures as empirical indicators of satisficing. Acquiescence or 'yeah-saying' (Olson and Bilgen, 2011) and choosing the first 'viable' response from the list (Krosnick and Alwin, 1987) are forms of weak satisficing. This occurs when one or more stages of the cognitive process of formulating a response are not executed properly. Strong satisficing prevails when respondents choose the 'Don't know' or 'No opinion' answer or when they do not differentiate their answers on the Likert scale (Kaminska *et al.*, 2011; Lipps, 2007). This happens when one or more stages of the response process are omitted altogether.

Another indicator of satisficing behaviour is when respondents provide a rounded response to questions which require a numerical answer (Roberts and Brewer, 2001). Heaping becomes apparent when data are aggregated over individuals and takes the form of visible spikes or 'heaps' at certain integer values. One example of a question that gives rise to heaping is the number of cigarettes smoked, where modal responses are 10 and 20 per day, equivalent to small and large cigarette packs (Bar and Lillard, 2010). Although many individuals will actually smoke exactly 20 cigarettes in a day, it is likely that many others smoke somewhat more or less than this amount but report 20 as a cognitively efficient heuristic (Burton and Blair, 1991). Another reason for providing a heaped answer is so called Digit Preference (DP) exhibited by respondents (Hultsman *et al.,* 1989) for numbers that end with 5 or 0. These round numbers are also referred to in literature as prototype – an approximation of the actual number (Huttenlocher *et al.*, 1990).

Satisficing is a particular form of measurement error and so it can originate from three distinct sources: the respondent, the task/instrument and the interviewer (Sudman and Bradburn, 1974). Question difficulty and the respondent's ability to provide a response are thought to be the main factors affecting satisficing (Krosnick, 1991). Additionally, respondent motivation is argued to play an important role in triggering satisficing behaviour (Callegaro *et al.,* 2009) with less motivated respondents more likely to satisfice. Because an important function of interviewers is to motivate sample members to complete surveys accurately, this implies that in the face-to-face data collection mode, interviewers may play an important mediating role in the generation of satisficing behaviour (Groves *et al.*, 2009).

Turner *et al.* (2012) showed that the respondent's propensity to satisfice can be positively affected by their interviewer's own propensity to satisfice. Interviewers can also affect the time it takes to conduct an interview as interviewers who get more familiar with the survey,

due to continuing fieldwork, are responsible for a faster pace of the interview which may result in a lower data quality (Olson and Peytchev, 2007). It is therefore likely that interviewers who exhibit satisficing response styles when completing questionnaires themselves will be more accepting of satisficing behaviour on the part of the respondents that they interview. According to the theory of survey satisficing, a satisficing response style should result in shorter response times at the expense of accuracy. In the empirical part of the paper this assumption is tested by combining indicators of satisficing with paradata on response times.

4.3 Response latencies as paradata

The term 'paradata' was coined by Couper (1998) and was originally intended to refer to information about the data collection process routinely recorded by computer software used in Computer and Telephone Assisted Personal Interviewing – CAPI and CATI. Examples of this type of paradata are keystrokes or time stamps. The term was soon expanded to include any information on the data collection process including information on interviewers, interviewers' observations, post-interview questionnaires on respondents' experiences and so on (Kreuter and Casas-Cordero, 2010; Kreuter *et al.*, 2007; Sinibaldi, 2009). Paradata are principally utilised to improve fieldwork by enabling real-time analysis of interim response rate and sample composition data (Groves and Heeringa, 2006; Laflamme *et al.*, 2008) and to better understand nonresponse in surveys (Blom *et al.,* 2011; Durrant, 2009; Durrant *et al.*, 2010). Response latencies in surveys are an example of automatically collected paradata (Couper and Kreuter, 2013).

In cognitive psychology, response latencies have been used to measure the strength of an individual's attitude towards an issue or object, with shorter response times taken to indicate attitudes which are strong or crystallised (Bassili and Fletcher, 1991; Fazio, 1990; Heerwegh, 2003). Of more direct relevance to this paper is the use of response latencies during questionnaire design, with long response times taken as indicative of questions that are potentially problematic, for instance badly worded or in some other way unclear (Bassili and Scott, 1996; Bassili, 1996). Generally, providing correct answers takes respondents the least time, while the non-substantive responses to the open-ended questions take the longest (Yan and Tourangeau, 2008). In web surveys, questions about salient facts take the least time to answer. They are followed by questions about less salient facts and simple attitude questions. The more complex attitude questions take respondents the longest to answer (Draisma and Dijkstra, 2004; Yan and Tourangeau, 2008). Kreuter and Couper (2010) built upon Yan and Tourangeau's (2008) work and extended it to the interviewer-administered questionnaire

context. Longer questions, more words, those that involved a showcard and open questions took significantly longer to answer.

The studies noted above suggest that if a survey question is difficult or complex, it takes more time to answer because greater thought and attention is required to determine a response. Using response latency as the indicator of the cognitive effort required to answer the question is a rather new concept that has not yet been fully explored. It stems from educational measurement research (Camara, 2002; Schnipke and Scrams, 2002; Wise *et al.*, 2006). Wise *et al.* (2006) showed that in a 'low-stake' test, examinees are more likely to employ 'rapid-guessing' behaviour when replying to multiple-choice test questions as a consequence of putting very little effort into taking the test. Wise and Kong (2005) examined the effort that examinees put into completing the test by developing and using a response-time effort (RTE) scale. The time was directly linked to the effort put into completing the tests and affecting the end result. Those who took more time were more likely to adopt a 'solution behaviour' of providing a correct answer as the opposite to a 'rapid-guessing' pattern (Wise and Kong, 2005). These concepts have been imported to the field of survey methodology by Callegaro *et al.* (2009), where they used the response latency as the indicator of cognitive effort needed to answer the survey question.

4.4 Response latencies and satisficing

While response latency is mostly affected by question type and difficulty, it is also influenced by respondent characteristics. For example, older respondents and those with lower educational qualifications in general take longer to answer survey questions than younger and better educated individuals (Couper and Kreuter, 2013; Yan and Tourangeau, 2008). Similarly, evidence shows that interviewers can affect the time it takes respondents to answer survey questions (Olson and Peytchev, 2007). This study looked at the overall interview time and found that the interview time per interviewer decreased as they conducted more interviews. This phenomenon was more apparent among the less experienced interviewers. Olson and Peytchev (2007) suggested that faster interviews may result in lower data quality as respondents are being rushed and discouraged to fully engage in the cognitive process required to provide a well thought through response. If interviewers negatively affect the interview time, they may encourage satisficing strategies among respondents. However, in models predicting response latencies, the interviewer variance component has been found to be small and not well explained by interviewer demographic characteristics such as age, gender, education and ethnicity (Couper and Kreuter, 2013).

Evidence exists that those who satisfice take less time to answer questions than those who optimise and they also score lower on the Need for Cognition scale (Cacioppo and Petty, 1982) which measures an individual's tendency to engage in and enjoy thinking (Oppenheimer *et al.*, 2009). Although this finding has an important implication for the study presented in this paper, it was not the main focus of Oppenheimer *et al.* (2009), who focused instead on developing new indicators of satisficing such as the Instructional Manipulation Check (described in more detail in Paper 2) and reduction in response time was merely an observation reported by the authors.

Response latency in research on satisficing was for the first time explicitly used by Callegaro *et al.* (2009). Their main research hypothesis was that optimisers would take more time answering questions than satisficers. Additionally, they expected that satisficers would take consistently less time per question as they proceeded through the questionnaire. This is because research shows that satisficing strategies are more apparent later in the questionnaire when the respondent's motivation is reduced (Krosnick *et al.,* 1996). They compared two groups of respondents completing a web-based recruitment questionnaire. One group were job applicants while the other group comprised current employees for whom the stakes were rather low, as they were told it was simply an exercise for the Human Resources department (Callegaro *et al.*, 2009). The authors did not observe satisficing behaviour directly but made an *a priori* assumption that the job applicants would be more motivated to complete the questionnaire accurately – hence they were optimisers – than existing employees, labelled as satisficers. Their results showed that in one section of the questionnaire, the job applicants indeed took significantly more time to answer the questions. However, the decrease in response times as they progressed through the questionnaire was the same for both groups (Callegaro *et al.,* 2009). While these results are consistent with the idea that longer response times in the employee group resulted from survey satisficing, the pattern of findings could also have arisen because of differences in the composition of the two groups with regard to demographic characteristics such as age, education and cognitive ability. Additionally, because these findings are not based on a general population sample and the questionnaire was administered online, there is a question as to how generalizable they are to face-to-face surveys of the general population.

A central axiom of the theory of satisficing is that satisficing respondents save themselves cognitive effort. If the cognitive effort needed to answer the question correctly is measured as the response time then satisficing respondents also save themselves time. Moreover, the stronger the satisficing, the less time it should take to answer the question. Based on the

literature presented in this section, three hypotheses have been proposed and tested in this paper:

- Hypothesis 1: Respondents who use a weak satisficing strategy will take less time to answer the question than optimisers
- Hypothesis 2: Respondents who use a strong satisficing strategy will take less time to answer the question than weak satisficers
- Hypothesis 3: Satisficing interviewers reduce the time taken by their respondents to answer the question

4.5 Data

The data used here come from the National Travel Survey (NTS) and were collected in 2011 during 12 months of continuous fieldwork. Certain information on interviewers is obtained from the Administrative Block of the survey. Detailed information on the NTS survey design has been presented in Chapter 1 and can also be obtained from the 2011 NTS technical report (Taylor *et al.,* 2012). The NTS questionnaire is the source of information on the type of answer each respondent has given to nine NTS items analysed here; demographic information on respondents': age, gender and educational attainment; and information on interviewers from the Administrative Block used to compute the interviewers' satisficing indicator. These were linked with additional sources of information including: the NTS Blaise audit trail – the source of response times; demographic information on interviewers; measures of interviewers' personality traits and skills and, finally, area characteristics at the MSOA level. The final dataset included 5,249 NTS 2011 English respondents nested in 971 MSOAs and interviewed by 159 interviewers. Only the complete records were used. The final number of cases used for each survey item varies slightly due to the treatment of outliers.

4.6 Measures

The classic approach to measuring response latency is to use so-called active timers, which measure the time that elapses between the moment in which the interviewers finish reading the question and the moment in which respondents start their answer (Yan and Tourangeau, 2008). The alternative approach is using latent timers that measure the total duration for each item, from the moment it appears on the screen until the respondent has finished answering and the interviewer has recorded the response; both measures tend to be highly

correlated (Mulligan *et al.*, 2003). To distinguish between the two, the latter approach which is the one used here, refers to response latencies as response times.

Response times are prone to outliers and therefore require editing. One of the major sources of outliers is execution error, which makes the measured response time longer than it should be (Barnett and Lewis, 1994). For instance, it is possible that the Blaise software remains on a certain field while the respondent is making a cup of tea or a phone call interrupts the interview. Another potential cause of outliers is when the interviewer does not close the computer session properly after the interview has finished (or has been suspended until another visit). Descriptive analysis of the dependent variables suggest that this is likely to have been the case, with response times of over 1,000 seconds for some questions. Such large outliers are usually spurious observations and so should be removed from the analysis (Anscombe and Guttman, 1960; Hawkins, 1980).

For the initial detection of outliers, the 'outlier labelling' approach was used (Bliss *et al.*, 1956; Hoaglin and Iglewicz, 1987; Hoaglin *et al.*, 1986). This method multiplies the interquartile range by a factor of 2.2 to obtain lower and upper thresholds for outliers. In the case of computer-generated response times, only the upper bound threshold applies. Once the questionably high values for each dependent variable were identified, they were cross-checked for each respondent. There were six questions about distance to local amenities and three questions about distance to public transport that belonged to the same question blocks, respectively. As the response times are not independent within respondents, it is possible that a particular value, which may appear to be an outlier across all respondents, is in line with the other responses given by a particular respondent. It was assumed that in each question block for each respondent, there can only be one outlier. The Grubbs criterion[22] was used to identify the outlier at the respondent level (Grubbs, 1950). Depending on the item, between 16% and 60% of the outliers identified in the initial sweep, were in fact genuine outliers, after the Grubbs criterion was applied. They were removed from the dataset and they constituted between 0.8% and 1.8% of the total sample.

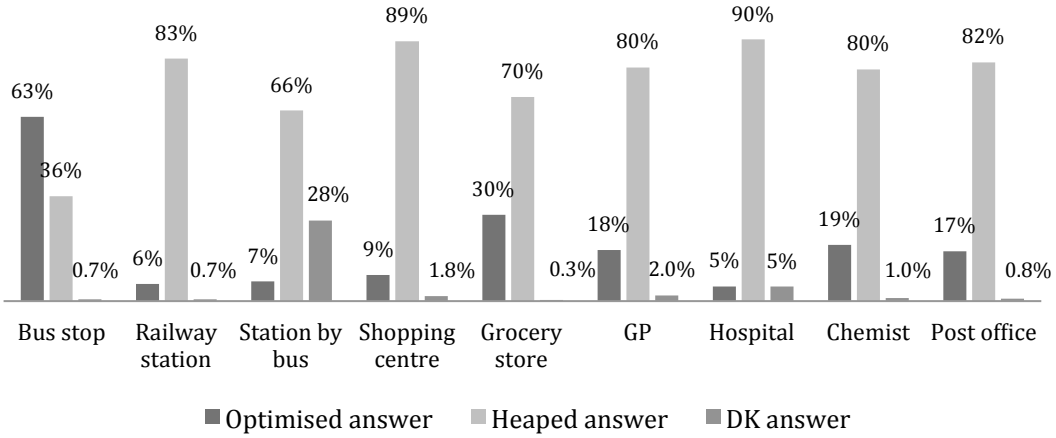
Nine NTS questionnaire items were used as dependent variables. Three questions were asked of all respondents about access to local means of transport – how long would it take to get to: nearest bus stop, nearest railway station and nearest railway station using a bus. Six questions were asked of a 50% random subsample about access to local amenities – how long would it take to get to the nearest: shopping centre, grocery store, post office, GP, hospital

[22] For the amenities variables (population of six) Grubbs' rejection rule was set at the 95% confidence level and for the transport variables (population of three) it was set at the 90% confidence level.

and chemist. For all nine items, respondents were asked to state the time in minutes. If they replied 'Don't know' then a follow up question was asked using a showcard with banded time intervals. These items are well suited to testing the research hypotheses because they are prone to heaping, which is used here as an indicator of weak satisficing. For each item, the following information was available from the audit trail: field name, time in seconds spent in the field, number of visits to the field, if the reference/help window was activated, if there was a refusal, if the 'Don't know' answer was given.

