

**The Impact of the Interviewer:
Nonresponse and Response Variance
in Social Surveys**

Pamela Comber Campanelli

**Department of Statistics
London School of Economics and Political Science**

Submitted for examination in November of 1998 for the degree,

Doctor of Philosophy

1999

UMI Number: U615576

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI U615576

Published by ProQuest LLC 2014. Copyright in the Dissertation held by the Author.
Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against
unauthorized copying under Title 17, United States Code.



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

THESES

F

7667



711746

ABSTRACT

Interviewer-based data collection is the norm for social and market research surveys in the United Kingdom and is likely to remain so for the foreseeable future. But what impact do interviewers have on survey results? Interviewers are often seen as valuable allies in the data collection process for their role in minimising many potential sources of survey error (e.g., through motivating the respondent and controlling the response process). Yet, at the same time, interviewers can also be one of the principal causes of nonresponse error and response variance in quantitative surveys. In terms of nonresponse, it is widely known that different interviewers have different response rates, but comparatively little is known about the extent to which this is due to differences among interviewers and their characteristics as opposed to differences among the respondents in those areas allocated to the interviewers. What is also unclear is the extent to which such interviewer differences persist over time in longitudinal surveys, the extent to which interviewers differ in their effectiveness at reducing the refusal as opposed to the non-contact component of nonresponse, and the extent to which the fieldwork strategy of matching the same interviewers to the same respondents, ‘interviewer continuity’, is useful for raising response rates in longitudinal surveys. In terms of response variance, it is rare to find studies in which both the complex sampling variance and the complex interviewer variance are both computed or in which the effects of interviewer continuity on response quality is examined. This thesis investigates these issues by using the interpenetrated sample design experiment designed by C. O’Muircheartaigh (my supervisor) and myself for implementation in Wave 2 of the British Household Panel Study. The analysis is facilitated by the use of cross-classified multilevel modelling in addition to other techniques. The thesis also looks at the issue of ‘interviewer continuity’ qualitatively, through the impressions of the interviewers themselves.

TABLE OF CONTENTS

	Page No.
ACKNOWLEDGEMENTS	16
EXECUTIVE SUMMARY	17
PART 1 BACKGROUND	29
CHAPTER 1 INTRODUCTION	29
1.1 Total Survey Error	29
1.2 Measuring Interviewer Effects	35
1.2.1 <i>The ANOVA Model</i>	36
1.2.2 <i>The Correlation Model</i>	39
1.2.3 <i>Simple and Correlated Response Variance</i>	42
1.2.4 <i>Further Extensions to the Correlation Model</i>	44
1.2.5 <i>Typical Values of ρ, the Intra-Interviewer Correlation Coefficient</i>	45
1.3 The Data Source	48
1.3.1 <i>Description of the British Household Panel Study</i>	48
1.3.2 <i>Description of the Interpenetrating Sample Experiment at Wave 2</i>	49
1.3.3 <i>Implications for Wave 3</i>	51
1.3.4 <i>Auxiliary Data</i>	53
1.4 Analysis Considerations for British Household Panel Study Data	54
1.4.1 <i>Two-way and Three-way Tables</i>	54
1.4.2 <i>Calculation of ρ with Hierarchical Analyses</i>	56

	<i>of Variance</i>	
1.4.3	<i>Logistic Regression Models for Data Reduction</i>	58
1.4.4	<i>Assumptions Behind the Logistic Regression Model</i>	60
1.4.5	<i>Cross-classified Multilevel Models</i>	62
1.4.6	<i>Implementation of the Cross-classified Multilevel Models</i>	65
1.4.7	<i>Calculation of ρ in the Cross-classified Multilevel Models</i>	67
1.4.8	<i>Technical Aspects of the Cross-classified Multilevel Models</i>	67
1.4.9	<i>Availability of British Household Panel Study Data and Multilevel Software</i>	70
1.4.10	<i>Options to Replicate Fellegi (1964)</i>	71
1.4.11	<i>The Use of the Interpenetrated Sub-Sample</i>	72
1.5	Specific Aims of the Thesis	73
1.6	Chapter Summary	75
CHAPTER 2	SURVEY RESEARCH INTERVIEWERS	76
2.1	The Role of the Interviewer	76
2.1.1	<i>A Neutral Collector of Accurate Data</i>	77
2.1.2	<i>Educating the Respondent as to His/Her Role</i>	81
2.1.3	<i>Conducting the Last Stages of Sampling</i>	82
2.1.4	<i>Fulfilling Administrative Duties</i>	83
2.1.5	<i>Gaining the Co-operation of the Respondent</i>	83
2.2	Interviewer Effects	88
2.2.1	<i>Background Characteristics</i>	90

2.2.2	<i>Psychological Factors</i>	93
2.2.3	<i>Behavioural Factors</i>	97
2.3	Types of Questions Prone to Interviewer Effects	100
2.3.1	<i>Effects Based on Question Content</i>	100
2.3.2	<i>Effects Based on Question Format</i>	101
2.3.3	<i>Interviewer Effects on Self-Completion Documents</i>	103
2.4	Chapter Summary	104
CHAPTER 3	EXPLORING NONRESPONSE ERROR	107
3.1	Definitions of Nonresponse	107
3.1.1	<i>Response Rate Calculation</i>	108
3.2	Nonresponse Bias	112
3.3	Nonresponse Variance	114
3.4	Theoretical Perspectives	116
3.4.1	<i>Social Context Variation</i>	116
3.4.2	<i>Respondent Level Variation</i>	117
3.4.3	<i>Interviewer Level Variation</i>	119
3.4.4	<i>An Integrated Approach</i>	120
3.5	Nonresponse Trends	121
3.6	Chapter Summary	122
CHAPTER 4	EXPLORING RESPONSE ERROR	126
4.1	Background	126
4.2	The Conceptual Perspective	127
4.2.1	<i>Interviewers, Respondents and the Survey Task</i>	127
4.2.2	<i>Interviewers and the Mode of Data Collection</i>	138

4.3	Measurement of Response Variance	140
	4.3.1 <i>Mathematical Treatment</i>	141
	4.3.2 <i>The Effects of Response Error on Estimates</i>	146
4.4	Chapter Summary	148
PART 2	EMPIRICAL INVESTIGATIONS INTO	150
	INTERVIEWERS AND NONRESPONSE	
	ERROR	
CHAPTER 5	CORRELATES OF NONRESPONSE	150
5.1	Background	150
	5.1.1 <i>Characteristics of Households and Individuals</i>	150
	5.1.2 <i>The Data: Moving Beyond Characteristics of</i> <i>Households and Individuals</i>	151
5.2	Returning to the Theoretical Model of Groves, Cialdini and Couper	153
5.3	Methods	158
	5.3.1 <i>Indicators of Survey Nonresponse</i>	158
	5.3.2 <i>The Explanatory Variables</i>	160
	5.3.3 <i>Overview of the Analysis Process</i>	160
	5.3.4 <i>Considerations with Respect to the Multiple</i> <i>Logistic Regression Models</i>	160
5.4	Results	166
	5.4.1 <i>Bivariate Analyses</i>	166
	5.4.2 <i>Multiple Logistic Regression Analyses</i>	176
	5.4.3 <i>Cross-Classified Multilevel Models</i>	185

5.5	Chapter Summary and Discussion	185
5.5.1	<i>Summary of the Findings about the Correlates of Nonresponse</i>	185
5.5.2	<i>Attrition Versus Initial Nonresponse</i>	190
5.5.3	<i>Interpreting Household and Individual Correlates from a Theoretical Perspective</i>	192
5.5.4	<i>Practical Implications of Nonresponse Correlates</i>	193
CHAPTER 6	ISOLATING INTERVIEWER EFFECTS ON NONRESPONSE	196
6.1	Background	196
6.2	Variation Between and Homogeneity Within	197
6.3	Results from the Cross-Classified Multilevel Models	199
6.4	Is a Household Level Needed?	202
6.5	Chapter Summary and Discussion	204
CHAPTER 7	EXPLORING INTERVIEWER AND AREA EFFECTS WITH A MULTINOMIAL APPROACH; EXPLORING THE PERSISTENCE OF INTERVIEWER AND AREA EFFECTS AT WAVE 3	207
7.1	Introduction	207
7.1.1	<i>The Relationship Between Refusals and Non-contacts</i>	207
7.1.2	<i>Interviewer and Area Effects at Wave 3</i>	209
7.2	The Data	210