NatCen granted restricted access to the actual NTS 2011 survey responses (see also Appendix 3). The survey data were therefore used to develop a satisficing indicator for each item for each respondent as follows. If a respondent responded with a non-rounded integer they were given value of 0 – 'optimised answer'. If their answer was a multiple of five, their answer was given score of 1 and was classed as a 'heaped answer' which is a measure of weak satisficing. If the respondent said 'Don't know' they were given a score of 2 and their answer was classed as 'DK answer', which is a measure of strong satisficing. Figure 4.1 shows the distribution of the three answer categories for each of the nine items in the order that they appear in the questionnaire. In every case, except for the 'station by bus' item, the proportion of 'Don't know' responses is very low (2% or below). And, with the exception of the 'nearest bus stop' item, the 'heaped' answer is the modal response.

Figure 4.1. Type of answer given to each survey item



Information about interviewers was obtained from two different sources. The basic demographic information came from the administrative records held at NatCen. These included: gender, age (banded), years of experience as an interviewer (banded) and interviewer grade. Interviewers' characteristics such as personality traits and skills measures

were developed using the responses to the 'Who is behind the NatCen ID card?' survey that is described in more detail in Section 1.7 in Chapter 1. The main section of the questionnaire consisted of 67 Likert scale items, statements describing the type of person interviewer is. These items were grouped into scales using the guidance from NatCen and the factor analysis method. Items that were coded in reverse were recoded. The measures of each particular personality trait and skill were developed as the straight mean scale scores. All scales had a Cronbach's Alpha coefficient of 0.7 or higher. The higher the mean score on each scale, the stronger the particular personality trait or skill. The measures used were: openness, neuroticism, conscientiousness, extraversion, empathy, efficiency, flexibility, attention to detail and communication skills including persuasion (see Appendix 2 for more detail).

In addition to the interviewer personality and skills variables, a measure of interviewer satisficing was derived as follows. Each respondent was required to complete a seven-day travel diary during an assigned 'travel week'. There are two items in the NTS Administrative Block, which asked interviewers to report, to the nearest minute, how much time it took to place the diaries and explain the documents and how much time it took to pick up and check the diaries. The time taken to complete each activity was recorded in two ways: the interviewers' own report, and also using the audit trail from the CAPI programme. This affords an opportunity to obtain some additional leverage on whether individual responses should be categorised as 'heaping' indicators or as accurate responses. Figures 4.2 and 4.3 show distributions of time for the placement interview as computed from the audit trail (Figure 4.2) and as reported by interviewers (Figure 4.3). It is apparent that when interviewers report time they round it to the nearest multiple of 5, causing visible 'heaps' at these points on the distribution.

Figure 4.2. Audit trail: how long did it take to place and explain the diary and other documents? Include any time spent preparing the diaries beforehand. Record to the nearest minute (N = 5422)

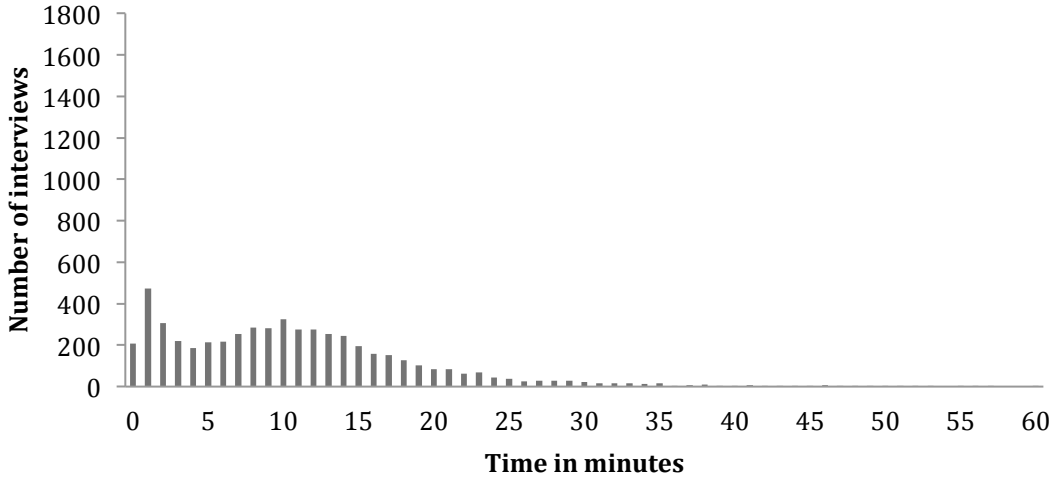
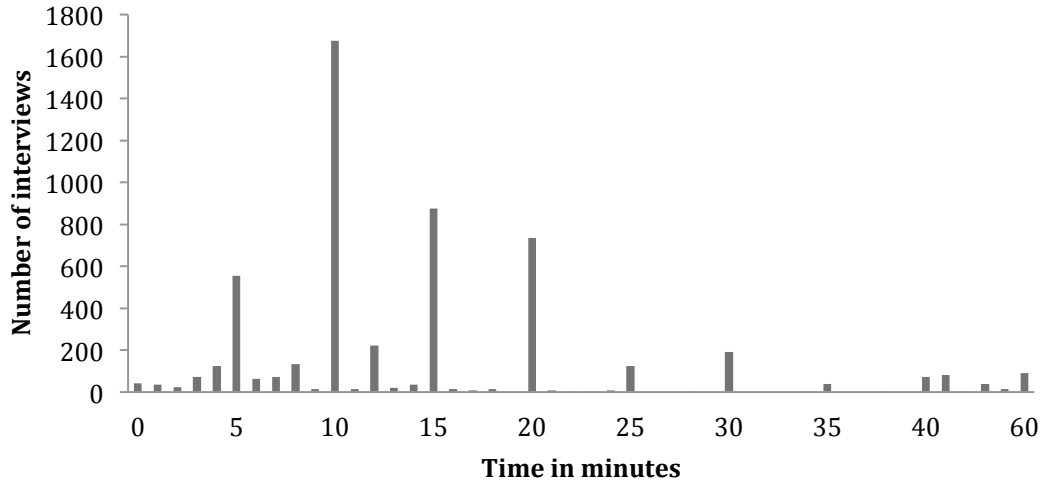


Figure 4.3. Reported by interviewers: how long did it take to place and explain the diary and other documents? Include any time spent preparing the diaries beforehand. Record to the nearest minute (N = 5422)



The satisficing indicator, based on providing a rounded response to the two Administrative Block items, was calculated for each interviewer. If an interviewer provided an answer that was a multiple of five they were given score 1 and 0 otherwise. These responses were then cross-checked with the audit trail values to identify false positives. If an interviewer reported a multiple of five, for example ten minutes, and the same value was generated by the audit trail, then the interviewer's score was changed from 1 to 0. This was the case for around 14% of responses to the diary placement question and around 8% of responses to the pick-up

interview question. The mean was taken over the two items giving interviewers a score of 0 if they did not report a multiple of five to either of the two questions, 0.5 if they heaped on one item and 1 if they heaped on both items. Then the mean was taken over the total number of interviews conducted. The minimum satisficing score was 0.058, the maximum was 1 and the mean was 0.716 across the 159 interviewers.

4.7 Model specification

The NTS data has a cross-classified structure, where respondents (Level 1) are nested within a cross-classification of interviewer and area (Level 2). The appropriate model for this type of hierarchical data is a cross-classified multilevel model (Rasbash and Goldstein, 1994). The multilevel approach is widely used in studies of interviewer effect as it allows controlling for non-random assignment of interviewers to respondents, a common problem with this type of research design (Hox, 1994; Lipps, 2007; O'Muircheartaigh and Campanelli, 1998; Schnell and Kreuter, 2005). Yan and Tourangeau (2008) used the multilevel cross-classified approach to control separately for respondent and question characteristics when analysing response times. Even though the interviewer effect is not the primary focus of this study, it is important to control for interviewer and area effects by adding relevant covariates. The study presented in Chapter Three (Paper Two) showed that both interviewers and areas can affect respondents' propensity to satisfice.

The multilevel models allow identification of separate variance components for respondent, area and interviewer as well as the incorporation of covariates at each level. The notation for the model used is:

$$y_{i(j,k)} = \alpha + u_j + v_k + e_{i(j,k)} \qquad (4.1)$$

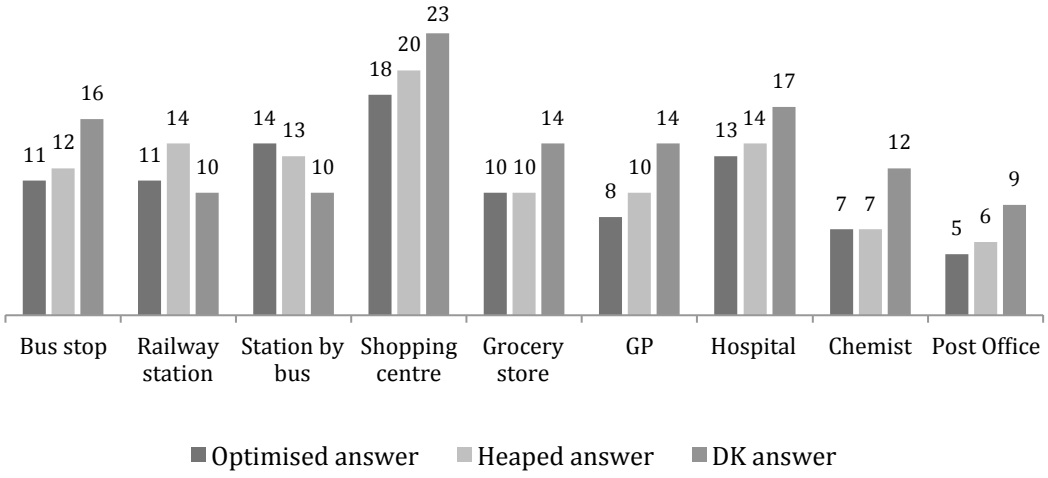
where $y_{i(j,k)}$ is a response time from the *i*th respondent in the *j*th area and interviewed by the *k*th interviewer expressed as the sample mean α (intercept) and three random error terms at the area (u_j), interviewer (v_k) and respondent ($e_{i(j,k)}$) levels. These error terms are assumed to be independent and identically distributed with zero mean and variances denoted as $\sigma_j^2,\ \sigma_k^2,\ \sigma_{i(j,k)}^2$ respectively. These variances can be used to calculate Variance Partition Coefficients (VPC) (Goldstein *et al.,* 2002) which are the ratio of each variance to the total variance expressed as percentages.

Models were estimated using MLwiN version 2.25 (Rasbash et al., 2009) using the Monte Carlo Markov Chain (MCMC) estimator (Browne 2009). The Deviance Information Criterion (DIC) was used to compare the fit of alternative models. A decrease in DIC suggests an improvement in model fit (Spiegelhalter *et al.,* 2002).

4.8 Results

Results are presented in the following order. First, Figure 4.4 shows the median response times for each item and each response type. Then six models are presented for the 'nearest bus stop' item. Table 4.1a shows the variance components model, Models 2 and 3 introduce the respondent and interviewer satisficing indicators respectively. Model 4 (Table 4.1b) introduces the remaining interviewers' characteristic variables, Model 5 the respondent covariates and Model 6 the area-level covariates. All models in this section are estimated without outliers[23]. Summary statistics from the models for the remaining eight items are then presented in Table 4.2. Finally, some additional analyses for items with very short response latencies are also discussed.

Figure 4.4. Median response times for each item and response type in seconds



For every item shown in Figure 4.4, with the exception of the 'station by bus' item, providing a heaped response takes the same time or longer compared to providing the optimised response. The 'Don't know' answer takes the longest. The 'station by bus' item generally shows a different pattern to all other items. The optimised response takes the longest at 14 seconds and the 'Don't know' response takes only 10 seconds. This is the only item that

[23]The same models were estimated using the full dataset with outliers and the results are substantially the same.

exhibits the pattern that would be expected when the theory of satisficing is considered. It may be due to a particular difficulty with this item and it is discussed further below.

Model 1, which does not include any covariates, shows that the mean time taken to answer the question on how long it takes to get to the nearest bus stop is 14.69 seconds. The area related variance is not significant. Interviewer-related variance accounts for 12.8% of the total variability and the residual variance is large, at 86%. Model 2 includes the key respondent variable, the satisficing indicator for this item. This shows the same pattern as Figure 4.4, the weak satisficers who provided a heaped answer took on average 2.17 seconds longer to answer this question than those who provided an optimised answer. The strong satisficers – those who chose the 'Don't know' response – took, on average, 7.3 seconds longer. Adding the satisficing indicator equally decreased the respondent and interviewer-related variances by 1.4%. Model 3 introduces the interviewer's heaping indicator. It has a negative coefficient, as expected, however it is not statistically significant.

Table 4.1a. Multilevel models for the 'nearest bus stop' item

	Model 1	Model 2	Model 3
Nearest bus stop time	14.696 (0.354)	13.90 (0.365)	14.818 (4.959)
Fixed effects			
Respondent's satisficing (Optimised answer)			
Heaped answer		2.175 (0.308)	2.189 (0.308)
Don't know		7.312 (1.753)	7.471 (1.727)
Interviewer's characteristics			
Interviewer's heaping score			-2.158 (1.730)
Random effects			
Interviewer (σ_k^2)	16.029 (2.393)	15.812 (2.406)	15.808 (2.392)
Area $\left(\sigma_j^2\right)$	1.311 (1.168)	1.533 (0.907)	1.430 (1.015)
Respondent $\left(\sigma_{i(j,k)}^2\right)$	107.641 (2.331)	106.135 (2.22)	106.177 (2.243)

Note: Estimates and standard errors are the means and standard deviations of 50,000 chains, with burn-in of 500. N = 5,176 respondents nested in cross-classification of 967 areas and 159 interviewers.

Model 4 includes controls at the interviewer level. The only interviewer characteristic that is statistically significant is 'efficiency'. Respondents take almost 1.5 seconds less on average if interviewed by a more efficient interviewer. However, this effect does not remain significant when the respondent characteristics are added in Model 5. Of the three respondent characteristics: gender, age and educational attainment, only age is a significant predictor of response latency; the older a respondent is, the longer it takes them to answer. Adding the respondent covariates reduces the residual variance by 2.4 percentage points.

Model 6 is the final one and includes area characteristics. Those who live in highly urban areas take slightly less time to answer the question, while those in ethnically diverse areas take on average 3.8 seconds longer to provide the answer. The final model implies that after controlling for interviewer, respondent and area characteristics, the main effect of weak and strong satisficing on response time to the 'nearest bus stop' question remains statistically significant. Both effects have the opposite directions to those specified in hypotheses one and two. Weak satisficers take longer to provide the answer than optimisers and strong satisficers take the longest of all three groups. The summary of results for the remaining eight items is presented in Table 4.2.