7.3	Methods	211
	7.3.1 <i>Cross-classified Multilevel Models</i>	211
	7.3.2 <i>Analysis Plan</i>	212
7.4	Results	217
	7.4.1 <i>Technical Aspects</i>	217
	7.4.2 <i>Substantive Aspects at Wave 2</i>	223
	7.4.3 <i>Wave 3 Interviewer/PSU Random Effects</i>	227
7.5	Chapter Summary and Discussion	228
CHAPTER 8	IN LONGITUDINAL STUDIES, DOES HAVING THE SAME INTERVIEWER MAKE A DIFFERENCE TO RESPONSE RATES?	232
8.1	Assessing the Effects of Interviewer Continuity	232
	8.1.1 <i>Background</i>	232
	8.1.2 <i>The Data</i>	237
	8.1.3 <i>Methods</i>	239
	8.1.4 <i>Results</i>	240
	8.1.5 <i>End of Section Summary</i>	246
8.2	Refining and Extending the Analysis of Interviewer Continuity	247
	8.2.1 <i>Background</i>	247
	8.2.2 <i>The Data</i>	248
	8.2.3 <i>Methods</i>	249
	8.2.4 <i>Results</i>	251
	8.2.5 <i>Chapter Summary and Discussion</i>	257

PART 3	EMPIRICAL INVESTIGATIONS INTO	260
	INTERVIEWERS AND RESPONSE	
	ERROR	
CHAPTER 9	WHAT TYPES OF RESPONSE VARIANCE DO	260
	INTERVIEWERS MANIFEST?	
9.1	Background	260
9.2	The Data and Methods	263
	9.2.1 <i>Estimation of ρ</i>	264
	9.2.2 <i>Multilevel Models</i>	265
9.3	Results	265
	9.3.1 <i>Findings from the Hierarchical Analyses of Variance</i>	265
	9.3.2 <i>Findings from the Multilevel Models</i>	269
9.4	Chapter Summary and Discussion	279
CHAPTER 10	DOES INTERVIEWER CONTINUITY AFFECT	283
	RESPONSE QUALITY?	
10.1	Background	283
10.2	Data and Methods	284
10.3	Results	285
	10.3.1 <i>Changes in Substantive Answers</i>	285
	10.3.2 <i>Other Indicators of Response Quality</i>	291
10.4	Chapter Summary and Discussion	292
PART 4	DISCUSSION AND INTEGRATION	294
CHAPTER 11	SUMMARY AND CONCLUSIONS	294

11.1	Integrating What the Various Strands of Empirical Work Have to Say	294
11.1.1	<i>The Background Literature</i>	294
11.1.2	<i>The Opportunities and Challenges Presented by the Data</i>	295
11.1.3	<i>Interviewers and Nonresponse</i>	296
11.1.4	<i>Interviewers and Response Error</i>	301
11.2	Implications for Survey Research Practice	303
11.2.1	<i>Implications for Fieldwork Strategies</i>	303
11.2.2	<i>Implications for Analysis Strategies</i>	308
11.3	Limitations of this Research and Suggestions for Future Research	309

REFERENCES	315
-------------------	------------

APPENDICES

APPENDIX A	Comparison of Interpenetrated Sub-Sample to Full British Household Panel Study Sample at Wave 2	347
APPENDIX B	Bivariate Correlates of Nonresponse: Nonresponse Rates and Logistic Regression Coefficients for Various Categories	349
APPENDIX C	Background Tables for the Calculation of Probabilities of Nonresponse by Respondent Co-operation and Contactability, Characteristics of the Respondent, and Location	357

APPENDIX D	Wave 2 Binary and Multinomial Cross-Classified Multilevel Regression Models using RIGLS, PQL and 2 nd Order Estimation	360
APPENDIX E	Wave 2 Binary and Multinomial Cross-Classified Multilevel Regression Models using RIGLS and MQL Estimation	364

LIST OF TABLES

Table 1	ANOVA Table for Calculation of Interviewer Variance	38
Table 2	Summary of Interviewer Variance Investigations Using ρ	47
Table 3	UK General Household Survey Annual Response Rates	125
Table 4	British Household Panel Study Response Rates for Waves 1 and 2	158
Table 5a	Multiple Logistic Regression Model Predicting Nonresponse	178
Table 5b	Predicted Probabilities of Nonresponse Depending on Respondent Co-operation and Contactability, Characteristics of the Respondent, and Location	179
Table 6	Cross-Classified Multilevel Logistic Regression Models: Fixed and Random Effects	186
Table 7	Estimates of ρ from Hierarchical Analyses of Variance Nonresponse Outcome Variable	200
Table 8	Cross-Classified Multilevel Logistic Regression Models: Variance Components Model	200
Table 9	Cross-Classified Multilevel Logistic Regression Models: Random Effects from Covariates Model (from Table 6)	200
Table 10a	Comparing an Individual Level Nonresponse Model With and	203

Without Household Level Variation: Variance Components
Model

Table 10b	Comparing an Individual Level Nonresponse Model With and Without Household Level Variation: Covariates Model	203
Table 11	Response Figures for Waves 2 and 3 of the Interpenetrated Sub-Sample	211
Table 12	Wave 2 Binary and Multinomial Cross-Classified Multilevel Regression: Variance Components Models using RIGLS, PQL and 2nd Order Estimation and Assuming Binomial/Multinomial Variation at Level 1	220
Table 13	Wave 2 Binary and Multinomial Cross-Classified Multilevel Regression: Covariates Models using RIGLS, PQL and 2nd Order Estimation and Assuming Binomial/Multinomial Variation at Level 1	221
Table 14	Combined Random Effects of Interviewers and Areas at Wave 3	228
Table 15	Variability of the Mean Difference in Nonresponse Rates Between the Same and Different Interviewer Groups by Geographic Pool at Wave 2	241
Table 16	Random Slopes Models for Interviewer Continuity at Wave 2: Modifications to the Variance Components Models Shown in Table 8, Chapter 6	242
Table 17a	Interviewer Responses to Debriefing Q22 – Outcome of Lack of Continuity (n=50)	244
Table 17b	Responses to Debriefing Q23 – Strategies to Cope with Lack	245

of Continuity (n=50)

Table 18	The Impact of Interviewer Continuity Across Waves of the British Household Panel Study for Original Wave 1 Respondents Who Have Not Moved Out of Their Original Areas	252
Table 19	Exploring the Possibility of Non-Random Interviewer Attrition using Loglinear Modelling: Three-way Tables - Interviewer Continuity by Individual Level Nonresponse by Region	255
Table 20	Random Slopes Models for Interviewer Continuity at Wave 3	257
Table 21	Cross-Classified Multilevel Logistic Regression Model: Response to the Interviewer Check Item ‘Whether Children Were Present’	271
Table 22	Cross-Classified Multilevel Logistic Regression Model: Response to the ‘Reads the Independent’ Question	274
Table 23	Cross-Classified Multilevel Logistic Regression Model: Response to the ‘Likely to Have More Children’ Question	276
Table 24	Month of Interview by Interviewer Continuity	286
Table 25	Differences in Substantive Answers Depending of Whether the Respondent Received the Same or a Different Interviewer	289
Table 26	The Relationship between Interviewer Continuity and Other Measures of Response Quality	292

LIST OF FIGURES

Figure 1a	Kish’s Geometric Representation of Total Survey Error	31
Figure 1b	Deming’s View of Total Survey Error	31

Figure 1c	Frankel and Dutka's View of Total Survey Error	31
Figure 2	Classification of Sources of Survey Error	34
Figure 3	Interpenetrated Elements of British Household Panel Study Data	51
Figure 4	Interpenetrated Options for Wave 3	52
Figure 5	Excerpt from U.S. Current Population Survey, October 1980 Version	79
Figure 6	Factors Affecting Survey Participation: Theoretical Model from the Work of Groves, Cialdini and Couper	123
Figure 7	Nonresponse by Type for the National Health Interview Survey, 1967-1985	124
Figure 8	A Model of Bias in the Interview	129
Figure 9	A Simple Model of the Research Interview	130
Figure 10	Diagram of Respondent's (R) Question-Answering Process	132
Figure 11a	Relations Among Characteristics of Questions, Interviewer, Respondent, and Interview-Context on the One Hand, and Proportion of Valid Responses and Distribution of Responses on the Other Hand	133
Figure 11b	Relations Among Characteristics of Questions, Interviewer, Respondent, and Interview-Context on the One Hand, and Some Characteristics of Response Behaviour on the Other Hand	134
Figure 12	A Model of the Survey Interaction Process	136
Figure 13	Highlights of the Esposito and Jobe (1991) Taxonomy	137
Figure 14a	Extending the Theoretical Model of Groves, Cialdini, and Couper	156

Figure 14b	Elements of the Theoretical Model for which Data are Available	157
Figure 15	Exploring the Normality of Residuals: Normal Scores versus Standardised Residuals - Illustrations for Two Models	219
Figure 16	Interviewer Continuity Items from British Household Panel Study Interviewer Debriefing Questionnaire	238
Figure 17	Intra-interviewer and Intra-cluster Correlations: Cumulative Distribution of ρ_i and ρ_s	267

ACKNOWLEDGEMENTS

The data used in this thesis were made available through the ESRC Data Archive. The data were originally collected by the ESRC Research Centre on Micro-social Change at the University of Essex. Neither the original collectors of the data nor the Archive bear any responsibility for the analyses or interpretations presented here. The initial investigation into separating interviewer effects from area effects for survey nonresponse (Chapters 5 and 6) overlap with Sub-Project 1 of a Social and Community Planning Research (SCPR) project designed and headed by myself and funded by the Economic & Social Research Council (ESRC) award number: R000235776.

I would like to acknowledge the special contribution of NOP Social and Political for supplying me with data on the characteristics of their interviewers. I would also like to acknowledge the supportive co-operation of the ESRC Centre on Micro Social Change and in particular their willingness to retrieve and key almost 3,000 Wave 1 and Wave 2 call records for the purpose of this project and for giving me special permission to link my subset of data to Census small area data.

I would like to thank the NOP interviewers who worked on the British Household Panel Study and all the men and women who were respondents and nonrespondents without whom this thesis would not be possible.

Most importantly, I would like to thank my supervisor, Colm O'Muircheartaigh, for his unfailing support, guidance, and encouragement.

EXECUTIVE SUMMARY

Background

- Interviewer-based data collection is the norm for social and market research surveys in the United Kingdom and is likely to remain so for the foreseeable future.
- This thesis explores the impact of the interviewer in social surveys. Two main areas are considered: (1) the effect of the interviewer on nonresponse (both variance and bias) and (2) the effect of the interviewer on response variance.
- Chapter 1 of this thesis provides a context for this research by
 - discussing the concept of total survey error,
 - describing the traditional measurement of interviewer effects as interviewer variance,
 - comparing and contrasting the ANOVA and correlational viewpoints on this source of error,
 - describing the data source, the British Household Panel Study (BHPS), and
 - describing the analytic plans.
- Survey samples in the UK typically assign one interviewer to each primary sampling unit (PSU) so that the effects of interviewer and area are confounded. This investigation into interviewer effects was made feasible through the interpenetrated sample design experiment designed by C. O’Muircheartaigh (my supervisor) and myself for implementation in Wave 2 of the BHPS.
- The BHPS is a multi-topic panel survey where interviews are conducted with all members of a household aged 16 and over. It is conducted by the ESRC Centre for Micro-Social

Change at the University of Essex. Beginning in 1991, the Wave 1 design consisted of a multistage stratified cluster sample covering all of Great Britain.

- For the interpenetrated sample design experiment, households were assigned at random to interviewers within 35 'geographic pools' across the country, with each 'geographic pool' containing 2 or 3 interviewers and 2 or 3 PSUs. The sub-sample included in the experiment included 1,406 co-operating and partially co-operating Wave 1 households which were issued for fieldwork at Wave 2 (approximately a quarter of the full BHPS sample).
- Auxiliary data were gathered from a number of sources. Data on the characteristics of interviewers were provided by NOP Social and Political. Data on the characteristics of areas were taken from 1991 Census small area statistics. Data on households and individuals were taken from Wave 1 of the BHPS and data on Wave 1 and 2 call records were specially keyed by BHPS staff for this project.
- Analytic strategies varied by topic and included chi-square tests of independence in two-way tables, loglinear modelling in three-way tables, hierarchical analysis of variance, multiple logistic regression and multinomial regression. The most advantageous approach was the use of cross-classified multilevel modelling, focusing on logistic regression and multinomial regression.

Interviewers and Nonresponse

- The literature on nonresponse was reviewed showing evidence to suggest that response rates to general population surveys have tended to fall over recent decades. Some surveys

have managed to maintain their response rates, but this has been through a higher investment of time and money (see Chapter 3).

- Nonresponse, response rates, nonresponse bias and nonresponse variance were defined and a theoretical model involving the social context, and interviewer and respondent factors was discussed (see Chapter 3)

- Also reviewed was the role of the interviewer and the ways in which interviewer background characteristics, expectations and attitudes, and actual behaviour are believed to affect their response rates (see Chapter 2).

- Specific goals for this part of the thesis were:
 - to examine the various characteristics of address residents that affect their inclusion in surveys, the impact of various interviewer characteristics on their ability to achieve good response rates, and the impact of the characteristics of areas which make survey research difficult, and to assess the usefulness of various measures of respondent co-operation and contactability (see Chapter 5),
 - to separate the interviewer effect on nonresponse from other effects such as the effect of the area and the effect of the characteristics of households and individual address residents (see Chapter 6),
 - to explore the covariance between the refusal and non-contact components of a multinomial model of nonresponse (see Chapter 7)
 - to explore combined interviewer/PSU effects at Wave 3 and see how these differ from Wave 2 (see Chapter 7),

- to assess experimentally the conventional survey wisdom which suggests that response rates are higher when the same interviewer is allocated at successive waves to the same respondents in longitudinal surveys (i.e., ‘interviewer continuity’) using the interpenetrated sample at Wave 2 (see Chapter 8), and
 - to assess experimentally the impact of interviewer continuity at Wave 3 and see how Wave 2 and 3 findings compare to the full sample results (see Chapter 8).
- Six main dependent variables were considered: Total nonresponse at both the household and individual levels and the two most frequent components of nonresponse (refusals and non-contacts) at both the household and individual levels.

Findings with Respect to Nonresponse

- Interviewer characteristics (such as age, gender, experience, and grade level) and area characteristics (such as population density, proportion of flats in the area, percentage of non-white residents, etc. taken from 1991 Census small area data) are important bivariate predictors of nonresponse. These effects, however, virtually disappear when respondent and household characteristics are controlled for (see Chapter 5).
- The random effects of interviewers and areas nonetheless are important with interviewers affecting individual and household level refusals and household level non-contacts; and areas affecting individual level non-contacts (see Chapters 6 and 7).
- This suggests that the easily measurable characteristics of interviewers and areas contained in the logistic regression models are not the key aspects of interviewers and areas which affect nonresponse (see Chapters 5 and 6).

- Given that a household level term has not been considered in the individual level models, these results for the random terms should be considered as indicative rather than definite.
- The combined effects of interviewer and area variation was found to persist over time, although their impact was less (see Chapter 7).
- Characteristics of individuals and households are important predictors of nonresponse. The current findings were very much in line with past research in terms of suggesting the types of households and individuals who are most likely to be missed in a survey. These included several indicators which suggest individuals and households that are less well off economically are both harder to find and harder to persuade. There was also some indirect evidence for the effects of urbanicity on nonresponse. Other characteristics of individuals and households which were associated with nonresponse, as identified by both the bivariate and multiple logistic regression analyses, include those who live in households without children, or with four or more adults of working age, males, persons of non-white origin, and the young and the old (see Chapter 5).
- Particularly useful indicators of attrition nonresponse are measures of respondents' co-operation and contactability from the previous wave. These suggest that respondents who were difficult to obtain an interview from in the first wave, but who nonetheless participated, are much more likely not to participate at all in the next wave (see Chapter 5).
- Using a polytomous dependent variable (refusals, non-contacts, and interviews), correlations between refusals and non-contacts within interviewers and within areas were also found (see Chapter 7).

- Interpenetrated sample data from Waves 2 and 3 showed little support for the believed benefits of interviewer continuity, however, effects were found in the full sample which was prone to non-random interviewer attrition (see Chapter 8).
- Interviewers have good suggestions about minimising the impact of interviewer discontinuity (see Chapter 8)

Interviewers and Response Error

- Response error, often called measurement error, derives from four specific sources in surveys: the respondent, the interviewer, the questionnaire, and the mode of data collection. Yet, in reality each of these influences is not separate. It is often the interaction between them which causes effects (e.g., a face-to-face interview with a black interviewer, a questionnaire focused on racial attitudes, and a white respondent - see Schuman and Converse, 1971)
- Theoretical models which describe this interaction process were reviewed. This review suggested that the perfect model is something to aim toward rather than something to achieve, given the great complexities of the survey interaction process (see Chapter 4).
- Response error was also visited from a mathematical perspective comparing and contrasting the viewpoints of psychometricians and sampling statisticians (see Chapter 4).
- Also reviewed was the role of the interviewer and the ways in which interviewers' background characteristics, expectations and attitudes, and actual behaviour are believed to

affect response quality (Chapter 2) and the types of questions which are most susceptible to interviewer effects (Chapter 2).

- Specific goals for this part of the thesis were:
 - to measure the complex variance due to both interviewers and areas (see Chapter 9),
and
 - to explore the relationships between interviewer continuity and response quality (see Chapter 10).