Table 4.1b. Statistical models for the 'nearest bus stop' item that include controls

	Model 4	Model 5	Model 6
Nearest bus stop time	14.818 (4.959)	14.541 (5.004)	14.403 (5.116)
Fixed effects			
Respondent's satisficing (Optimised answer)			
Heaped answer	2.189 (0.308)	1.647 (0.307)	1.595 (0.306)
Don't know	7.471 (1.727)	6.825 (1.734)	6.664 (1.756)
Interviewer's characteristics			
Interviewer's heaping score	-1.464 (2.037)	-1.77 (2.012)	-1.766 (1.984)
Openness	-0.302 (0.673)	-0.258 (0.653)	-0.284 (0.662)
Conscientiousness	0.888 (0.684)	0.764 (0.697)	0.874 (0.667)
Empathy	0.398 (0.765)	0.301 (0.76)	0.231 (0.757)
Efficiency	-1.426 (0.695)	-1.348 (0.714)	-1.388 (0.696)
Flexibility	-0.129 (0.677)	-0.135 (0.657)	0.013 (0.661)
Attention	0.675 (0.664)	0.716 (0.653)	0.559 (0.659)
Communication skills	-0.302 (0.773)	-0.341 (0.772)	-0.251 (0.812)
Neuroticism	0.025 (0.375)	-0.011 (0.367)	-0.072 (0.396)
Extraversion	0.711 (0.377)	0.687 (0.388)	0.558 (0.391)
Female (Male)	-1.287 (0.799)	-1.306 (0.805)	-1.191 (0.808)
Interviewer's grade (A)			
D	-0.311 (2.789)	-0.324 (2.777)	0.153 (2.87)
C	-0.467 (2.715)	-0.627 (2.633)	-0.089 (2.727)
B	0.33 (2.169)	0.362 (2.128)	0.357 (2.163)
S	-1.229 (2.864)	-1.249 (2.767)	-0.637 (2.925)
T	-2.746 (3.43)	-3.065 (3.554)	-2.304 (3.64)
Interviewer's age (30<)			
30-39 years	0.516 (5.461)	1.161 (5.554)	0.74 (5.609)
40-49 years	1.237 (4.819)	1.424 (4.919)	1.448 (5.081)
50-59 years	2.035 (4.743)	2.254 (4.81)	2.552 (4.986)
60-69 years	2.98 (4.779)	3.15 (4.812)	3.125 (5.07)
70+ years	1.697 (5.148)	1.568 (5.196)	1.754 (5.335)
Interviewer's experience (1<)			
1yr + N mths	-2.183 (2.064)	-2.218 (2.035)	-2.548 (2.133)
2yrs + N mths	-1.287 (2.701)	-1.215 (2.566)	-1.693 (2.681)
3/4yrs + N mths	-2.016 (2.477)	-2.076 (2.413)	-2.597 (2.51)
5/6 yrs + N mths	-2.98 (2.701)	-2.719 (2.609)	-3.456 (2.695)
7-10 yrs + N mths	-2.936 (2.599)	-2.558 (2.525)	-3.112 (2.627)
11 yrs +	-1.937 (2.89)	-1.645 (2.918)	-2.398 (2.971)
Respondent's characteristics			
Female (Male)		0.314 (0.296)	0.327 (0.289)
Age		0.101 (0.009)	0.099 (0.01)
No (yes, qualifications with a certificate)		0.387 (0.37)	0.481 (0.371)
Area characteristics			
disadvantage			-0.301 (0.258)
urbanisation			-0.527 (0.246)
social mobility			-0.019 (0.204)
age profile			-0.162 (0.2)
housing structure			-0.177 (0.201)
crime rate			-0.109 (0.382)
ethnic diversity			3.84 (1.694)
Random effects			
Interviewer (σ_k^2)	15.778 (2.613)	15.443 (2.444)	15.398 (2.557)
Area (σ_j^2)	1.214 (0.744)	0.433 (0.538)	0.969 (0.836)
Respondent $(\sigma_{i(j,k)}^2)$	106.369 (2.207)	103.792 (2.15)	103.368 (2.142)

Note: Estimates and standard errors are the means and standard deviations of 50,000 chains, with burn-in of 500. N = 5,176 respondents nested in cross-classification of 967 areas and 159 interviewers.

For all nine questions, with the exception of 'nearest bus stop', providing a heaped answer takes the respondent longer than providing the optimised answer. This relationship is statistically significant in four of nine models, including the 'nearest bus stop' model. Providing the 'Don't know' answer has the longest response latency in five final models out of nine, and is significant in four instances. The interviewer satisficing indicator did not reach statistical significance levels in any of the nine models. However, in all models the coefficient for this variable was negative, in line with the expectation stated in the third hypothesis.

Table 4.2. The effect of weak and strong satisficing on the response time for the remaining eight items

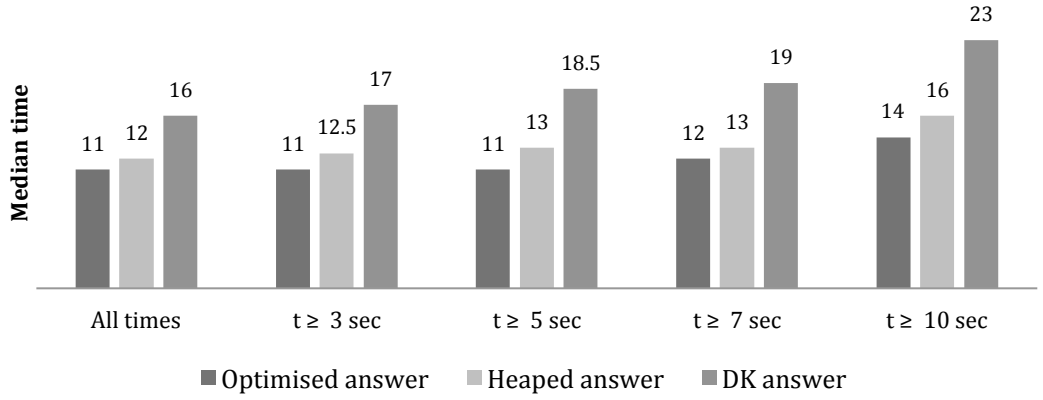
	Railway station	Station by bus	Shopping centre	Grocery store	GP	Post Office	Hospital	Chemist
Respondent's satisficing (Optimised answer)								
Heaped answer	2.147	-0.120	0.124	0.863	2.012	0.890	0.192	0.325
	(0.713)	(0.625)	(0.926)	(1.973)	(0.491)	(0.335)	(1.076)	(1.337)
Don't know	-1.584	-4.236	3.367	-0.228	7.393	6.292	-0.997	3.145
	(1.040)	(0.667)	(2.339)	(3.723)	(1.427)	(1.569)	(1.581)	(1.571)
Interviewer's satisficing								
Interviewer's heaping score	-3.306	-2.256	-3.394	-3.603	-1.207	-1.571	-1.720	-1.162
	(2.441)	(2.925)	(3.654)	(1.838)	(2.065)	(1.406)	(2.913)	(1.585)

It is clear then that providing a heaped answer and saying 'Don't know' do not save the respondent any time. Indeed, these results suggest that these apparent satisficing response styles actually take significantly more time to formulate on the majority of items examined here. On only one item was either response style associated with a significantly lower response latency compared to an optimised response.

One possible explanation for these counterintuitive results may be the assumption that those who readily know the answer will reply very quickly, and are more likely to optimise. An example would be a respondent who travels to work every day and knows it takes her exactly eight minutes to get to the bus stop because she does it frequently and she does not wish to spend more time waiting at the bus stop than is necessary. Figure 4.5 shows an example for

the 'nearest bus stop' item for various samples with very short response times excluded, of three seconds and more, response times of five seconds and more etc.

Figure 4.5. Median times for the 'nearest bus stop' item for various truncated samples



This shows that the same pattern for each response type is present, regardless of whether the instant, possibly optimised responses, are excluded or not. This is also true for the remaining eight items.

4.9 Discussion

Response latencies have been utilised in methodological research for a long time but there has been a renewal of interest in them recently due to the development of the concept of survey paradata. Response times automatically collected by computer software have become more accessible. This study took advantage of information provided by the NTS 2011 audit trail. This is the first time that NTS computer-generated paradata have been extracted in a usable format and have been used in methodological research. More paradata were utilised as well, including information on interviewers, mainly used to control for the interviewer effect when analysing response times.

The theory of survey satisficing implies that some respondents provide a 'good enough' response only as an effect of using 'cognitive shortcuts' and limiting the effort needed to provide a 'well-thought-through' response. The theory distinguishes forms of weak and strong satisficing and in the case of both, it is expected that those who satisfice save themselves time (Oppenheimer *et al.,* 2009; Callegaro *et al.*, 2009). In this paper, two indicators of satisficing were used at the question level: the 'Don't know' answer as the measure of strong satisficing and heaping as the measure of weak satisficing. Both indicators were compared with optimised responses. The expectation for the results was in line with the satisficing theory, that those who give satisficed response save themselves cognitive effort

measured as time. Three hypotheses have been tested: weak satisficers will take less time to answer than optimisers, strong satisficers will take less time than weak satisficers and satisficing interviewers will shorten the response time of their respondents.

The results showed a consistent pattern that the rounded/heaped response took longer than the optimised response and the 'Don't know' response in most cases took the longest. The first two hypotheses were hence not supported. As for the third hypothesis, although the observed interviewer satisficing effect was indeed negative, it was not statistically significant and also the null hypothesis could not be rejected. The results of this study are counterintuitive and in opposition to the theory of satisficing. However, there is one question that is an exception to the above pattern.

It was noticed that the awkwardly phrased question – 'railway station by bus' – followed the response pattern in line with the original expectations and theory of satisficing. This question asks how long would it take to get to the nearest railway station by bus. This is a complex concept and presumably only those who travel frequently by bus to a railway station would readily know the answer. The question is possibly quite difficult to answer for the majority of individuals. The mean response time for this question is in line with other questions in the same block but it has a much higher proportion of 'Don't know' responses than any other question investigated in this paper. Additionally, unlike the other questions, both satisficed responses take less time than the optimised response. One possible explanation is that only with this item is 'true' satisficing observed. The question is so difficult that most of the respondents do not even attempt to answer it and go straight for the 'Don't know' option or the very rough estimate. If this is true, then the satisficing indicators used in this study could possibly be indicators of optimising behaviour instead.

It is impossible to generalise this result as it is only based on one item as in the literature a respondent is considered to adopt the satisficing strategy when choosing the 'Don't know' response across a number of items rather than for one question only. It is unlikely that respondents would adopt such tactic over the items analysed in this paper. This is because each 'Don't know' response to the main question triggers a follow up question and presumably after the initial few items satisficing respondents would abandon this strategy, possibly in favour of heaping instead.

However, at the same time it is possible that respondent can 'switch' between response styles and satisfice only with certain items such as difficult questions. What speaks in favour of this idea is the pattern observed for the presumably easiest question analysed in this study – 'How long would it take to walk to the nearest bus stop?' This question is also different from

the rest as it generated the highest proportion of optimised responses (63%). It seems that if the question is 'easy' then respondents are more likely to optimise and if the question is 'difficult', like the 'station by bus' item, then respondents are more likely to satisfice. These results potentially suggest that with easy questions and highly motivated respondents, the commonly used satisficing indicators may indeed be appropriate indicators of optimising. Those who do not instantly know the answer will undertake a mental exercise to provide an estimate, that takes longer, and only if they cannot come up with that either will they resort to the 'Don't know' option, which takes the longest. Conversely, in more difficult questions, the satisficing indicators indeed show shorter response times and indicate that the respondent did not engage fully in the cognitive process of formulating the answer, going instead for the quicker and 'easier' response option such as 'Don't know'. This may be supported by the fact that these questions are asked at the beginning of the interview when the motivation of respondents is at its highest. There is a potential for studies on satisficing that should employ some sort of objective measures of question difficulty.

The study presented in this paper has also shown how automatically collected paradata can be used for methodological purposes. In fact, it is worth noticing that no actual survey results have been used in this study and all variables in the models presented here were paradata. The emphasis should be put on the format of automatically collected paradata as already mentioned in Section 1.5.1. The NTS 2011 dataset was the first NTS dataset to be accompanied by the audit trail file in a usable format. Nevertheless, it became obvious that automatically collected time stamps are prone to measurement error. It turned out that Blaise software does not provide an error-free way of collecting paradata. Before the response times could be used, the problem with outliers had to be solved. There is no 'catch all' method for dealing with outliers and the approach presented in this study is only one of many that could potentially be used.

One major advantage of automatically collected time stamps is that they provided information that validated the interviewer heaping score. Thanks to the automatically recorded times for the placement interviews, false positive heaping scores among interviewers could be identified. Indeed, it became apparent that in 14% of cases the time reported by interviewer was a genuine five or ten-minute spell, rather than a rounded or heaped answer. Putting the measurement error of paradata to one side, this gives an important insight to the prevalence of 'true' heaping.

While this research based purely on paradata has proven to be very useful, the lack of the actual survey responses made it impossible to control for the 'remoteness' of survey

respondents. It has already been demonstrated that the distance to the amenities/transport links can affect the extent to which respondents satisfice. Models estimated in this study lacked this control variable and relied on area-related variables like urbanisation as a proxy for the 'remoteness'.

5. Conclusions

Social surveys remain the main source of many indicators relating to modern society such as the unemployment rate, level of crime and quality of life, to name but a few prominent examples. Most of the UK-wide social surveys rely on a face-to-face mode of data collection. Social changes in recent decades in the UK such as the increase of single-person households and increased working hours, mean that it has become harder to find someone at home with whom to conduct an interview; also potential respondents have less time to devote to interviewers. As a result, survey nonresponse has become an important issue for survey methodologists and most of their research efforts have been concentrated on understanding and reducing nonresponse errors in surveys. While the strand of methodological research on nonresponse remains strong, simultaneously research is being conducted on measurement errors in surveys, essentially on what happens after the interview has been secured. The research presented in this thesis has been concerned with measurement error in surveys and has focused particularly on the interviewer as its main source.

5.1 Measurement error in surveys

Measurement error in surveys is basically a difference between the true value that is being measured and the value provided by a respondent – their survey response. It has the form of a systematic bias when respondents under-report their age or number of sexual partners and it has the form of a random variance when the respondent's response would vary from one survey administration to another, if the same respondent were to be asked the same question twice, on different occasions. Face-to-face data collection is perceived to result in high data quality as the crucial role of the interviewer means that the respondent can be helped with providing a response and can be appropriately motivated. While this is certainly the case, it is also possible that interviewers deviate from the standard interviewing procedures and themselves become a source of measurement error. Interviewers can make mistakes in recording responses or read survey questions wrongly. Some interviewers are better suited than others to interviewing survey respondents: attention to detail, good communication skills, empathy or openness are all potentially linked with interviewer performance and influence the quality of the data they collect.