- To achieve the first goal, all variables in the BHPS dataset with 700 or more cases were analysed with hierarchical analyses of variance to estimate the intraclass correlation coefficient for interviewers (see Chapter 1). This is a portable measure of the homogeneity of randomly assigned respondents within an interviewer's workload.

- Also considered were intraclass correlation coefficients for areas, representing the homogeneity of respondents within areas.

- Through the measures *inteff* and *deff* (see Chapter 9), one can assess the extent of variance inflation of the mean due to interviewer and sample clustering, respectively.

- To achieve the second goal, again all of the variables in the BHPS dataset were cross-tabulated with interviewer continuity to see if substantive answers were affected.

- Also constructed were counts of the total number of refusal and don't know responses across the various questionnaires as well as counts of missing and wild values, and the

imputation flags with respect to the income variables. These indicators of response quality, along with whether or not the interviewer had seen the respondent's payslip were examined in relation to interviewer continuity and in relation to the co-operativeness of the respondent.

Findings with Respect to Response Error

- Across the 820 variables or categories of variables considered in the study, there was evidence of a significant impact of both population clustering and the clustering of households/individuals in interviewer workloads (see Chapter 9).
- The sample design effects and interviewer effects were comparable in impact, with overall inflation of the variance as great as five times the unadjusted estimate, with a median effect of an 80 percent increase in variance (see Chapter 9).
- The magnitude of the intra-class correlation coefficient for interviewers was comparable across question types, though the most susceptible items tended to be the interviewer check items (see Chapter 9).
- There was a tendency for variables that were subject to large design effects to be sensitive also to large interviewer effects (see Chapter 9).
- Across the 695 variables considered in Chapter 10, there was virtually no evidence to suggest that interviewer continuity influences response quality.

Recommendations from the Research

- Multilevel modelling is a very useful tool for the survey analyst. For example:
 - Although hierarchical analyses of variance can be used to estimate interviewer variability, the use of cross-classified multilevel logistic regression models are essential for the accurate estimation of the random effects of interviewers and areas (see Chapters 6, 7, and 9).
 - Cross-classified multilevel multinomial regression is essential for looking at the covariance between categories of the dependent variable (e.g., refusals and non-contacts) within random terms (e.g., within interviewers or within areas) (see Chapter 7).
 - Cross-classified multilevel modelling allows for the incorporation of the complex variance-covariance structure, present in almost all survey data, directly into the substantive analyses (see Chapters 6, 7, and 9).
 - Different conclusions can be reached when this complex structure is ignored (take, for example, the difference in conclusions about the significance of fixed effects in a single-level versus a cross-classified multilevel multinomial model (see Chapter 7), or the difference in conclusions about the impact of interviewer characteristics in explaining response variance under a single level versus a cross-classified multilevel logistic model (see Chapter 9).
 - Multilevel models facilitate the direct incorporation of explanatory variables to help explain any observed random effects (see Chapters 6, 7, and 9).
- This research has also suggested facets that will be useful to field staff from survey organisations. For example:

- that the visible characteristics of interviewers (such as their age and gender) do not have an impact on nonresponse in general-purpose household surveys so there should be no specific concern for interviewer recruitment for such studies (see Chapter 5),
- that there is a particular need for more nonresponse research on those aspects of interviewers not directly studied in this thesis, such as interviewers' attitudes and expectations or interviewers' doorstep skills (see Chapter 6),
- that interviewers in this study tended to be either good at both aspects of their nonresponse work (i.e., minimising refusals and minimising non-contacts) or poor at both aspects, although this may not be true for all interviewer populations (see Chapter 7),
- that there is a trade-off with respect to individual level nonresponse within a household: Higher refusals can mean lower non-contact rates and lower refusals can mean higher non-contact rates (see Chapter 7). (This is consistent with the hypothesis that non-contacts within households can be hidden refusals),
- that targeting difficult households and individuals from previous waves to receive special attention at subsequent waves can be a useful strategy for improving response rates (see Chapter 5),
- that there is the possibility of revisiting past non-contacts and refusals at future waves of a panel study as these households and individuals may now participate (see Chapter 5),
- the types of characteristics that it would be useful to measure with a 'nonresponse form' during data collection on a one-off survey to facilitate nonresponse weighting (see Chapter 5),
- that it is probably wise to continue with an interviewer continuity approach despite the lack of evidence to support its usefulness (see Chapters 8),

- that when it is not possible to maintain interviewer continuity, then as a minimum give interviewers the name of the previous interviewer, instruct them to mention the name, and instruct them to give regards from the previous interviewer and explain why the previous interviewer is not there (see Chapter 8),
- that there is a need for regular measurement of interviewer variance as well as other aspects of interviewers' work (see Chapters 6 and 9), and
- that due to the presence of interviewer effects on substantive variables, smaller dedicated interviewer forces with large assignments should be avoided (see Chapter 9).

Limitations of this Research and Suggestions for Future Research

- The intraclass correlation coefficient, ρ , is an estimate. Although standard errors of ρ were not calculated in this thesis, it would be useful to calculate standard errors for ρ whenever possible. Research on how best to calculate ρ in the case of a multilevel logistic regression also is needed.
- Results with respect to the random effects of interviewers and areas should be considered as indicative rather than definite as a larger sample is ideally needed for the household level investigations and the individual level investigations have omitted household level variation.
- No explicit cost models have been considered. The experience of the BHPS, where randomisation only took place within geographic pools made up of adjacent PSUs, however, suggests that the costs of interpenetration can be negligible in comparison to a total survey budget.

- The influence of interviewers on scales in addition to individual items has not been considered. Although considered by O’Muircheartaigh and Wiggins (1981), this is still a useful topic for further research.
- The results of this thesis only generalise to the 153 PSUs of the BHPS whose centroid was no more than 10 kilometres from any other PSU. There is a slight urban bias in the interpenetrated sample.
- This is just one study. Some of the unexpected findings with respect to interviewer characteristics (such as age and experience) are probably best explained by the particular interviewers who were included in the study. A replication of the work on interviewer characteristics would be useful.
- Research into what specific nonresponse strategies work with particular kinds of nonrespondents would be useful.
- Further experimental research is needed with respect to the affects of interviewer continuity with different interviewers, different survey topics and different survey populations.
- Ideally one would like a research design in which all the elements of the full theoretical model of nonresponse (see Chapter 5) are included and can be estimated.
- In addition, future research on the interviewer should ideally be motivated from the perspective of total survey error so that the interviewer is seen as one piece of a highly interactive process.

PART 1 BACKGROUND

CHAPTER 1 INTRODUCTION

1.1 Total Survey Error

The increasing popularity and use of survey methodology is very evident in social research today. However, as Dijkstra and van der Zouwen (1982, preface) suggest, "*the criticisms on this data collection method are also becoming more articulate and convincing.*" This thesis represents an exploration of the survey research interviewer's contribution to error in survey data, more precisely to nonresponse error and response variance.

Nonresponse error and response variance are components of 'total survey error'. This introductory section gives a brief overview of total survey error, thereby providing a context for the specific study of interviewer effects (see Chapter 2), nonresponse error (see Chapter 3) and response error (see Chapter 4).

Most of the models developed to evaluate total survey error are generalisations of the mean squared error formulation. The mean squared error formulation from mathematical statistics is as follows:

$$\begin{aligned}MSE(y) &= E(y_c - Y_{true})^2 \\ &= E(y_c - E(y))^2 + (E(y) - Y_{true})^2\end{aligned}\tag{1}$$

where y_c = the sample estimate,
 Y_{true} = the true population value, and
 $E(y)$ = the mean value of y_c taken over all possible samples under the same sample design.

Thus, the first part of the formula represents the sampling variance and the second part, the statistical bias. As Andersen, Kasper and Frankel (1979, p. 3) suggest, this formulation is limited in that “*it assumes perfect implementation of the sample design, perfect measurement of the variable values for all sample elements, and perfect sampling frames.*” Further, as suggested by Simonoff (1993, p. 3) for some purposes it can be misleading because “*it does not address the relative importance of bias and variability, and the differing effects of negative bias and positive bias.*” Following on from Andersen and his colleagues’ concerns, the mean square error in survey statistics is typically broadened to include the sum of all the biases squared and the sum of all the variable errors. Take, for example, a general model for the error specification in the sample mean as synthesised and presented by Kish (1965, p. 516) based on the work of Cochran (1963), Hansen, Hurwitz, and Madow (1953), Sukhatme (1954) and Zarkovich (1963). This formulation is as follows:

$$(\bar{y} - \bar{Y}_{true})^2 = \left(\sum_g \beta_g \right)^2 + \sum_v \frac{s_v^2}{m_v} \quad (2)$$

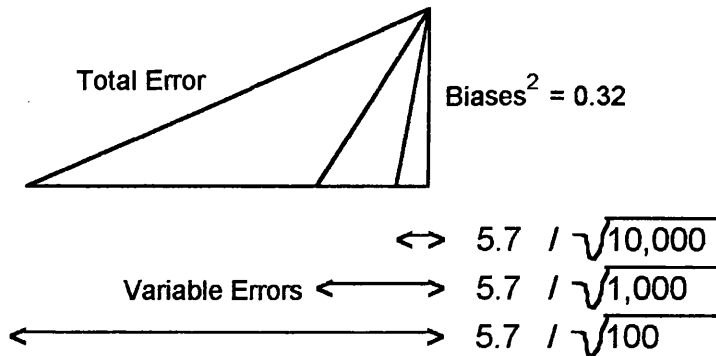
where the first term = the square of all the bias terms and

the second term = the sum of all the variance terms,

with both of these from diverse sources.