Respondents can also be a source of measurement error due to survey satisficing (Krosnick, 1991). Satisficing occurs when respondents do not put enough effort into answering survey questions correctly and instead use a range of strategies to provide only a 'good enough' response. These strategies are not desirable from a methodological point of view and can

substantially skew the final survey results. Respondents' cognitive skills, interest in the survey subject and motivation to participate in the survey are key triggers of such behaviour. These, on the other hand, also depend on the difficulty of the task, design of the questionnaire and on interviewers and their skills to appropriately motivate their respondents. While a lot of attention in the research literature has been given to respondents and questionnaires as the sources of measurement error in survey, the interviewer effect has been less well researched.

There are at least two reasons why the interviewer effect has been somewhat neglected by researchers in comparison to other sources of measurement error. First, there is not much detailed information on interviewers available for methodological research. Second, the use of complex sample designs in modern surveys means that interviewer effect is difficult to estimate as it is confounded with the area effect. The studies presented in this thesis address these two issues by taking advantage of the recent development of survey paradata and of innovations in multilevel modelling. So far, survey paradata have mostly been used to investigate nonresponse, for example in the responsive survey design when interim response rate has been analysed and the data collection process has been adjusted appropriately in order to maximise the final response rate. Research in this thesis has shown that both automatically generated paradata, such as time stamps, and paradata collected as a separate exercise, such as surveys of interviewers, are useful in research on measurement error.

It is difficult to assess interviewer-related error as a systematic bias without any external validation criteria. It is, however, more straightforward to estimate the interviewer effect as a proportion of the total variance of survey estimates. The multilevel approach allows disaggregation of the total variance of a survey estimate to show what proportion of can be attributed to the interviewer. The same statistical approach takes away the problem of non-random assignment of interviewers and respondents by accounting for the area effect and respondents' characteristics hence controlling for the potential differences between interviewers' assignments.

5.2 Key findings

The main purpose of the research presented in Paper One (Chapter Two) was to estimate the interviewer-related variance of survey estimates and investigate whether or not interviewers' characteristics explain some of the variance they introduced. Paper Two (Chapter Three), also concerned with measurement error, investigated whether certain interviewers encourage undesirable satisficing behaviour among their respondents and, if so,

what are their characteristics. The third paper focused on the indicators used to detect satisficing behaviour among respondents. Commonly used satisficing indicators were evaluated using response time as a measure of cognitive effort invested by respondents in formulating their answer. Research presented here showed that paradata are useful in understanding measurement error in surveys. Both the automatically collected paradata, such as time stamps, and paradata collected specifically for methodological purposes offered a valuable insight, especially when explaining measurement error originating from the interviewer.

5.2.1 Interviewer-related variance

All 13 NTS items investigated in Paper One turned out to be prone to the survey design effect and some of this was attributable to the interviewer. On average, interviewers were responsible for 4.6% of the total variance. The 'satisfaction' items were more prone to the interviewer effect, i.e. proportionally more variance was attributed to interviewers than with the 'duration' items, which were more factual in nature and less open to interpretation. This finding is in line with other research in this area (Lipps, 2007; Hox, 1994; Schnell and Kreuter, 2005).

Six paradata variables relating to the interviewer were used to partition the variance they introduced: satisfaction with pay, determination, NTS experience, level of communication skills, conscientiousness and extroversion. Interviewers not satisfied with their pay, the less determined ones, those with reasonable NTS experience (three to five administrations) and those with good communication skills were found to introduce, on average, more variance than their counterparts. The results observed for personality traits were somewhat counterintuitive. Those responsible for the most of total variance were interviewers with high conscientiousness and low extroversion levels.

5.2.2 Interviewers' effect on satisficing

In Paper Two, satisficing was operationalised as heaping, specifically providing a rounded answer that is a multiple of five. This satisficing indicator could also be computed for interviewers thanks to the availability of the Administrative Block questions asking interviewers to state 'time the placement and pick-up interviews took'. Four hypotheses were proposed and tested, all involving interviewers' characteristics and their effect on respondents' propensity to satisfice. A strong, statistically significant relationship was found between respondents' satisficing and interviewers' own propensity to satisfice. Respondents interviewed by satisficing interviewers were more likely to satisfice themselves. This result

suggests that interviewers' own response style can affect their respondents. Moreover, interviewers who are more determined to secure the interview (disagree with the statement that the refusal should be accepted of reluctant respondents) elicit more satisficing behaviour among their respondents.

5.2.3 Satisficing and response latencies

The theory of satisficing states that satisficers, by saving themselves the mental effort needed to answer the question correctly, also save themselves time by arriving at an 'acceptable' answer more quickly. In Paper Three (Chapter Four), the response times for nine NTS items asking about time taken to access local transport links and amenities were analysed using a cross-classified multilevel model. The key independent variable was the satisficing indicator with three categories: optimised answer, heaped answer and 'Don't know' answer. A range of controls at respondent, interviewer and area levels was added to the model.

The results showed that providing the optimised response took respondents less time than providing a satisficed response. Providing the 'Don't know' response consistently took the longest. The relationship between the satisficed response and time that was uncovered in this study is in contradiction to that postulated in the theory of satisficing. There was one exception to this pattern, a more difficult and awkwardly phrased question – 'railway station by bus'. This question potentially triggered different response formulation mechanisms. It seemed that respondents not readily knowing the answer chose a heaped or 'Don't know' answer and this took them less time than providing an optimised response. It is possible that only responses to this item exhibit 'true' satisficing.

5.2.4 The value of paradata

This thesis demonstrated that paradata are useful in research on measurement error. The interviewer-related paradata used in this study are examples of information collected separately from the methodological exercise. The advantage of such a methodological study is that interviewers' personality traits, skills and attitudes can be used to inform research using different survey data conducted by the same survey agency and their interviewer network. Moreover, these can be linked with interviewers' performance indicators, such as contact and cooperation rates, to better understand the effectiveness of their work, regardless of the survey they work on. Finally, such detailed information on interviewers can be used in large-scale nationally representative surveys while many older studies on the interviewer effect tend to use small experimental designs. Obviously, such additional studies are expensive and gradually will become out of date due to turnover of the interviewer workforce.

While collecting 'external' paradata can be expensive, there are many types of paradata that are simply by-products of computer systems such as Blaise supporting the data-collection process. While these cost virtually nothing to collect, there is a nontrivial cost of specific software and staff time associated with extracting this information into a usable format. Research presented in Paper Three is a good example of the use of audit trails. It provided information on response times to each survey item in the NTS as well as some additional information such as indicators of whether or not requests for information were refused, or how many times each field was visited. Each survey item was associated with a range of audit information in an ordered manner and it allowed analyses with minimal data manipulation. One downside to this approach was lack of information about the order in which these questions were asked of a respondent. Many variables were sorted in the file by their name rather than their position in the questionnaire. This could be a major drawback of the extracted audit trail if researchers are interested in investigating the response times and other data quality indicators in relation to passing time of an interview, e.g. to investigate 'speeding up' towards the end of an interview.

Both types of paradata used in this thesis are prone to error. The 'Who is behind NatCen ID card?' is as subject to errors as any other survey. Some missing data affected how the PCA was conducted. It needs to be remembered that all psychological measures used in the study are self-reported and rely on a shortened battery of questions concerning interviewers' personality traits. Response times recorded by the Blaise software are subject to the outlier problem. Making interviewers' aware of these underlying automated processes, so they suspend the session or close the session properly when the interview is being interrupted or completed, could potentially alleviate the problem of outliers.

Even though the response times were subject to measurement error they successfully served as the validation criterion for the interviewers' heaping measure in Paper Three. Times reported by interviewers could be cross-checked with the actual times recorded by the Blaise software. Heaping is a highly prevalent data problem and there are very limited ways in which the 'true' round value can be distinguished from 'heaped' round value. Using automatically collected paradata is one way of dealing with that problem. In this thesis it has been identified that around 14% of round values can be seen as 'genuine' and the remaining round values are the result of heaping.

5.2.5 Statistical modelling

The multilevel models employed in this thesis successfully allowed interviewer and area effects to be disentangled and their separate variance components to be estimated. Firstly, the structure of the NTS data is cross-classified, which means that each interviewer worked in more than one PSU. This is not always the case in social surveys and usually happens if the data have been collected during continuous fieldwork. NTS is one example of such a survey design, the Crime Survey for England and Wales being another. The continuous fieldwork is an important feature of these designs due to the seasonality of the phenomena measured in these surveys.

Secondly, the lack of interpenetrated design means that potentially the effect of different interviewer assignments needs to be controlled for statistically. As already stated, multilevel models make it possible to include covariates at respondent, interviewer and area levels simultaneously. While this is the main advantage of this type of model, it can only be utilised if the covariates are available. Respondent level covariates were easily available from the demographic section of the individual NTS questionnaire. Similarly, basic interviewer characteristics were also easy to obtain as they are routinely collected by research agencies (as opposed to personality traits and skills). The area-level variables were also available as certain PSU characteristics are included in the NTS dataset. However, these variables were rather basic hence MSOA level variables were used instead.

5.2.6 MSOA-level variables

Switching from PSUs, which were postcode sectors, to MSOA-level variables enhanced the naturally occurring cross-classification in the data. Moreover, using MSOA as a unit allowed linkage of much more information on the particular areas using the 2001 census indicators. MSOA-level area characteristics were used for statistical control in every model presented in this study. Furthermore, in Paper Three, area characteristics such as urbanisation or ethnic diversity were used as proxies for distance to amenities and transport links. Since there seems to be a positive relationship between satisficing/heaping and distance, these were used instead of the unavailable 'remoteness' indicator.

5.3 Limitations of the study

The studies presented in this thesis have a number of limitations, one of them being the lack of an interpenetrated design. While the cross-classified multilevel models commonly used in similar research are one way of dealing with this problem, this does not guarantee that the

area and interviewer effect are completely separated. Moreover, it does not provide a clear indication of what proportion of the interviewer effect is due to variability and what is a systematic bias such as the social desirability effect.

Nevertheless, the results presented here are in line with research that used an interpenetrated design, for example the 'satisfaction' items are more prone to the interviewer effect than 'duration' items. Also, the area-related variance component is consistently larger than the interviewer-related component, which again is to be expected considering the nature of the NTS. Finally, the expensive interpenetrated design is rarely used beyond the experimental methodological research, meaning that results from a study like this one better relate to 'real life' survey scenarios.

Another limitation of the research presented in this thesis is the nature of questions in the NTS. Research on measurement error suggests that the survey items most prone to that effect are sensitive questions, opinion and attitudinal questions. This suggests that questions of this type would be the most suitable to investigate the interviewer effects. The NTS questionnaire includes mainly questions about salient facts regarding travel, hardly a sensitive or difficult topic. On the one hand, a large interviewer effect is not a prerequisite for further investigation, but on the other hand, interviewers' personality traits, skills or characteristics such as gender and age could potentially explain more of the interviewer-related variance if these were questions about opinions or attitudes. Nevertheless, it needs to be noted that some NTS questions investigated in these studies use showcards and in doing so increase the interviewers' involvement and potentially their effect. Indeed, these 'satisfaction' items saw a larger interviewer-related variance component than the 'duration' items.

Despite these limitations, the NTS uses a rolling design where natural cross-classification of areas and interviewers occurs. Moreover survey data are accompanied by extensive paradata both automatically collected and external sources as MSOA level variables and interviewers' survey. The NTS questionnaire includes a fair number of questions to which the response is an integer. For that reason, the NTS dataset provides enough information to develop the new measure of satisficing – heaping. Unfortunately, the NTS does not include other question formats that would provide more traditional indicators of satisficing such as non-differentiation or acquiescence. For that reason, in Paper Two only one measure of satisficing has been used. Also, because 'digit preference' is such a strong phenomenon, some of the items investigated in Paper Three have a very high proportion of rounded/heaped answers, up to 90% for the 'hospital' item, making some of the analyses potentially less robust.

The approach in Paper One describes general patterns across 13 items, but would benefit from more items being investigated. This was not possible due to a limited number of continuous items in the NTS questionnaire. Similarly, in Paper Three more items characterised by various levels of difficulty should be investigated to support the initial finding that providing the 'Don't know' response may not be a straightforward satisficing indicator. It is possible that with difficult questions it is a measure of satisficing, while with easy questions it may be an indicator of optimising. This conclusion is based on one question only – travel by bus to the nearest railway station – hence the suggested explanation needs to be treated with caution.

Some analyses presented in Paper Two suggest that rounding/heaping on the NTS items may be linked to distance between the household and transport links/amenities. The greater the distance, i.e. the more remote the area in which the respondent lives, the more likely they are to round their answer. Statistical models in Paper Two include the 'remoteness' variable to account for it. Unfortunately, the same variable was not available in the data used in Paper Three and it was only possible to introduce some limited area controls such as urbanisation and ethnic diversity to the model in an attempt to compensate for this. Although the MSOA-level area variables proved to be useful in the models used in this thesis, it should not be forgotten that they were initially computed to inform research on fear of crime.

Survey satisficing is linked with the cognitive abilities of respondents. The lower the cognitive skills, the higher chance of satisficing. Respondents' educational attainment is one variable commonly used by researchers as a proxy measure for cognitive abilities. Models in Paper Three contained the binary educational attainment variable (Do you hold qualifications confirmed with a certificate? Yes/No), but it was not a statistically significant predictor of the time used to provide a response. One reason for this may be that this variable was not the most appropriate measure of educational attainment.

5.4 Implications for survey practice

In order to provide survey statistics of the highest quality, research agencies need to focus not only on nonresponse error but also on measurement error. Two out of three main sources of measurement error can be manipulated by the survey agency; these are the survey instrument and interviewer. The findings emerging from this study provide more insight into survey measurement error and the interviewer as its source and have important implications for survey practice. Moreover, they are based on a large-scale nationally representative survey, making them widely applicable.

Research presented in this thesis has shown that the interviewer is responsible for introducing variance into survey estimates, even for survey questions traditionally less prone to interviewer effect, such as NTS factual questions. Thanks to the multilevel modelling techniques, quantifying this variance is fairly easy, meaning that it could potentially be used as one of the interviewers' performance indicators alongside cooperation rates and other measures. These could be calculated across items within one survey or across all surveys on which the interviewer worked and potentially used to identify 'good' and 'bad' interviewers. It has the potential to inform new methods of reducing survey measurement error by making assignments dependent on interviewer characteristics or compensating for interviewer assignment at the post-survey adjustment stage. For example, interviewers identified as more prone to heaping would not be assigned to conduct surveys that involve a large number of numerical responses. Interviewers' propensity to satisfice, measured using the interviewer questionnaire or Administrative Block questions completed by interviewers, could also be used as an indicator of interviewers' performance.

Interviewers' own propensities to satisfice, measured by heaping, could be used as a staff selection tool or could also be used to identify potentially problematic interviewers within the existing network. Potential candidates could be asked to complete recruitment questionnaires designed to assess their propensity to satisfice. These questionnaires could include many different indicators of satisficing. Most certainly the negative implications of heaping should be incorporated into the training materials for interviewers, as the damaging effect of rounding is not apparent at the interview level. Results presented in papers Two and Three clearly show that interviewers round their responses even when they are specifically asked to record time to the nearest minute. That suggests they are not very good at following questionnaire instructions themselves.