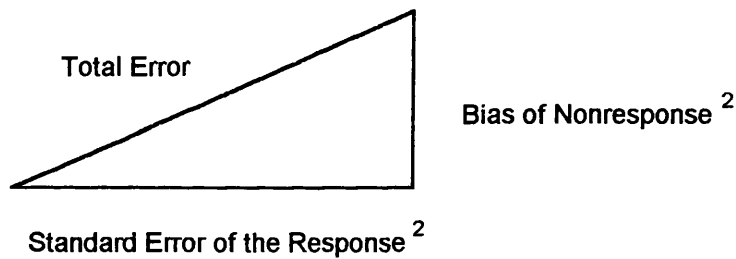
Note that each of the variance terms is expressed as a unit variance as suggested by division by m_v , the number of units. Kish (1965) continues on to describe how this simple, general model can be broadened to include the finite population correction factor (i.e., one minus the sampling fraction), or include covariance between elements which can arise from sample clustering, interviewers, and coders, etc. Kish (1965) also suggests the use of triangle diagrams as a tool in comparing the relative importance of the variable and bias components of total survey error. For example, in a right triangle, the height represents bias, the base

Figure 1a: Kish's Geometric Representation of Total Survey Error



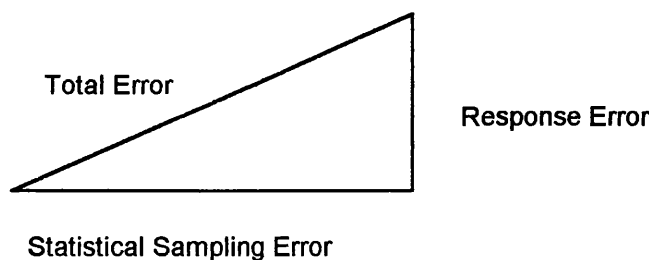
Source: Kish, L. (1965), *Survey Sampling*, New York: John Wiley & Sons, p. 521.

Figure 1b: Deming's View of Total Survey Error



Source: Deming, W.E. (1953), On a Probability Mechanism to Attain an Economic Balance Between the Resultant Error of Response and the Bias of Nonresponse, *Journal of the American Statistical Association*, 48, 743-772.

Figure 1c: Frankel and Dutka's View of Total Survey Error



Source: Frankel, L.R. and Dutka, S. (1983), Survey Design in Anticipation of Nonresponse and Imputation, in: W.G. Madow and I. Olkin (eds) *Incomplete Data in Sample Surveys, Vol 3, Proceedings of the Symposium*, New York: Academic Press, p. 69-83.

represents variable error and the hypotenuse measures total error (see Figure 1a). These diagrams clearly illustrate that as sample size increases (variable error decreases), but more importantly even a moderate sized bias can dominate the mean square error.

Note that other breakdowns have also been suggested as ways of viewing total survey error. Deming's view (1953) is presented in Figure 1b. As described by Lessler and Kalsbeek (1992), under Deming's conception the standard error of response clearly incorporates all other biases and all other variable errors. Frankel and Dutka's view (1983) in Figure 1c makes a distinction between the sampling error and response errors. Their usage of the term response error apparently includes all other sources of variable error and bias (in contrast to Andersen, Kasper and Frankel, 1979, below).

Building on the work of Kish (1965), Andersen, Kasper, and Frankel (1979) show a useful diagram to categorise the many sources of both variable error (variance) and bias in survey data. A modified version of Andersen and his colleagues' model is shown as Figure 2. As can be seen, after the primary divide between variable errors (variance) and bias errors, there is a further subdivision into *sampling* and to *nonsampling* error. Here nonsampling error is simply defined as "*an error in sample estimates which cannot be attributed to sampling fluctuations*" (Kendall and Buckland, 1982, p. 137). Nonsampling error can then be seen to derive from errors which occur through the measurement process itself (i.e., errors of *observation*) and those which occur because the sampled unit can not be observed (i.e., errors of *non-observation*). Errors of observation can further be divided into those errors which occur during the data collection phase (labelled as *field* errors in the diagram) and those errors which occur back in the office (labelled *processing* errors in the diagram). These data collection/field errors have traditionally been called

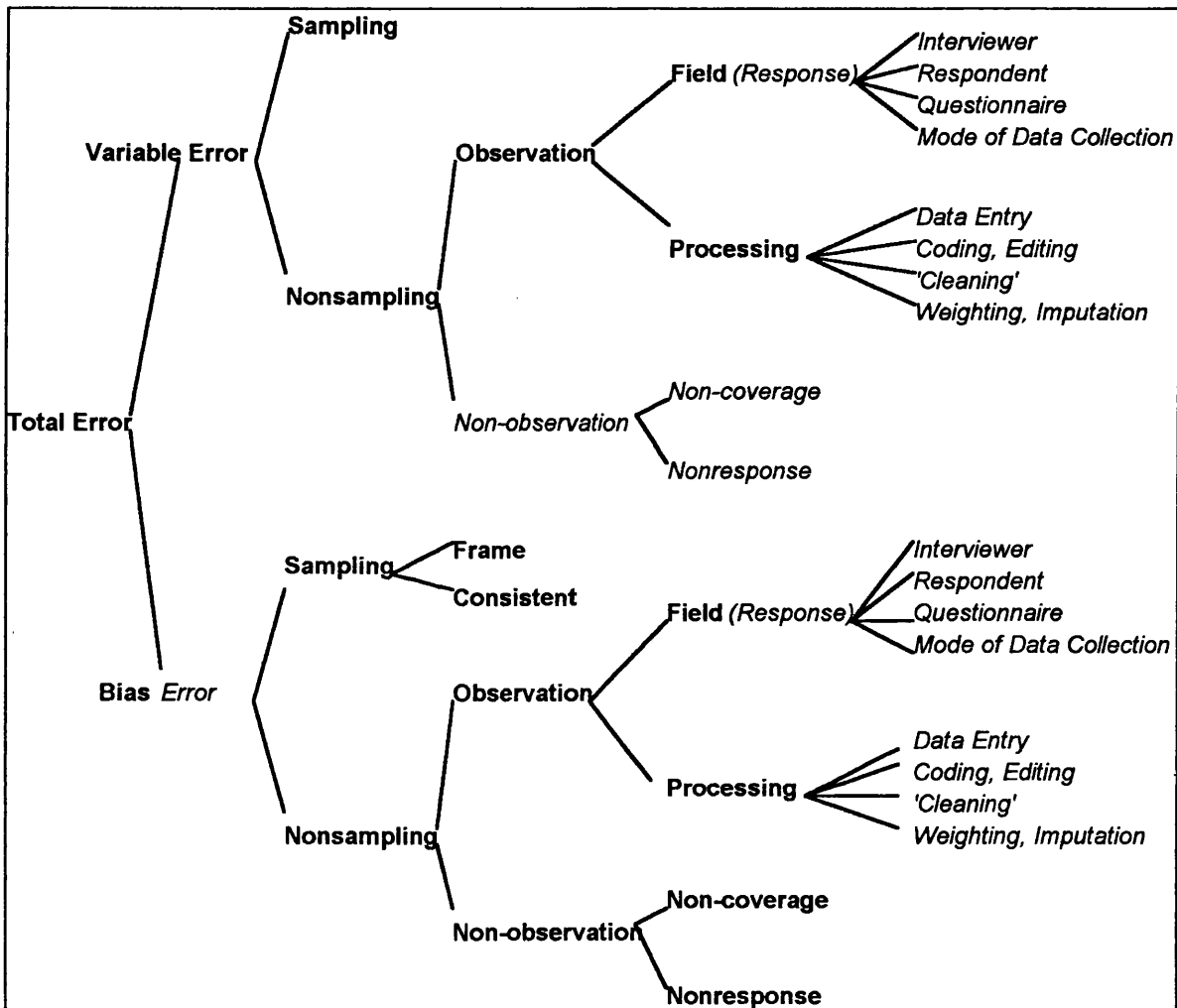
‘response errors’ (see Hansen, *et al*, 1951) and have also been referred to as ‘measurement errors’ (see, for example, Biemer *et al*, 1991). Errors of non-observation include the errors of *non-coverage* and *nonresponse*. Field errors are conveniently categorised into four categories, those arising from the interviewer, the respondent, the questionnaire, and the mode (i.e., method) of data collection (see Groves, 1989). Processing errors represent all those types of errors which can occur while translating the data into machine readable form and analysing it.