The biggest implication of the research presented in Paper Three is the question raised about the validity of the 'Don't know' response as an indicator of satisficing behaviour. Many studies on satisficing use the 'Don't know' response as data quality indicator. The satisficing respondents are those who choose 'Don't know' response across many items in the questionnaire. There is no study on satisficing that does not use this indicator; yet, it has not been properly scrutinised. It is possible that, depending of the level of difficulty of the question and other factors, the 'Don't know' response may actually indicate an optimiser rather than a satisficer. Potentially, even if a respondent chooses 'Don't know' across a number of items, in some cases it may be due to genuine lack of knowledge, while in others it may be due to question difficulty leading to satisficing on this particular item. Similarly,

rounding an answer is a problematic response strategy that can substantially distort the data but it is not synonymous with saving time or cognitive effort by a respondent.

Finally, the study has important implications in relation to paradata. It showed that information about the survey process can be useful for methodological research in the area of survey measurement. This is believed to be the first time that such extensive paradata on interviewers have been used for research on measurement error and they provided tangible outcomes that could inform survey practice. In fact, models in Paper Three use only paradata, and provide useful results without even considering the actual survey responses. This thesis also stressed the importance of extracting audit trail data so they can be used in methodological research and showed how the problem with outliers can be addressed. While computers have been used in face-to-face interviews for many years and have been collecting process data in the background, these data still remain a largely untapped resource. Research presented in Paper Three shows that investing some staff time into extracting the audit trail provides highly usable data. At the same time it uncovered some issues with the formatting, suggesting that more information should be extracted, including the position of each question in the survey. This is particularly important when questionnaires involve complicated routing and no two interviews involve the same number of questions.

5.5 Further research directions

Quite a lot of attention has previously been given to interviewers as a source of nonresponse error and paradata have been used to inform research on nonresponse; interviewers' personality traits and skills have been used to investigate their contact and cooperation rates. However, the use of interviewer paradata to research measurement error has received much less attention in the research literature. Its focus largely remains in the area of question types and respondent characteristics. Interviewers' personality traits, attitudes and skills as shown in Paper One turned out not to be as useful in explaining the interviewer effect as expected. One reason for this may be the factual nature of the NTS items. The same paradata should be used with other surveys that include questions about attitudes and opinions. One study that uses the same interviewer network and contains such questions is 'Understanding Society'.

Moreover, although a strong, statistically significant positive association between interviewers' and respondents' satisficing has been found in Paper Two, it is not clear what mechanisms underlie this process. In other words, it is unknown why interviewers who satisfice tend to get higher rates of satisficing among their respondents. It is possible that such interviewers are more inclined to tolerate satisficing behaviour of their respondents or,

equally, they may actively encourage it through inappropriate probing; for example, encouraging struggling respondents to provide an estimate rather than an exact number. Without knowing the reasons behind the observed relationship, it is possible that it can arise as a methodological artefact due to a non-random assignment of respondents to interviewers. One possibility would be the already mentioned 'remoteness' of a respondent and potentially other such characteristics that could be 'external' causes of satisficing. An experimental design study using cognitive interviewing could establish what mechanisms cause such a relationship and whether appropriate interviewer training might potentially decrease the negative effect of interviewers' own response styles on the quality of survey data. Additionally, the study presented in Paper Two could be extended by using additional satisficing indicators, both for the interviewer and respondent.

As suggested in Paper Three, more research should be done on the commonly used indicators of satisficing. All the hypotheses tested that were consistent with the line of reasoning of the theory of satisficing were rejected. Response times for optimised and satisficed responses should be compared against some objective measure of question difficulty. This type of analysis, undertaken for a substantial number of items, would give more insight as to whether current satisficing indicators indeed measure satisficing rather than optimising. Heaping is the new measure of satisficing that emerged in this thesis, which creates potential for other researchers to use it in further studies on satisficing. Rounding numerical values and 'digit preference' are very common among respondents with a particularly damaging consequences for descriptive statistics, which means that investigating and better understanding of heaping has a lot of potential for survey methodology and research into measurement error.

Appendices

Appendix 1: Who is behind the NatCen ID card?

NatCen
National Centre for Social Research

P2779

Confidentia
lity
The information you give us will be treated as strictly confidential.
Your name and address will not be connected to your answers in any way.

WHO'S BEHIND THE NATCEN ID CARD?

A research study carried out by the Survey Methods Unit and the University of Essex.

May 2008

Thank you for taking part in this study which has been funded by the Economic and Social Research Council.

1 **Have you <u>ever</u> worked as an interviewer for an organisation other than NatCen? Please include any work you might be doing now.**

◆ Please indicate your answer by circling the appropriate code.

◆ Code first that applies.

Yes, currently	1
Yes, since 1st January 2006	2
Yes, but <u>not</u> since 1st January 2006	3
No, never	4

1013

2 **How long in total have you worked as an interviewer on <u>social</u> surveys, which involves interviewing people in their own homes?**

It is not important whether this is your main employment or just part-time. Please include your current work and work you may have done earlier, including for other organisations.

Years 1017 ____ and months ____ 1014-

3 **Apart from interviewing people in their own homes for a social survey, have you ever done any of the following activities?**

◆ Please indicate your answer by circling the appropriate code.

	Yes	No	
3.1 Other survey interviewing (over the phone or market research)	1	2	1018
3.2 Other non-survey interviewing (for recruitment, benefit office, etc.)	1	2	1019
3.3 Activities involving interaction with the general public	1	2	1020
3.4 Activities involving 'cold calling' at people's homes	1	2	1021
3.5 Activities where you needed to persuade people (sales job, fund raising, etc.)	1	2	1022

4 **Thinking about jobs in general, how important do <u>you personally</u> think each of the following aspects are in a job?**

◆ Please indicate your answer by circling the appropriate code.

How important is …		Very important	Important	Not important	Not important at all	
4.1	**…a job that offers good pay?**	1	2	3	4	1023
4.2	**…a job that is interesting?**	1	2	3	4	1024
4.3	**…a job that allows me to work independently?**	1	2	3	4	1025
4.4	**…a job that involves interaction with people?**	1	2	3	4	1026
4.5	**…a job that allows me to decide my times or days of work?**	1	2	3	4	1027
4.6	**…a job that allows me to choose how much work to take on?**	1	2	3	4	1028

5 **Now thinking about <u>your interviewing job at NatCen</u>, how satisfied are you with the following aspects of the job?**

◆ Please indicate your answer by circling the appropriate code.

How satisfied are you with…		Very satisfied	Satisfied	Not satisfied	Not satisfied at all	
5.1	**…the amount of pay you receive?**	1	2	3	4	1029
5.2	**…the interesting nature of the job?**	1	2	3	4	1030
5.3	**…working independently?**	1	2	3	4	1031
5.4	**…the interaction you have with people?**	1	2	3	4	1032
5.5	**…the level of flexibility to choose your times and days of work?**	1	2	3	4	1033
5.6	**…the level of flexibility to choose how much work to take on?**	1	2	3	4	1034
5.7	**…the need to work during the 'evening and weekend' hours?**	1	2	3	4	1035

6 **Below follows a series of statements on persuading respondents. To what extent do you agree or disagree with each of these. Interviewers may differ in their opinions about these strategies. There are no right or wrong answers. We are interested in your opinion, based on your experience as an interviewer.**

◆ Please indicate your answer by circling the appropriate code.

		Strongly agree	Agree	Disagree	Strongly disagree	
6.1	**Reluctant respondents should always be persuaded to participate.**	1	2	3	4	1036
6.2	**With enough effort, even the most reluctant respondent can be persuaded to participate.**	1	2	3	4	1037
6.3	**An interviewer should respect the privacy of the respondent.**	1	2	3	4	1038
6.4	**If a respondent is reluctant, a refusal should be accepted.**	1	2	3	4	1039
6.5	**One should always emphasise the voluntary nature of participation.**	1	2	3	4	1040
6.6	**It does not make sense to contact reluctant target persons repeatedly.**	1	2	3	4	1041
6.6	**If you catch them at the right time, most people will agree to participate.**	1	2	3	4	1042
6.7	**Respondents persuaded after great effort do not provide reliable answers.**	1	2	3	4	1043

☐7 **The following questions are about how you see yourself as a person. Please circle the number which best describes how you see yourself where 1 means 'does not apply to me at all' and 7 means 'applies to me perfectly'.**

Please describe yourself as you generally are now, not as you wish to be in the future. Describe yourself as you honestly see yourself, in relation to other people you know of the same sex as you and roughly your age.

I see myself as someone who . . .	1=Does not apply to me at all ◄——► 7=Applies perfectly to me							
								1044
7.1	**…is sometimes rude to others**	1	2	3	4	5	6	7	1045
7.2	**…does a thorough job**	1	2	3	4	5	6	7	1046
7.3	**…is talkative**	1	2	3	4	5	6	7	1047
7.4	**…worries a lot**	1	2	3	4	5	6	7	1048
7.5	**…is original, comes up with new ideas**	1	2	3	4	5	6	7	1049
7.6	**…has a forgiving nature**	1	2	3	4	5	6	7	1050
7.7	**…tends to be lazy**	1	2	3	4	5	6	7	1051
7.8	**…is outgoing and sociable**	1	2	3	4	5	6	7	1052
7.9	**…gets nervous easily**	1	2	3	4	5	6	7	1053
7.10	**…values artistic, aesthetic experiences**	1	2	3	4	5	6	7	1054
7.11	**…is considerate and kind to almost everyone**	1	2	3	4	5	6	7	1055
7.12	**…does things efficiently**	1	2	3	4	5	6	7	SPARE 1056-1060

I see myself as someone who . . .	1=Does not apply to me at all ◄——► 7=Applies perfectly to me							
7.13	**…is reserved**	1	2	3	4	5	6	7	1061
7.14	**…is relaxed and handles stress well**	1	2	3	4	5	6	7	1062
7.15	**…has an active imagination**	1	2	3	4	5	6	7	1063
7.16	**…prefers to stick with things that I know**	1	2	3	4	5	6	7	1064
7.17	**…is very aware of my surroundings**	1	2	3	4	5	6	7	1065
7.18	**…is good at sensing what others are feeling**	1	2	3	4	5	6	7	1066
7.19	**…is interested in many things**	1	2	3	4	5	6	7	1067
7.20	**…anticipates the needs of others**	1	2	3	4	5	6	7	1068
7.21	**…can tell a lot about people from how they live**	1	2	3	4	5	6	7	1069
7.22	**…has trouble guessing how others will react**	1	2	3	4	5	6	7	1070
7.23	**…senses others' wishes**	1	2	3	4	5	6	7	1071
7.24	**…prefers variety to routine**	1	2	3	4	5	6	7	1072

	I see myself as someone who . . .	1=Does not apply to me at all ← → 7=Applies perfectly to me							
7.25	…tends to miss things that other people notice	1	2	3	4	5	6	7	1073
7.26	…knows what to say to make people feel good	1	2	3	4	5	6	7	1074
7.27	…respects the viewpoints of others	1	2	3	4	5	6	7	1075
7.28	…tends to use people's body language to help me understand what they mean	1	2	3	4	5	6	7	1076
7.29	…often tells myself that I'll do an activity tomorrow, even when I have enough time to complete it today	1	2	3	4	5	6	7	1077
7.30	…is always prepared	1	2	3	4	5	6	7	1078
7.31	…avoids difficult reading material	1	2	3	4	5	6	7	1079
7.32	…is always aware of how I present myself	1	2	3	4	5	6	7	1080
7.33	…is never at a loss for words	1	2	3	4	5	6	7	1081
7.34	…remains calm under pressure	1	2	3	4	5	6	7	1082
7.35	…feels that others don't understand what I'm trying to say	1	2	3	4	5	6	7	1083
7.36	…uses body language to help me get my point across	1	2	3	4	5	6	7	1084

SPARE 1085-1090

	I see myself as someone who . . .	1=Does not apply to me at all ← → 7=Applies perfectly to me							
7.37	…sometimes realises that I'm not paying attention when others are speaking to me	1	2	3	4	5	6	7	1091
7.38	…finds it difficult to persuade others	1	2	3	4	5	6	7	1092
7.39	…sticks up for myself	1	2	3	4	5	6	7	1093
7.40	…feels overwhelmed by the amount of things that need to be done	1	2	3	4	5	6	7	1094
7.41	…tends to mislay things	1	2	3	4	5	6	7	1095
7.42	…likes to tackle complex problems	1	2	3	4	5	6	7	1096
7.43	…catches on to things quickly	1	2	3	4	5	6	7	
7.44	…can talk my way out of anything	1	2	3	4	5	6	7	
7.45	…finds it difficult to organise tasks and activities	1	2	3	4	5	6	7	
7.46	…adapts easily to new situations	1	2	3	4	5	6	7	
7.47	…avoids 'small talk'	1	2	3	4	5	6	7	
7.48	…listens to others, even if I disagree	1	2	3	4	5	6	7	

7.49	**…says 'no' to requests from others at times, without feeling guilty**	1	2	3	4	5	6	7
7.50	**…sets goals and priorities for myself**	1	2	3	4	5	6	7
7.51	**…frequently forgets to do things**	1	2	3	4	5	6	7
7.52	**…comes up with new ways to do things**	1	2	3	4	5	6	7
7.53	**…likes to take time making decisions**	1	2	3	4	5	6	7
7.54	**…can't help but look upset when something bad happens**	1	2	3	4	5	6	7
7.55	**…expresses myself easily**	1	2	3	4	5	6	7
7.56	**…can talk others into doing things**	1	2	3	4	5	6	7
7.57	**…lets others make the decisions**	1	2	3	4	5	6	7
7.58	**…gets upset if others change the way that I have arranged things**	1	2	3	4	5	6	7
7.59	**…jumps into things without thinking**	1	2	3	4	5	6	7
7.60	**…pays little attention to my appearance**	1	2	3	4	5	6	7

I see myself as someone who . . .

1=Does not apply to me at all ⟷ 7=Applies perfectly to me

7.61	**…is hard to convince**	1	2	3	4	5	6	7
7.62	**…regularly fills my time with unimportant tasks**	1	2	3	4	5	6	7
7.63	**…can manage many things at the same time**	1	2	3	4	5	6	7
7.64	**…likes to follow standard routines**	1	2	3	4	5	6	7
7.65	**…quickly bounces back from setbacks**	1	2	3	4	5	6	7
7.66	**…tends to finish what I start**	1	2	3	4	5	6	7
7.67	**…is good at explaining things to people**	1	2	3	4	5	6	7

This questionnaire has been sent to all interviewers who have carried out interviewing work for NatCen since 1 January 2006. The following 2 sections are only applicable to those interviewers who have not stopped working for NatCen as an interviewer.

Have you stopped working for NatCen as an interviewer?

◆ Please indicate your answer by circling the appropriate code.

Yes	1	Go to	
No	2	Go to	1122

Apart from your interviewing work (at NatCen and elsewhere), do you <u>currently</u> do any other paid work?

◆ Please indicate your answer by circling the appropriate code.

Yes	1	Go to	
No	2	Go to	1123

Is this other paid work usually at fixed times and on fixed days?

◆ Please indicate your answer by circling the appropriate code.

Yes	1	
No	2	1124

Apart from paid work, do you <u>currently</u> have any other commitments at fixed times and on fixed days?

Other commitments include unpaid work, volunteer work, religious services, caring for children/elderly/disabled relatives, study, classes, sports and other leisure activities.

◆ Please indicate your answer by circling the appropriate code.

Yes	1	
No	2	1125

Did you answer 'yes' at either question 10 or question 11?

◆ Please indicate your answer by circling the appropriate code.

Yes	1	Go to	
No	2	Go to	1126

Please indicate using a tick (9) the fixed times and days on which you are… (a) usually doing other paid work (other than interviewing work) or (b) usually carrying out other commitments.

- Activities outside core interviewing hours (9:00-21:00) do not need to be marked on the grid.
- Include partial hours

	09:00 to 10:00	10:00 to 11:00	11:00 to 12:00	12:00 to 13:00	13:00 to 14:00	14:00 to 15:00	15:00 to 16:00	16:00 to 17:00	17:00 to 18:00	18:00 to 19:00	19:00 to 20:00	20:00 to 21:00	
Monday	1127						1133						1127-1138
Tuesday	1139						1145						1139-1150
Wednesday	1151						1157						1151-1162
Thursday	1163						1169						1163-1174
Friday	1175						1181						1175-1186
Saturday	1187						1193						1187-1198
Sunday	1199						1205						1199-1210

Are there any other times and days that you are usually unwilling to carry out interviewing work for NatCen?

- Please indicate your answer by circling the appropriate code.

Yes	1	Go to →	15
No	2	Go to →	16

1211

15 Please indicate using a tick (9) the fixed times and days on which you are usually unwilling to carry out interviewing work for NatCen.

- Times outside core interviewing hours (9:00-21:00) do not need to be marked on the grid.
- Include partial hours

	09:00 to 10:00	10:00 to 11:00	11:00 to 12:00	12:00 to 13:00	13:00 to 14:00	14:00 to 15:00	15:00 to 16:00	16:00 to 17:00	17:00 to 18:00	18:00 to 19:00	19:00 to 20:00	20:00 to 21:00
Monday	1212						1218				1212-1223	
Tuesday	1224						1230				1224-1235	
Wednesday	1236						1242				1236-1247	
Thursday	1248						1254				1248-1259	
Friday	1260						1266				1260-1271	
Saturday	1272						1278				1272-1283	
Sunday	1284						1290				1284-1295	

16 How supportive do you find the following people when you have a problem or need some advice…

- Please indicate your answer by circling the appropriate code.

		Very supportive	Fairly supportive	Not very supportive	Not at all supportive
16.1	**Your supervisor?**	1	2	3	4 1296
16.2	**Your team leader?**	1	2	3	4 1297
16.3	**The project teams?**	1	2	3	4 1298
16.4	**Your area manager?**	1	2	3	4 1299
16.5	**Your deputy area manager?**	1	2	3	4 1300

All in all, how satisfied are you with your job as a NatCen interviewer?

◆ Please indicate your answer by circling the appropriate code.

Very satisfied	1
Quite satisfied	2
Neither satisfied nor dissatisfied	3
Quite dissatisfied	4
Very dissatisfied	5

1301

If there is anything else you would like to tell us, please write in the box below.

Appendix 2: Factor analyses

1. Personality traits

   Four factored scores for each interviewer were devised using the Principal Component Analysis. Direct Oblimin rotation was used as personality traits are rarely independent. Factor scores were saved using the Anderson-Rubin method. The agreeableness personality trait did not achieve a reliable Cronbach's Alpha coefficient.

   Factors are as follows:

   Openness

Cronbach's Alpha = 0.730

q715bg5o	has an active imagination
q710bg5o	values artistic, aesthetic experiences
q75big5o	is original, comes up with new ideas
q719bg5o	is interested in many things
q724bg5o	prefers variety to routine

Neuroticism

Cronbach's Alpha = 0.755

q74big5n	worries a lot
q79big5n	gets nervous easily
q714bg5n_rev	is relaxed and handles stress well

Conscientiousness

Cronbach's Alpha = 0.703

q7l2bg5c	does things efficiently
q72big5c	does a thorough job
q711bg5a*	is considerate and kind to almost everyone

Extraversion

Cronbach's Alpha = 0.739

q78big5e	is outgoing and sociable
q73big5e	is talkative
q713bg5e_rev	is reserved

2. Other personality traits

Ability to pick up cues, empathy and ability to read other people were all merged into one factor only as follows:

Empathy, understanding others

Cronbach's Alpha = 0.781

q728cue	uses people's body language to help understand what they mean
q720emp	anticipates the needs of others
q723emp	senses others' wishes
q727emp	respects the viewpoints of others
q726read	knows what to say to make people feel good

3. Person skills

There were five characteristics in this section: time management, organisation/attention to detail, problem solving/reasoning, presentation of self, adaptability/flexibility, resilience. In total, seven factors were devised but only three were chosen for a further investigation.

Efficiency

Cronbach's Alpha = 0.702

q765res	quickly bounces back from setbacks
q733flex	is never at a loss for words
q734res	remains calm under pressure
q763org	can manage many things at the same time

Flexibility

Cronbach's Alpha = 0.724

q743flex	catches on to things quickly
q742prob	likes to tackle complex problems
q746flex	adapts easily to new situations
q752prob	comes up with new ways to do things

Attention to detail

Cronbach's Alpha = 0.754

q741org	tends to mislay things
q751org	frequently forgets to do things
q730org	is always prepared
q745org	finds it difficult to organise tasks and activities
q740time	feels overwhelmed by the amount of things that need to be done
q766org	tends to finish what he or she starts
q759org	jumps into things without thinking

4. Interpersonal skills

Communication skills (incl. persuasion)

Cronbach's Alpha =0.763

q756pers	can talk others into doing things
q744comv	can talk his or her way out of anything
q755comv	expresses himself or herself easily
q767comv	is good at explaining things to people
q738pers	finds it difficult to persuade others
q736comn	uses body language to help get his or her point across

Appendix 3: Data acknowledgment

The NTS data including paradata were released by NatCen on two occasions. The initial data release included selected survey variables from seven NTS cohorts between 2002 and 2008. This dataset included:

- selected survey responses at the household level
- selected survey responses at the individual level
- complete Administrative Block section for each interview
- demographic questions responses at the individual level

Survey data were linked by NatCen with MSOA indicators (as described in the data section), supplied directly to NatCen by Ian Brunton-Smith to prevent the disclosure of the actual MSOAs and preserve the anonymity of respondents, i.e. the area where they live. NatCen also supplied dataset containing complete survey responses to the 'Who is behind the NatCen ID card?' questionnaire. The dataset included only NatCen interviewers who worked on the NTS. The dataset was anonymised and included the ID variable that allowed linking interviewers with their respondents in the NTS dataset. Finally, a file containing administrative information on the NTS interviewers such as age, gender and tenure was also supplied. All data in the initial data release were prepared and checked by NatCen researcher Susan Nunn.

The second data release included the Blaise audit trail for the NTS 2011 administration. Each case in the dataset contained field name, time spent in the field (in seconds), whether the help window was brought up, whether the information was refused, whether the 'Don't know' answer was provided and finally the number of visits to the field. The Blaise audit trail was supplied for the household reference person at the household and individual questionnaire levels. The mean number of cases per respondent was 318. The actual survey responses were not provided but instead access was given to the NatCen data enclave and NTS 2011 survey results. Using the actual survey responses satisficing indicators were calculated (as described in Papers Two and Three), which were linked with the Blaise audit trail. Responses to demographic questions were also used to obtain certain respondents characteristics such as gender, age and educational attainment. As in the first data release these were linked with paradata on interviewers and MSOA level variables. Gerry Nicolaas at NatCen conducted the final data linking and checking.

NatCen provided the NTS datasets for 2002-2008 and 2011, including associated paradata, under the agreement on access to commercially sensitive data to third parties. The data obtained from NatCen are for the research purpose of this thesis only and the use of these data will not prejudice the commercial interests of NatCen. The data is kept securely and treated with confidentiality and will be destroyed on the conclusion of the research project. Any outputs from the data satisfy relevant disclosure control policies. The author of this thesis will supply NatCen with any bibliographic details of any published work based on this thesis and data provided by NatCen.

List of References

Anderson, B.A., B.D. Silver, and P. Abramson. 1988. "The Effects of the Race of Interviewer on Race-Related Attitudes of Black Respondents." *Public Opinion Quarterly* 52 (3): 289–324.

Anscombe, F. J., and I. Guttman. 1960. "Rejection of Outliers." *Technometrics* 2 (2): 123–147.

Bailar, B. A. 1976. "Some Sources of Error and Their Effect on Census Statistics." *Demography* 13 (2): 273–286.

Bailar, B. A., and L. Bailey. 1977. "Measures of Interviewer Bias and Variance." *Journal of Marketing Research* 14 (3): 337–343.

Bancroft, G., and E. Welch. 1946. "Recent Experience with Problems of Labour Force Measurement." *Journal of the American Statistical Association* 41 (235): 303–312.

Bar, H., and D. Lillard. 2010. "A Heap of Trouble ? Accounting for Mismatch Bias in Retrospectively Collected Data on Smoking." http://www.york.ac.uk/media/economics/documents/herc/wp/10_20.pdf. (Accessed 14.04.2013)

Barge, S., and H. Gehlbach. 2011. "Using the Theory of Satisficing to Evaluate the Quality of Survey Data." *Research in Higher Education* 53 (2): 182–200.

Barnett, V., and T. Lewis. 1994. *Outliers in Statistical Data*. Chichester: Wiley.

Bassili, J. N. 1996. "The How and the Why of Response Latency Measurement in Telephone Surveys." In *Answering Questions: Methodology for Determining Cognitive and Communicative Process in Survey Research*, edited by N. Schwarz and S. Sudman, 319–346. San Francisco: Jossey-Bass.

Bassili, J. N., and J. F. Fletcher. 1991. "Response-Time Measurement in Survey Research A Method for CATI and a New Look at Nonattitudes." *Public Opinion Quarterly* 55 (3): 331–346.

Bassili, J. N., and B. S. Scott. 1996. "Response Latency as a Signal to Question Problems in Survey Research." *Public Opinion Quarterly* 60 (3): 390–399.

Beaman, J., J. J. Vaske, and C. A. Miller. 2005. "Cognitive Processes in Hunters' Recall of Participation and Harvest Estimates." *Journal of Wildlife Management* 69 (3): 967–975.

Belson, W. A. 1981. *The Design and Understanding of Survey Questions*. Aldershot: Gower Publishing.

Berg, G.M. van der, E. de Feber, and P. de Graaf. 1992. "Analysing Statistical Data Processing." In *New Technologies and Techniques for Statistics*, 194–203. Bonn: Eurostat.

Bethlehem, J. 1997. “Integrated Control Systems for Survey Processing.” In *Survey Measurement and Process Quality*, edited by L. E. Lyberg, P. P. Biemer, M. Collins, E. De Leeuw, C. S. Dippo, N. Schwarz, and D. Trewin, 371–392. New York: John Wiley & Sons.

Biemer, P., and L. E. Lyberg. 2003. *Introduction to Survey Quality*. New Jersey: John Wiley & Sons, Inc.

Biemer, P., and L. Stokes. 1985. “Optimal Design of Interviewer Variance Experiments in Complex Surveys.” *Journal of the American Statistical Association* 80 (389): 158–166.

Blasius, J., and V. Thiessen. 2012. *Assessing the Quality of Survey Data*. SAGE Publications Ltd.

Bliss, C. I., W. G. Cochran, and J. W. Tukey. 1956. “A Rejection Criterion Based Upon the Range.” *Biometrika* 43 (3/4): 418–422.

Blom, A. 2009. “Nonresponse Bias Adjustments: What Can Process Data Contribute?” www.iser.essex.ac.uk/files/iser_working_papers/2009-21.pdf. (Accessed 14.04.2013).

Blom, A., Leeuw E. De, and J. Hox. 2011. “Interviewer Effect on Nonresponse in the European Research Survey.” *Journal of Official Statistics* 27 (2): 359–377.

Browne, W. J. 2009. *MCMC Estimation in MLwiN, Version 2.13*. University of Bristol: Centre for Multilevel Modelling. http://www.bristol.ac.uk/cmm/software/mlwin/download/manuals.html. (Accessed 20.10.2010).

Browning, M., T. F. Crossley, and G. Weber. 2003. “Asking Consumption Questions in General Purpose Surveys.” *The Economic Journal* 113 (491): 540–567.

Brunton-Smith, I. 2008. “Local Areas and Fear of Criminal Victimisation: Applying Multilevel Models to the British Crime Survey”. University of Surrey.

Brunton-Smith, I., and P. Sturgis. 2011. “Do Neighborhoods Generate Fear of Crime? An Empirical Test Using the British Crime Survey.” *Criminology* 49 (2): 331–369.

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

Burton, S., and E. Blair. 1991. “Task Conditions, Response Formulation Processes, and Response Accuracy for Behavioral Frequency Questions in Surveys.” *Public Opinion Quarterly* 55 (1): 50–79.

Cacioppo, J. T., and R. E. Petty. 1982. “The Need for Cognition.” *Journal of Personality and Social Psychology* 42 (1): 116–131.

Callegaro, M., Y. Yang, D. S. Bhola, D. A. Dillman, and T.-Y. Chin. 2009. “Response Latency as an Indicator of Optimizing in Online Questionnaires.” *Bulletin De Méthodologie Sociologique* 103 (1): 5–25.

Camara, W. 2002. “Examinee Behavior and Scoring of CBTs.” In *Computer Based Testing Building the Foundation for Future Assessments*, edited by C. N. Mills, M. T. Potenza, J. J. Fremer, and W. C. Ward, 275–286. New Jersy: Routledge.

Campanelli, P., P. Sturgis, and S. Purdon. 1997. “Can You Hear Me Knocking? An Investigation into the Impact of Interviewers on Survey Response Rates”. London. http://eprints.soton.ac.uk/80198/. (Accessed 14.04.2013).

Catania, J. A., D. Binson, J. Canchola, L. M. Pollack, W. Hauck, and T. J. Coates. 1996. “Effects of Interviewer Gender, Interviewer Choice, and Item Wording on Responses to Questions Concerning Sexual Behavior.” *Public Opinion Quarterly* 60 (3): 345–375.

Chang, L., and J. A. Krosnick. 2010. “Comparing Oral Interviewing with Self-Administered Computerized Questionnaires An Experiment.” *Public Opinion Quarterly* 74 (1): 154–167.