The italicised text in Figure 2 represent modifications (additions) to Andersen, Kasper and Frankel’s original diagram. The italicised text was added to extend and to balance the two halves of the tree because this thesis examines the variable side of both nonresponse and response error. Both Kish (1965) and Andersen, Kasper, and Frankel (1979) suggest that it is conceptually and operationally possible to have variable errors and biases arising from both sampling and nonsampling causes, although these are not illustrated in their diagrams perhaps because of the view that ‘sampling errors account for most of the variable errors of a survey, and biases arise chiefly from nonsampling sources’ (Kish, 1965, p. 509). Note that with the addition of the italicised text, Figure 2 is analogous to the template diagram used by Groves (1989) in his first chapter where he presents a very thorough review of how the different disciplines involved in the social sciences view sources of error.

Studying variable errors is facilitated by considering certain elements of the design to be fixed and others to vary over replications while under the same ‘essential survey conditions’. (The ‘essential survey conditions’ were defined by Hansen, Hurwitz, and Bershad (1961) (see also Kish, 1965). As outlined by O’Muircheartaigh (1982), these

include such aspects as the subject matter, the data collection and recording methods, the timing and sponsorship, the type or class of interviewers and coders to be used in an interview survey, etc. It is also useful to note that thinking of a design as repeatable is not necessarily at odds with the fact that a given survey **may not be repeatable** (Fellegi, 1964).

Figure 2: Classification of Sources of Survey Error



Key to terms:

- Variables errors, measured by the variance of a statistic, arise from variability in the achieved values.
- Bias is a constant error in the sense it occurs over all possible surveys using the same design.
- Nonsampling error refers to errors which cannot be attributed to sampling fluctuations.
- Frame biases are the consequence of using inappropriate selection procedures to choose the sample.
- Consistent biases are the consequence of using biased but consistent estimators.
- Observation errors are caused by obtaining and recording observations incorrectly.
- Non-observation errors arise from failure to obtain observation on some segment of the population.
- Field (response) errors arise in the collection of observations.
- Processing errors occur while translating the data into machine readable form and analysing it
- Non-coverage denotes failure to include some elements of the defined survey population in the actual operational sampling frame.
- Nonresponse refers to the failure to obtain observation on some elements selected and designated for the sample.

Conceiving of the same error from both sides of the tree can be illustrated with the example of nonresponse. Although nonresponse is typically treated as a bias, it can also be conceptualised as a variable error. If, over repeated surveys under the same ‘essential survey conditions’ and the same sample, the same persons are always nonrespondents, this could result in nonresponse bias. In contrast, if the sample members are actually variable in their participation decisions, this represents nonresponse variance. (The subject of nonresponse variance is discussed further in Section 3.3).

Figure 2 clearly illustrates the importance of considering nonsampling errors in all confidence intervals, as these are believed to account for the largest source of total survey error (see, among others, Kish, 1965; Andersen, Kasper, and Frankel, 1979; O’Muircheartaigh, 1977; Groves, 1989; Biemer *et al*, 1991; Lyberg *et al*, 1997). Yet despite the fact that the current interest in ‘quality’ (cf continuous quality improvement, Imai, 1986) has come to the survey world, nonsampling errors are still often ignored. This is in part due to the difficulty in obtaining proper estimates of their magnitudes.

1.2 Measuring Interviewer Effects

This thesis is about the impact of the interviewer. In quantitative sample surveys the interviewer can be seen as both an aid and a hindrance. The presence of an interviewer is seen to minimise many sources of response error (e.g., through motivating the respondent and controlling the response process). Yet, at the same time, the interviewer is also seen as one of the principal sources of error in data collected from structured face-to-face interviews.

The literature on the measurement of interviewer *effects* typically focuses on interviewer *variance* (at the aggregate level) rather than the measurement of interviewer *bias*. One of the

prime reasons for this is the practical and logistical constraints on the measurement of bias. Bias can only be assessed by comparing survey data to external validation data, which can be difficult to obtain and simply does not exist for many variables of interest. On the other hand, the impact of variance can be studied with the approaches of *interpenetration* and *re-enumeration*. Interpenetration requires the randomisation of the allocation of elements to conditions. In the context of studying interviewer variance, interpenetration was pioneered by Mahalanobis (1946) under the name of ‘interpenetrating samples’. Re-enumeration requires multiple measurements or trials on the same element (see, for example, Hansen, Hurwitz, and Bershad, 1961; Fellegi, 1964; O’Muircheartaigh, 1977, 1982).

Through making use of interpenetrated sample designs, survey statisticians have expressed interviewer variance in formal statistical models of two kinds. In the analysis of variance (ANOVA) framework the errors are seen as net biases for the individual interviewers and the interviewer effect is seen as the increase in variance due to the variability among these biases. The alternative approach is to consider the interviewer effect to arise from the creation of positive correlations among the response deviations contained in (almost all) survey data; the increase in the variance of a mean is due to the positive covariance among these deviations. Each of these two approaches is described in detail below. (The reader is directed to Section 4.3 for a general introduction to response error models.)

1.2.1 The ANOVA Model

The ANOVA approach to the calculation of interviewer variance was expounded by Kish (1962) and developed by Hartley and Rao (1978) and others. The model views interviewer variance as a component of the total variance per respondent as follows:

$$\sigma^2 = \sigma_a^2 + \sigma_b^2 \quad (3)$$

where $\sigma^2 =$ total variance,

$\sigma_a^2 =$ the between interviewer variability, and

$\sigma_b^2 =$ the within interviewer variability.

The interviewer effect is measured by the proportion of the total variability which is due to interviewers and is designated by ρ , the intraclass correlation coefficient:

$$\rho = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_b^2} \quad (4)$$

Using Kish's (1962) notation, an estimate can be obtained from sample observations as

$$\hat{\rho} = \frac{s_a^2}{s_a^2 + s_b^2} \quad (5)$$

Given an interpenetrated sample design, Kish (1962) demonstrates that $\hat{\rho}$ can be obtained from a one-way analysis of variance in which the survey characteristic of interest (y) is the dependent variable and interviewers (i) are the explanatory variables. The ANOVA table shown by Kish (1962) is presented as Table 1,

where $n =$ the total sample size,

$n_i =$ the size of the i -th interviewer's workload,

$\alpha =$ the number of interviewers, and

$k =$ the average interviewer workload, i.e., $k = n / \alpha$.

Table 1: ANOVA Table for Calculation of Interviewer Variance

Source of Variation	Degrees of Freedom	Sum of Squares (SS)	Mean Square	Components of the Mean Squares
Among interviewers	a-1	$\sum y_i^2 / n_i - y^2 / n$	$V_a = \frac{SS(a)}{a-1}$	$s_b^2 + ks_a^2$
Within interviewers	n-a	$\sum \sum y_{ij}^2 - \sum y_i^2 / n_i$	$V_b = \frac{SS(b)}{n-a}$	s_b^2

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

As described by Groves (1989) using Kish's original notation, $\hat{\rho}$ is therefore calculated as follows:

$$\hat{\rho} = \frac{\frac{V_a - V_b}{k}}{\frac{V_a - V_b}{k} + V_b} \quad (6)$$

Through substitution of terms and some simple algebra, we can see that this is equivalent to Kish's original formulation as follows:

$$\begin{aligned} & \frac{(s_b^2 + ks_a^2) - (s_b^2)}{k} = \frac{(ks_a^2)}{k} = \frac{s_a^2}{s_a^2 + s_b^2} \\ & = \frac{(s_b^2 + ks_a^2) - (s_b^2)}{k} + (s_b^2) = \frac{(ks_a^2)}{k} + (s_b^2) \end{aligned}$$

It can also be seen that the formulation by Groves (1989) is equivalent to the more commonly used formula among survey researchers (see, for example, Freeman and Butler, 1976; Fowler and Mangione, 1990; Hox, 1994):

$$\hat{\rho} = \frac{MSB - MSW}{MSB + (k-1)MSW} \quad (7)$$

where $MSB =$ the mean square between (among) interviewers, and
 $MSW =$ the mean square within interviewers.