Clarke, P., K. Sproston, and R. Thomas. 2003. “An Investigation into Expectation-led Interviewer Effects in Health Surveys.” *Social Science & Medicine (1982)* 56 (10): 2221–8.

Cleary, P. D., D. Mechanic, and N. Weiss. 1981. “The Effect of Interviewer Characteristics on Responses to a Mental Health Interview.” *Journal of Health and Social Behavior* 22 (2): 183–193.

Cornfield, J. 1942. “On Certain Biases in Samples of Human Populations.” *Journal of the American Statistical Association* 37 (217): 63–68.

Couper, M. 1998. “Measuring Quality in CASIC Environment.” In *Section on Survey Research Methods American Statistical Association*, 41–49. http://www.amstat.org/sections/srms/proceedings/papers/1998_006.pdf. (Accessed 14.04.2013).

———. 2009. “The Role of Paradata in Measuring and Reducing Measurement Error in Surveys.” In *Why Use Paradata in Social Survey Research? NatCen Seminar (London School of Economics)*. London. http://www.natcen.ac.uk/ncrm-paradata-network#info. (Accessed 25.10.2010).

Couper, M., and F. Kreuter. 2010. “Using Item-Level Paradata to Explore Response Time and Data Quality.” In *Recent Advances in the Use of Paradata (Process Data) in Social Survey Research (Royal Statistical Society)*. London. http://www.southampton.ac.uk/s3ri/events/workshops/Files/Couper&Kreuter 2010-11-13.pdf. (Accessed 14.04.2013).

———. 2013. “Using Paradata to Explore Item Level Response Times in Surveys.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176 (1): 271–286.

Davis, P., and A. Scott. 1995. “The Effect of Interviewer Variance on Domain Comparisons.” *Survey Methodology* 21 (2): 99–106.

DeMaio, T. J. 1984. “Social Desirability and Survey Measurement: A Review.” In *Surveying Subjective Phenomena*, edited by C. F. Turner and E. Martin, 2:257–282. New York: Russell Sage.

Deming, W. 1944. “On Errors in Surveys.” *American Sociological Review* 9 (4): 359–369.

Denic, S., F. Khatib, and H. Saadi. 2004. “Quality of Age Data in Patients from Developing Countries.” *Journal of Public Health Oxford England* 26 (2): 168–171.

Dijkstra, W. 1987. “Interviewing Style and Respondent Behavior: An Experimental Study of the Survey-Interview.” *Sociological Methods & Research* 16 (2): 309–334.

Dippo, C. S. 1997. “Survey Measurement and Process Improvement.” In *Survey Measurement and Process Quality*, edited by L. E. Lyberg, P. P. Biemer, M. Collins, E. De Leeuw, C. S. Dippo, N. Schwarz, and D. Trewin, 475–475. New York: John Wiley and Sons Inc.

Dippo, C. S., and B. Sundgren. 2000. “The Role of Metadata in Statistics.” In *International Conference on Establishment Surveys II*. Buffalo, New York. http://www.amstat.org/meetings/ices/2000/proceedings/S57.pdf. (Accessed 14.04.2013)

Draisma, S., and W. Dijkstra. 2004. “Response Latency and (Para)Linguistic Expressions as Indicators of Response Error.” In *Methods for Testing and Evaluating Survey Questionnaires*, edited by S. Presser, J. M. Roghtgeb, M. Couper, J. T. Lessler, E. Martin, J. Martin, and E. Singer, 131–147. Hoboken, New Jersey: John Wiley & Sons.

Durrant, G. B. 2009. “Using Paradata to Predict Best Times of Contact Conditioning on Household and Interviewer Influences.” In *Why Use Paradata in Social Survey Research? NatCen Seminar (London School of Economics)*. London. http://www.natcen.ac.uk/ncrm-paradata-network#info. (Accessed 20.10.2010).

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

Fazio, R. H. 1990. “Multiple Processes by Which Attitudes Guide Behavior: The Mode Model as an Integrative Framework.” *Advances in Experimental Social Psychology* 23: 75–109.

Fellegi, I. P. 1964. “Response Variance and Its Estimation.” *Journal of the American Statistical Association* 59: 1016–1041.

———. 1974. “An Improved Method of Estimating the Correlated Response Variance.” *Journal of the American Statistical Association* 69 (346): 496–501.

Fowler, F. J. 1992. “How Unclear Terms Affect Survey Data.” *Public Opinion Quarterly* 56 (2): 218–231.

Fowler, F. J., and T. Mangione. 1990. *Standardized Survey Interviewing. Minimizing Interviewer-Related Error*. Newbury Park: SAGE Publications.

Fricker, S., M. Galesic, R. Tourangeau, and T. Yan. 2005. “An Experimental Comparison of Web and Telephone Surveys.” *Public Opinion Quarterly* 69 (3): 370–392.

Fuchs, M. 2009. “Asking for Numbers and Quantities: Visual Design Effects in Paper&Pencil Surveys.” *International Journal of Public Opinion Research* 21 (1): 65–84.

Gabler, S., S. Häder, and P. Lynn. 2006. “Design Effect for Multiple Design Samples.” *Survey Methodology* 32 (1): 115–120.

Goldberg, L. R. 1990. “An Alternative ‘Description of Personality’: The Big-five Factor Structure.” *Journal of Personality and Social Psychology* 59 (6): 1216–1229.

Goldstein, H., W. J. Browne, and J. Rasbash. 2002. “Partitioning Variation in Multilevel Models.” *Understanding Statistics* 1 (4): 223–231.

Green, M. C., J. A. Krosnick, and A. L. Holbrook. 2001. “The Survey Response Process in Telephone and Face-to-Face Surveys: Differences in Respondent Satisficing and Social Desirability Response Bias.” http://www.clas.ufl.edu/users/kenwald/pos6757/spring02/tch62.pdf. (Accessed 14.04.2013).

Groves, R. M. 2004. *Survey Errors and Survey Costs*. New Jersey: Wiley.

Groves, R. M., R. Cialdini, and M. Couper. 1992. “Understanding the Decision to Participate in a Survey.” *Public Opinion Quarterly* 56: 475–495.

Groves, R. M., and M. Couper. 1998. *Nonresponse in Household Interview Surveys*. *Wiley Series in Probability and Statistics. Survey Methodology Section*. New York ; Chichester: Wiley.

Groves, R. M., F. J. Jr. Fowler, M. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau. 2009. *Survey Methodology*. *Wiley Series in Survey Methodology*. 2nd ed. Hoboken, New Jersey: John Wiley & Sons, Inc.

Groves, R. M., and S. Heeringa. 2006. “Responsive Design for Household Surveys: Tool for Actively Controlling Survey Errors and Costs.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169 (3): 439–457.

Groves, R. M., and L. E. Lyberg. 2010. “Total Survey Error. Past, Present and Future.” *Public Opinion Quarterly* 74 (5): 849–879.

Grubbs, F. E. 1950. “Sample Criteria for Testing Outlying Observations.” *The Annals of Mathematical Statistics* 21 (1): 27–58.

Hansen, M. H., W. N. Hurwitz, and M. A. Bershad. 1961. “Measurement Errors in Censuses and Surveys.” *Bulletin of the International Statistical Institute* 38: 359–374.

Hansen, M. H., W. N. Hurwitz, E. S. Marks, and W. P. Mauldi. 1951. “Response Errors in Surveys.” *Journal of the American Statistical Association* 46 (254): 147–190.

Hartley, H., and J. Rao. 1978. “The Estimation of Nonsampling Variance Components in Sample Surveys.” In *Survey Sampling and Measurement*edited by N.K. Namboodri,. New York: Academic Press.

Hawkins, D. M. 1980. *Identification of Outliers*. London; New York: Chapman and Hall.

Heeringa, S., and J. Liu. 1998. “Complex Sample Design Effects and Inference in Mental Health Survey Data.” *International Journal of Methods in Psychiatric Research* 7 (1): 56–65.

Heerwegh, D. 2003. “Explaining Response Latencies and Changing Answers Using Client-Side Paradata from a Web Survey.” *Social Science Computer Review* 21 (3): 360–373.

Heerwegh, D., and G. Loosveldt. 2008. “Face-to-Face Versus Web Surveying in a High-Internet-Coverage Population: Differences in Response Quality.” *Public Opinion Quarterly* 72 (5): 836–846.

Heitjan, D. F., and D. B. Rubin. 1991. “Ignorability and Coarse Data.” *Annals of Statistics* 19 (4): 2244–2253.

Hoaglin, D. C., and B. Iglewicz. 1987. “Fine Tuning Some Resistant Rules for Outlier Labeling.” *Journal of American Statistical Association* 82 (400): 1147–1149.

Hoaglin, D. C., B. Iglewicz, and J. W. Tukey. 1986. “Performance of Some Resistant Rules for Outlier Labeling.” *Journal of the American Statistical Association* 81 (396): 991–999.

Holbrook, A. L., M. C. Green, and J. A. Krosnick. 2003. “Telephone Versus Face-to-Face Interviewing of National Probability Samples with Long Questionnaires : Comparisons of Respondent Satisficing and Social Desirability Response Bias.” *Public Opinion Quarterly* 67 (1): 79–125.

Hox, J. 1994. “Hierarchical Regression Models for Interviewer and Respondent Effects.” *Sociological Methods and Research* 22 (3): 300–318.

———. 2002. *Multilevel Analysis.Techniques and Applications*. *Quantitative Methodology Series*. London: Lawrence Erlbaum Associates.

Hultsman, W. Z., J. T. Hultsman, and D. R. Black. 1989. “‘Response Peaks’ as a Component of Measurement Error: Assessment Implications for Self-reported Data in Leisure Research.” *Journal of Leisure Research* 10 (21): 310–315.

Huttenlocher, J., L. V. Hedges, and N. M. Bradburn. 1990. “Reports of Elapsed Time: Bounding and Rounding Processes in Estimation.” *Journal of Experimental Psychology Learning Memory and Cognition* 16 (2): 196–213.

Jäckle, A., P. Lynn, J. Sinibaldi, and S. Tipping. 2011. “The Effect of Interviewer Personality , Skills and Attitudes on Respondent Co-operation with Face-to-face Surveys.” https://www.iser.essex.ac.uk/publications/working-papers/iser/2011-14. (Accessed 13.04.2013).

Jäckle, A., C. Roberts, and P. Lynn. 2010. “Assessing the Effect of Data Collection Mode on Measurement.” *International Statistical Review* 78 (1): 3–20.

Johnson, R., and Y. Price. 1988. “The Relationship of the Jackson Personality Traits of US Census Bureau Current Survey Interviewers to Work Performance and Turnover”. DC.

Kaminska, O., A. L. McCutcheon, and J. Billiet. 2011. “Satisficing Among Reluctant Respondents in a Cross-National Context.” *Public Opinion Quarterly* 74 (5): 956–984.

Kish, L. 1962. “Studies of Interviewer Variance for Attitudinal Variables.” *Journal of the American Statistical Association* 57 (297): 92–115.

———. 1965. *Survey Sampling*. New York: John Wiley & Sons.

Knowles, E. S., and C. A. Condon. 1999. “Why People Say ‘ Yes ’: A Dual-Process Theory Of Acquiescence.” *Journal of Personality and Social Psychology* 77 (2): 379–386.

Koch, G. 1973. “An Alternative Approach to Multivariate Response Error Models for Sample Survey Data with Applications of Estimators Involving Subclass Means.” *Journal of the American Statistical Association* 68 (344): 906–913.

Kreuter, F., and C. Casas-Cordero. 2010. “Paradata.” *German Council for Social and Economic Data (RatSWD)*. http://www.ratswd.de/download/RatSWD_WP_2010/RatSWD_WP_136.pdf. (Accessed 13.04.2013).

Kreuter, F., M. Lemay, and C. Casas-Cordero. 2007. “Using Proxy Measures of Survey Outcomes in Post-Survey Adjustments: Examples from the European Social Survey (ESS).” In *Section on Survey Research Methods American Statistical Association*. http://www.amstat.org/sections/srms/proceedings/y2007/Files/JSM2007-000527.pdf. (Accessed 14.04.2013).

Kreuter, F., K. Olson, J. Wagner, J. Yan, T. Ezzati-Rice, C. Casas-Cordero, C. Lemay, M. Peytchev, R. M. Groves, and T. Raghunathan. 2010. “Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Nonresponse: Examples from Multiple Surveys.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 173 (2): 389–407.

Kreuter, F., S. Presser, and R. Tourangeau. 2009. “Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity.” *Public Opinion Quarterly* 72 (5): 847–865.

Krosnick, J. A. 1991. “Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys.” *Applied Cognitive Psychology* 5 (3): 213–236.

Krosnick, J. A., and D. F. Alwin. 1987. “An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement.” Edited by Norbert Schwarz and Seymour Sudman. *Public Opinion Quarterly* 51 (2): 201–219.

Krosnick, J. A., A. L. Holbrook, M. K. Berent, R. T. Carson, W. M. Hanemann, R. J. Kopp, R. C. Mitchell, et al. 2002. “The Impact of ‘No Opinion’ Response Options on Data Quality: Non-Attitude Reduction or an Invitation to Satisfice?” *Public Opinion Quarterly* 66 (3): 371–403.

Krosnick, J. A., S. Nuruyan, and W. R. Smith. 1996. “Satisficing in Surveys : Initial Evidence.” *New Directions for Evaluation* 1996 (70): 29–44.

Laflamme, F., M. Maydan, and A. Miller. 2008. “Using Paradata to Actively Manage Data Collection.” In *Joint Statistical Meetings, American Statistical Association, Section on*

Survey Research Methods, 630–637. Denver. http://www.amstat.org/sections/srms/proceedings/y2008/Files/300608.pdf. (Accessed 14.04.2013).

Lee, S. 2009. "Responsive Design of Random Digit Dial Surveys Using Paradata." In *Why Use Paradata in Social Survey Research? NatCen Seminar (London School of Economics)*. London. http://www.natcen.ac.uk/ncrm-paradata-network#info. (Accessed 12.11.2010).

Lelkes, Y., J. A. Krosnick, D. M. Marx, C. M. Judd, and B. Park. 2012. "Complete Anonymity Compromises the Accuracy of Self-Reports." *Journal of Experimental Social Psychology* 48 (6): 1291–1299.

Lipps, O. 2007. "Interviewer and Respondent Survey Quality Effects in a CATI Panel." *Bulletin De Méthodologie Sociologique* 95 (1): 5–25.

Locker, T. E., and S. M. Mason. 2006. "Digit Preference Bias in the Recording of Emergency Department Times." *European Journal of Emergency Medicine Official Journal of the European Society for Emergency Medicine* 13 (2): 99–101.

Lyberg, L. E. 2009. "The Paradata Concept in Survey Research." In *Why Use Paradata in Social Survey Research? NatCen Seminar (London School of Economics)*. London. http://www.natcen.ac.uk/ncrm-paradata-network#info. (Accessed 12.11.2010).