Starting with the Groves formulation, this equivalence is shown as follows:

$$\begin{aligned} \hat{\rho} &= \frac{\frac{V_a - V_b}{k}}{\frac{V_a - V_b}{k} + V_b} = \frac{\frac{MSB - MSW}{k}}{\frac{MSB - MSW}{k} + MSW} \\ &= \frac{\frac{MSB - MSW}{k}}{\frac{MSB - MSW}{k} + MSW} \times \frac{k}{k} \\ &= \frac{MSB - MSW}{MSB - MSW + k(MSW)} \\ &= \frac{MSB - MSW}{MSB + (k - 1)MSW} \end{aligned}$$

Given that ρ is not directly dependent on sample size, it allows for the comparison of interviewer effects across different surveys. It should be noted, however, that if interviewers alter the homogeneity of their workload through nonresponse, then ρ reflects this as well (see Groves and Magilavy, 1986).

1.2.2 The Correlation Model

The correlation model was first presented by Hansen, Hurwitz and Bershada (1961) and is often called the Census Bureau model. The model is based on the idea of the response deviation, defined as

$$d_{jt} = x_{jt} - P_j \tag{8}$$

where $x_{jt} =$ an observation on the j -th unit in the survey on the t -th trial, and

$$P_j = E_{G_j}(x_{jt}) \quad (9)$$

is the expected value of over all possible samples and trials conditioned on the j -th unit and certain fixed aspects of the survey design (G). (It is interesting to note that Hansen, Hurwitz and Bershada (1961) define the response deviation in relation to the expected value rather than the unknown true value. This is feasible because the unknown true value is only needed for the measurement of bias.)

The response deviations are seen to become correlated through such things as “*an interviewer’s misunderstanding of his instructions, carelessness, or a tendency to introduce his own judgements in a survey*”, causing “*his results to differ from those of other interviewers*” (Hansen, Hurwitz and Bershada, 1961, p. 366). Or more specifically as suggested by Groves (1991, p. 2-3), correlated response deviations resulting from the interviewer can occur because of the interviewer’s “*failure to read the question correctly (leading to response errors by the respondent), delivery of the question with an intonation that influences the respondent’s choice of answer, and failure to record the respondent’s answer correctly.*”

Under the correlation model, the intraclass correlation is defined:

$$\rho_w = \frac{E(d_{hjt}d_{hj't})}{\sigma_d^2}, \quad (10)$$

where h refers to a particular interviewer’s workload (stratum),

$j \neq j'$ refer to different units within the same (t -th) trial, and

$\sigma_d^2 = E(d_{jt}^2)$ is the effect of the variance of the individual response

deviations over all possible trials and all individuals.

The σ_d^2 is often referred to as the simple response variance. The simple response variance is described further in Section 1.2.3.

The simple response variance, σ_d^2 can be broken down as

$$\sigma_d^2 = \sigma_\alpha^2 + \sigma_\varepsilon^2$$

where σ_α^2 = the variance due to interviewers and

σ_ε^2 = remaining error variance.

Thus, using the notation of O'Muircheartaigh (1977), ρ_w can be re-expressed as

$$\rho_w = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2} \quad (11)$$

and $\hat{\rho}_w$ can be estimated as

$$\hat{\rho}_w = \frac{s_\alpha^2}{s_\alpha^2 + s_\varepsilon^2} \quad (12)$$

How does this ρ_w compare with ρ of the ANOVA model? ρ from the ANOVA model as given in equation (4) was

$$\rho = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_b^2}$$

Here we can see σ_a^2 which is the interviewer effect from the ANOVA model is equivalent to

σ_α^2 from the correlation model. However, σ_b^2 is equal to $\sigma_y^2 + \sigma_\varepsilon^2$. Thus, ρ represents the

proportion of total variability which is due to interviewers and ρ_w represents the *proportion of*

response variability which is due to interviewers. Both measures have their value. But it should be noted that only ρ is portable, allowing for comparisons across studies and that $\sigma_y^2 + \sigma_e^2$ can only be separated (and thus ρ_w can only be calculated) when there are multiple trials (i.e., re-enumeration) present in the design.

1.2.3 Simple and Correlated Response Variance

In the presence of response deviations, the total variance of a given mean (\bar{y}) can be expressed as

$$Var(\bar{y}) = E(\bar{y}_t - \bar{y})^2 + 2E(\bar{y}_t - \bar{y})(\bar{y} - \bar{Y}) + E(\bar{y} - \bar{Y})^2 \quad (13)$$

where $E(\bar{y}_t - \bar{y})^2$ = the response variance over trials t and

$E(\bar{y} - \bar{Y})^2$ = the sampling variance.

Note that this formulation for the variance of \bar{y} assumes simple random sampling, sampling with replacement, and no other variable errors and no biases. Appropriate additions can be made to the equation when this does not prove to be the case. For the simplicity of explanation, this assumption will be made for the remainder of this Section.

In turn the response variance can be expressed as

$$\sigma_{d_t}^2 = E(\bar{d}_t^2) = \frac{1}{n} \sigma_d^2 + \frac{n-1}{n} \rho_w \sigma_d^2 \quad (14)$$

where σ_d^2 = the simple response variance (the effect of the variance of the

individual response deviations over all possible trials and all sample individuals),

$\rho_w \sigma_d^2$ = the correlated response variance, and

ρ_w = is a measure of the correlation between response deviations as defined in equation (10).

The response variance is more typically expressed as

$$\sigma_{d_i}^2 = \frac{1}{n} \sigma_d^2 (1 + \rho_w (k - 1)) \quad (15)$$

where k = the average interviewer workload.

Thus either component can be seen to increase the total variance. However, re-enumeration is necessary in the design in order to measure the simple response variance due to interviewers and interpenetration is necessary in the design in order to measure the correlated response variance due to interviewers. And both are necessary in order to estimate both sources of variance (see O’Muircheartaigh, 1977).

In a simple random sample with no survey variable errors or biases the variance of (\bar{y}) is given as

$$Var(\bar{y}) = \frac{\sigma_y^2}{n} \quad (16)$$

In the presence of simple response variance due to interviewers (see O’Muircheartaigh, 1977), the variance needs to be re-expressed as

$$Var(\bar{y}) = \frac{\sigma_y^2 + \sigma_d^2}{n} \quad (17)$$

Note, however, in the case of proportions, the sum of both the sampling variance and simple response variance cannot exceed pq/n (see Hansen, Hurwitz and Bershad, 1961, p 365).

In the presence of correlated response variance, the variance of (\bar{y}) becomes:

$$Var(\bar{y}) = \frac{\sigma_y^2}{n} [1 + \rho_w (k - 1)] \quad (18)$$

Note that either ρ (or ρ_w) can be used and that even small values of ρ (or ρ_w) can have a large effect on the variance of the desired mean or proportion depending on the size of the interviewer's workload (k). This inflation factor due to interviewers is also analogous to a design effect for sampling clustering (see Kish, 1965, and Chapter 9).

Finally, in the presence of both simple and correlated response variance due to interviewers the variance of (\bar{y}) becomes

$$Var(\bar{y}) = \frac{\sigma_y^2}{n} + \frac{\sigma_d^2}{n} [1 + \rho(k - 1)], \quad (19)$$

1.2.4 Further Extensions to the Correlation Model

In his interpenetrated, repeated design, Fellegi (1964) further extends the model of Hansen, Hurwitz and Bershad (1961). Fellegi divided his sample into k sub-samples, denoted by S_1, S_2, \dots, S_k . Each sub-sample was then randomly paired with another sub-sample, so that each sub-sample was selected twice (the re-enumeration). Each pair was then randomly assigned to an interviewer (the interpenetration). Thus each interviewer interviews at two different sub-samples or phrased differently, during the repeat survey, each respondent receives a different interviewer. In addition, to the correlation of response deviations obtained by the same interviewer in the same survey with different units (ρ_w), Fellegi (1964) explores several types of correlation between response deviations (see also O'Muircheartaigh, 1982), i.e., the correlation of response deviations obtained:

- by different interviewers in the same survey (different units),
- in the two surveys for the same unit, different interviewers (zero if both measurements can be considered independent),
- by the same interviewer in different surveys (different units),
- for different units by different interviewers in different surveys, but the same sub-sample, and
- for different units by different interviewers in different surveys and different sub-samples.

Fellegi (1974) further expands the model to provide an estimator of correlated response variance based on the work of all interviewers (i.e., those working in the interpenetrated sub-sample as well as those working in the non-interpenetrated sub-sample.) This results in a estimator of the correlated response variance with a smaller variance than the Hansen, Hurwitz, and Bershada (1961) estimator (see equation (14) above).

1.2.5 Typical Values of ρ , the Intra-Interviewer Correlation Coefficient

Given the need for designs with both interpenetration and re-enumeration for the measurement of ρ_w , and only the need for designs with interpenetration for the measurement of ρ , studies of ρ are much more prevalent in the literature. Also as mentioned in Section 1.2.1, ρ has the advantage of being more portable than either ρ_w or the variance components themselves. Thus ρ values are often used to compare interviewer effects between studies.