Lynn, P. 2003. "PEDAKSI: Methodology for Collecting Data About Survey Non-respondents." *Quality & Quantity* 37 (4): 239–261.

Lynn, P., and S. Gabler. 2005. "Approximation of Beta in the Prediction of Design Effects Due to Clustering." *Survey Methodology* 31 (1): 101–104.

Mahalanobis, P. 1946. "Recent Experiments in Statistical Sampling in the Indian Statistical Institute." *Journal of the Royal Statistical Society* 109: 325–370.

Martin, D. J. 2001. "Geography for the 2001 Census in England and Wales". London. www.statistics.gov.uk/geography/downloads/georoadshowpaper.pdf. (Accessed 14.04.2013).

Matsuo, H., and G. Loosveldt. 2011. "Usefulness of Paradata in Cross-national Surveys: Evaluation of Dwelling and Neighborhood Characteristics." In *4th Conference of the European Survey Research Association (ESRA)*. Lausanne.

McCulloch, S., F. Kreuter, and S. Calvano. 2010. "Interviewer Observed Vs. Reported Respondent Gender: Implications on Measurement Error." In *Annual Conference of the American Association of Public Opinion and Research (AAPOR)*. Chicago.

Meter, K. M. van. 2005. "Studying Survey Interviewers: A Call for Research and an Overview." *Bulletin De Méthodologie Sociologique* 88 (1): 51–71.

Moore, J. C., L. L. Stinson, and E. J. Jr. Welniak. 1999. "Income Reporting in Surveys: Cognitive Issues and Measurement Error." In *Cognition and Survey Research*, edited by G. M. Sirken, J. D. Herrmann, S. Schechter, N. Schwartz, M. J. Tanur, and R. Tourangeau, 155–174. New York: John Wiley & Sons, Inc.

Morton-Williams, J. 1993. *Interviewer Approaches*. Dartmouth: Dartmouth Publishing Company Limited.

Mulligan, K., G. J. Tobin, S. T. Mockabee, and J. Q. Monson. 2003. "Response Latency Methodology for Survey Research: Measurement and Modeling Strategies." *Political Analysis* 11 (3): 289–301.

Narayan, S., and J. A. Krosnick. 1996. "Education Moderates Some Response Effects in Attitude Measurement." *Public Opinion Quarterly* 60 (1): 58.

Neyman, J. 1934. "On the Two Different Aspects of the Representative Method: The Methods of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society* 97 (4): 558–606.

O'Muircheartaigh, C. 1997. "Measurement Error in Surveys: A Historical Perspective." In *Survey Measurement and Process Quality*, edited by L. E. Lyberg, P. P. Biemer, M. Collins, E. De Leeuw, C. S. Dippo, N. Schwarz, and D. Trewin, 1–25. New York: John Wiley and Sons.

O'Muircheartaigh, C., and P. Campanelli. 1998. "The Relative Impact of Interviewer Effects and Sample Design Effects on Survey Precision." *Journal of the Royal Statistical Society Series A Statistics in Society* 1 (161): 63–77.

Olson, K. 2009. "Paradata for Dynamic Response Propensities and Nonresponse Adjustment." In *Why Use Paradata in Social Survey Research? NatCen Seminar (London School of Economics)*. London. http://www.natcen.ac.uk/ncrm-paradata-network#info. (Accessed 11.10.2010).

Olson, K., and I. Bilgen. 2011. "The Role Of Interviewer Experience on Acquiescence." *Public Opinion Quarterly* 75 (1): 99–114.

Olson, K., and R. M. Groves. 2009. "The Lifecycle of Response Propensities in Fertility and Family Demography Surveys." In *Annual Meeting of the Population Association of America*. Detroit. http://paa2009.princeton.edu/papers/90840. (Accessed 12.04.2013).

Olson, K., and A. Peytchev. 2007. "Effect of Interviewer Experience on Interview Pace and Interviewer Attitudes." *Public Opinion Quarterly* 71 (2): 273–286.

Oppenheimer, D. M., T. Meyvis, and N. Davidenko. 2009. "Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power." *Journal of Experimental Social Psychology* 45 (4): 867–872.

Pickering, R. M. 1992. "Digit Preference in Estimated Gestational Age." *Statistics in Medicine* 11 (9): 1225–1238.

Pickery, J., and G. Loosveldt. 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.

Pickery, J., G. Loosveldt, and A. Carton. 2001. "The Effects of Interviewer and Respondent Characteristics on Response Bahavior in Panel Surveys: A Multilevel Approach." *Sociological Methods and Research* 29 (4): 509–523.

Pudney, S. 2008. “Heaping and Leaping : Survey Response Behaviour and the Dynamics of Self- Reported Consumption Expenditure.” https://www.iser.essex.ac.uk/publications/working-papers/iser/2008-09. (Accessed 05.05.2012).

Raper, J. F., D. W. Rhind, and J. W. Shepherd. 1992. *Postcodes: The New Geography*. London: Longman Scientific & Technical.

Rasbash, J., and W. J. Browne. 2008. “Non-Hierarchical Multilevel Models.” In *Handbook of Multilevel Analysis*, edited by J. de Leeuw and E. Meijer, 301–334. New York: Springer.

Rasbash, J., and H. Goldstein. 1994. “Efficient Analysis of Mixed Herarchical and Cross-Classified Random Structures Using a Multilevel Model.” *Journal of Educational and Behavioral Statistics* 19 (4): 337–350.

Rasbash, J., F. Steele, W. J. Browne, and H. Goldstein. 2009. *A User’s Guide to MLwiN. Version 2.10*. University of Bristol: Centre for Multilevel Modelling. http://www.bristol.ac.uk/cmm/software/mlwin/download/manuals.html.(Accessed 01.09.2010).

Reinecke, J., and P. Schmidt. 1993. “Explaining Interviewer Effects and Respondent Behavior: Theoretical Models and Empirical Analysis.” *Quality and Quantity* 27 (3): 219–247.

Rice, S. 1929. “Contagious Bias in the Interview.” *American Journal of Sociology* 35 (3): 420–423.

Roberts, C. 2007. “Mixing Modes of Data Collection in Surveys : A Methodological Review ESRC National Centre for Research Methods.” http://eprints.ncrm.ac.uk/418/.(Accessed 14.01.2012).

Roberts, J., and D. Brewer. 2001. “Measures and Tests of Heaping in Discrete Quantitative Distributions.” *Journal of Applied Statistics* 28 (7): 887–896.

Robinson, D., and S. Rohde. 1946. “Two Experiments with an Anti-Semitism Poll.” *Journal of Abnormal Social Psychology* 41 (2): 136–144.

Scheuren, F. 2001. “Macro and Micro Paradata for Survey Assessment.” In *United Nations Working Sessions. Statistical Metadata*. Washington DC. http://www.unece.org/fileadmin/DAM/stats/documents/2000/11/metis/crp.10.e.pdf. (Accessed 14.04.2013).

———. 2005. “The Role of Paradata at All Stages of Survey: Going from Concept to Completion.” In *Statistics Canada International Symposium on Methodological Issues*. http://www5.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=11-522-X20050019432&lang=eng. (Accessed 14.04.2013).

Schnell, R., and F. Kreuter. 2005. “Separating Interviewer and Sampling-Point Effects.” *Journal of Official Statistics* 21 (3): 389–410.

Schnipke, D. L., and D. J. Scrams. 2002. “Exploring Issues of Examinee Behavior: Insights Gained from Response-time Analyses.” In *Computer-Based Testing: Building the*

Foundation for Future Assessments, edited by C. N. Mills, M. T. Potenza, J. J. Fremer, and W. C. Ward, 237–266. New Jersey: Lawrence Erlbaum Associates.

Schuman, H., and J. M. Converse. 1981. "The Effect of Black and White Interviewers on Black Responses in 1968." *Public Opinion Quarterly* 35 (1): 44–68.

Schwartz, B., A. Ward, J. Monterosso, S. Lyubomirsky, K. White, and D. R. Lehman. 2002. "Maximising Versus Satisficing: Happiness Is a Matter of Choice." *Journal of Personality and Social Psychology* 83 (5): 1178–1197.

Simon, H. 1957. *Models of Man*. New York: Wiley.

Singer, E., M. R. Frankel, and M. B. Glassman. 1983. "The Effect of Interviewer Characteristics and Expectations on Response." *Public Opinion Quarterly* 47 (1): 68–83.

Singer, E., and L. Kohnke-Aguirre. 1979. "Interviewer Expectation Effects : A Replication and Extension." *Public Opinion Quarterly* 43 (2): 245–260.

Sinibaldi, J. 2009. "Do Interviewer Characteristics Predict Their Contacting Behaviour and Success? An Analysis Using Paradata from Call Records and an Interviewer Survey." In *Why Use Paradata in Social Survey Research? NatCen Seminar (London School of Economics)*. London. http://www.natcen.ac.uk/ncrm-paradata-network#info. (Accessed 14.09.2010).

———. 2010. "Measurement Error in Objective and Subjective Interviewer Observations." In *Annual Conference of the American Association of Public Opinion and Research (AAPOR)*. Chicago. http://www.aapor.org/AM/Template.cfm?Section=65th_Annual_Conference&Template=/CM/ContentDisplay.cfm&ContentID=2385. (Accessed 14.04.2013).

Sinibaldi, J., C. Casas-Cordero, S. Eckman, S. McCulloch, F. Kreuter, and B. West. 2010. "Assessment of the Quality of Interviewer Observations." In *Recent Advances in the Use of Paradata (Process Data) in Social Survey Research (Royal Statistical Society)*. London. http://www.natcen.ac.uk/ncrm-paradata-network#info.(Accessed 14.09.2010).

Sirken, G. M., and S. Schechter. 1999. "Interdisciplinary Survey Methods Research." In *Cognition and Survey Research*, edited by G. M. Sirken, J. D. Herrmann, S. Schechter, N. Schwarz, M. J. Tanur, and R. Tourangeau, 1–10. New York: John Wiley & Sons, Inc.

Skinner, C., and M. Vieira. 2005. "Design Effects in the Analysis of Longitudinal Survey Data." *Southampton Statistical Sciences Research Institute Methodology Working Papers*. http://eprints.soton.ac.uk/15012/1/15012-01.pdf. (Accessed 12.04.2013).

Smit, H. J., W. Dijkstra, and J. van der Zouwen. 1997. "Suggestive Interviewer Behaviour in Surveys: An Experimental Study." *Journal of Official Statistics* 13 (1): 19–28.

Snijders, T, and R Bosker. 2004. *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modelling.* London: SAGE Publications.

Spiegelhalter, D., N. Best, B. Carlin, and A. van der Linde. 2002. "Bayesian Measures of Model Complexity and Fit (with Discussion)." *Journal of the Royal Statistical Society Series B* 64 (4): 583–640.

Sturgis, P. 2004. "Analysing Complex Survey Data: Clustering, Stratification and Weights." *Social Research Update*. http://sru.soc.surrey.ac.uk/SRU43.pdf. (Accessed 13.04.2013).

Sturgis, P., and P. Campanelli. 1998. "The Scope of Reducing Refusals in Household Surveys: An Investigation Based on Transcripts of Tape-Recorded Doorstep Interactions." *Journal of the Market Research Society* 40 (2): 121–139.

Sudman, S., and N. M. Bradburn. 1974. *Response Effects in Surveys*. Chicago: Aldine Publishing Company.

Sudman, S., N. M. Bradburn, E. Blair, and C. Stocking. 1977. "Modest Expectations: The Effects of Interviewers' Prior Expectations on Responses." *Sociological Methods & Research* 6 (2): 171–182.

Sukhatme, P., and B. Sukhatme. 1970. *Sampling Theory of Surveys with Application*. 2nd ed. Ames, IA: Iowa State University Press.

Summers, G. F., and E.M. Beck. 1973. "Social Status and Personality Factors in Predicting Interviewer Performance." *Sociological Methods & Research* 2 (1): 111–122.

Taylor, B. L. 2008. "The 2006 National Health Interview Survey (NHIS) Paradata File: Overview and Applications." In *American Statistical Association, Proceedings of the Section on Survey Research Methods*. http://www.amstat.org/sections/srms/proceedings/y2008/Files/301266.pdf. (Accessed 12.04.2013).

Taylor, E., A. Humphrey, K. Pickering, and S. Tipping. 2012. "National Travel Survey 2011 Technical Report." https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/36548/nts2011-technical.pdf. (Accessed 13.12.2012).

Tobler, W. 1970. "A Computer Movie Simulating Urban Growth in the Detroit Region." *Economic Geography* 46 (2): 234–240.

Tourangeau, R., L. J. Rips, and K. A. Rasinski. 2000. *The Psychology of Survey Response*. Cambridge, UK: Cambridge University Press.

Turner, M., P. Sturgis, and D. J. Martin. 2012. "Survey Satisficing: How Interviewer Personality Traits, Attitudes and Response Styles Influence Data Quality." In *67th Annual Conference of the American Association of the Public Opinion Research*. Orlando, Florida.

Wang, H., and D. F. Heitjan. 2008. "Modeling Heaping in Self-reported Cigarette Counts." *Statistics in Medicine* 27 (19): 3789–804.

Wang, K., J. Murphy, R. Baxter, and J. Aldworth. 2005. “Are Two Feet in the Door Better Than One? Using Process Data to Examine Interviewer Effort and Nonresponse Bias.” In *Federal Committee on Statistical Methodology Research Conference*. http://www.fcsm.gov/05papers/Wang_Aldworth_etal_VIB.pdf. (Accessed 14.04.2013).

West, B. T. 2010. “An Examination of the Quality and Utility of Interviewer Estimates of Household Characteristics in the National Survey of Family Growth.” In *Annual Conference of the American Association of Public Opinion and Research (AAPOR)*. Chicago.

Wise, S. L., D. S. Bhola, and S. Yang. 2006. “Taking the Time to Improve the Validity of Low-Stakes Tests: The Effort-Monitoring CBT.” *Educational Measurement Issues and Practice* 25 (2): 21–30.

Wise, S. L., and X. Kong. 2005. “Response Time Effort: A New Measure of Examinee Motivation in Computer-based Tests.” *Applied Measurement in Education* 18 (2): 163–183.

Yan, T., and R. Tourangeau. 2008. “Fast Times and Easy Questions: The Effects of Age, Experience and Question Complexity on Web Survey Response Times.” *Applied Cognitive Psychology* 22 (1): 51–68.

Zelin, A., and R. Stubbs. 2005. “Cluster Sampling: A False Economy?” *International Journal of Market Research* 47 (5): 503–524.

Zouwen, J. van der., and W. Dijkstra. 1988. “Types of Inadequate Interviewer Behaviour in Survey Interviews: Their Causes and Effects.” *Bulletin the Méthodologie Sociologique* 18 (1): 5–20.