This section reviews the literature with respect to ρ . During the past thirty years or so evidence has accumulated about the order of magnitude of ρ for interviewers in sample surveys in the US and elsewhere. Though it is impossible to generalise with complete

confidence, the evidence suggests that values of ρ for interviewers greater than 0.1 are uncommon and the majority of values tend to be less than 0.02. A summary of past research employing an interpenetrated design is given in Table 2. Note that in most cases the standard error of ρ is often large, thus negative values can occur. As suggested by Groves and Magilavy (1986) the standard error of ρ can be reduced by including more interviewers in the design or through a pooled standard error created across replications.

Although the typical values of ρ are extremely small in comparison to what would be considered meaningful values of the Pearson's product moment correlation coefficient, ρ can still have an important effect on the overall precision of survey results. As shown in equation (19), the effect of ρ is seen to multiply the sum of the sampling and simple response variances by a factor of $(1 + \rho (\text{average interviewer workload} - 1))$. Thus, if the value of ρ is as little as "0.04 and the average interviewer workload is $m = 26$, then the total variance is increased by a factor of $(1 + (0.04) (25))$, i.e. the total variance is doubled due to the correlation among response deviations for interviewers" (O'Muircheartaigh, 1977, p. 224-225).

The variability in values of ρ across studies can be due to variability in their standard errors and due to different populations of interviewers. Some researchers have also suggested that ρ will vary in size based on the content and format of the questions used in the questionnaire. This topic is explored in Section 2.3.

Table 2: Summary of Interviewer Variance Investigations Using ρ ‡

<u>Study</u>	<u>Values of ρ_i</u>	<u>Mean</u>
Neighbours noise/illness (UK) Gray (1956)	-0.018 to 0.10 †	0.015 †
TV habits (UK) Gales and Kendall (1957)	(0.00) to 0.05, 0.19 ‡	NA ϕ
Census (US) Hanson and Marks (1958)	-0.00 to 0.061 ‡	0.011 ‡
Blue Collar Workers (US) Kish (1962)		
First study	-0.031 to 0.092	0.020
Second study: Interview	-0.005 to 0.044	0.014
Second study: Self-Completion	-0.024 to 0.040	0.009
Census (Canada) Fellegi (1964)	(0.00) to 0.026	0.008
Health Survey (Canada) Feather (1973)	-0.007 to 0.033	0.006
Mental Retardation (US) Freeman & Butler (1976)	-0.296 to 0.216	0.036
Aircraft Noise (UK) O'Muircheartaigh & Wiggins (1981)	(0.00) to 0.09	0.020
Consumer Attitude Survey (UK) Collins & Butcher (1982)	-0.039 to 0.119	0.013
9 Telephone Surveys (US) Groves and Magilavy (1986)	-0.042 to 0.171	0.009

‡ All studies based on interpenetrated designs.

† Calculated from F-Ratios using formula supplied by Kish (1962).

‡ Numbers available through Kish (1962).

ϕ Mean can not be computed. Paper does not report all variables analysed.

1.3 The Data Source

As described in Section 1.2, an interpenetrated sample design is essential to the study of interviewer variance. Although in studying interviewer and respondent effects, some researchers (see, for example, Hox, 1994), argue that the use of a multilevel regression model with the proper covariates, introduces adequate statistical control to make up for the lack of experimental control.

Given the need for an interpenetrated sample design, the ideal data source for this thesis was the British Household Panel Study (BHPS). The BHPS contains an interpenetrated sample design experiment designed by C. O'Muircheartaigh and myself, when I was a member of the BHPS staff. This design was implemented at Wave 2 of the panel study.

1.3.1 Description of the British Household Panel Study

The BHPS is funded by the UK Economic and Social Research Council. It was designed to enable researchers to describe and analyse how individuals, families and households experience changes in their socio-economic environment and how they act in relation to these changes. Interviewing on the survey began in 1991 and has continued in annual waves with the fieldwork being conducted by NOP Social and Political, London. The survey used a multistage stratified cluster design covering all of Great Britain. The survey instrument comprised a short household level questionnaire followed by a face-to-face 45 minute interview and short self-completion schedule with every adult in the household. These questionnaires covered various substantive areas; household organisation, income and wealth, labour market experience, housing costs and conditions, health issues, consumption behaviour, education and training, and socio-economic values. More information on the objectives of the BHPS can be found in Rose *et al* (1991).

1.3.2 Description of the Interpenetrating Sample Experiment at Wave 2

The interpenetrating design was implemented in a sample of PSUs in Wave 2 of the survey. In order to minimise interviewer travel costs, we suggested a constrained form of randomisation in which addresses were allocated to interviewers at random within ‘geographic pools’. Groves (1989, p. 361) points out that randomisation within geographic clusters has the added advantage of randomising “*assignments among interviewers who in any likely replication of the survey would be given interviewing duties in the same areas.*” Thus in the BHPS case, London interviewers are given London areas and Scottish interviewers are given Scottish areas and so on. This maintains realistic design options and counters the effects that could occur simply from interviewers working in areas in which they are unaccustomed.

These geographic pools consisted of 2 or 3 nearby PSUs. All PSUs whose centroid was a minimum of 10 kilometres away from the centroid of at least one other PSU were eligible for inclusion in the design. One hundred and fifty three of the 250 PSUs in the BHPS sample were eligible. Mutually exclusive and exhaustive combinations of these 153 eligible PSUs were then formed. This process resulted in 70 pools of two or three PSUs each. A systematic sample of 35 pools was then selected for inclusion in the interpenetrating sample design.

Twenty-five of the 35 geographic pools included two interviewers and two PSUs, five included three interviewers and three PSUs, four proved to be ineligible as the same interviewer was needed to cover all of the PSUs in the pool and one proved to be effectively ineligible for analysis as one interviewer was needed to cover 3/4 of the

geographic pool. This resulted in 30 usable geographic pools for analysis (65 interviewers and 65 PSUs). Within a given pool, households were assigned randomly to the interviewers working in those PSUs. For example, this random assignment means that Interviewer A who worked only in PSU 1 in Wave 1, now has a random half of her/his assignment in PSU 1 and a random half in PSU 2 at Wave 2 and Interviewer B who worked only in PSU 2 in Wave 1, now has a random half of her/his assignment in PSU 2 and a random half in PSU 1 at Wave 2, with PSUs 1 and 2 constituting Geographic Pool I. This mixing of workloads can be seen clearly in Figure 3. Thus at Wave 2, the interpenetrated design creates a hierarchical cross-classification in which households or individuals are seen to be nested within the cross-classification of interviewer by PSU, which are in turn nested with geographic pools.

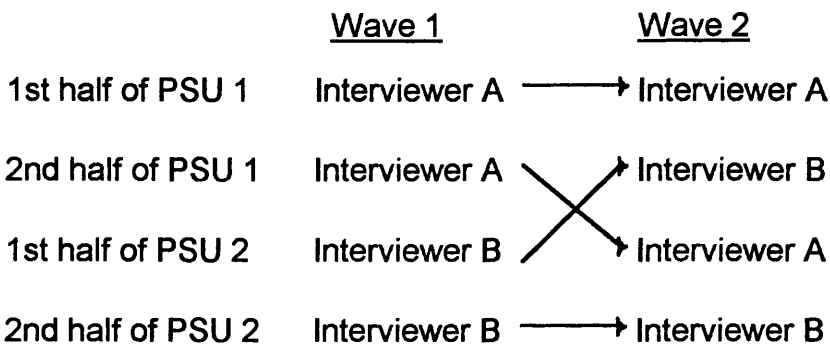
The interpenetrated sample design involved approximately a quarter of the BHPS Wave 2 sample. More specifically, 1406 co-operating and partially co-operating Wave 1 households were issued for Wave 2 fieldwork. Given changes in people's circumstances, this resulted in 1,493 households at Wave 2. 1,283 of these were responding households, 189 were non-responding households and 21 were ineligible. This translated into 2,842 eligible individuals (including both Wave 1 respondents, Wave 1 non-responding individuals in partially co-operating households, and new entrants). Of these, 2,324 co-operated fully at Wave 2 and 518 were proxied or were non-respondents. Restricting the sample to just original Wave 1 sample members who had completed full interviews, the figures are 2,110 and 311, respectively.

As the interpenetration in the survey was implemented at Wave 2, the design implicitly allows for the assessment of the differential effects of sending or not sending back the

same interviewer at a subsequent wave of a panel study (see Figure 3). Throughout this thesis, this will be referred to as the issue of ‘interviewer continuity’.

Figure 3: Interpenetrated Elements of British Household Panel Study Data

Example of a 2 x 2 geographic pool:



1.3.3 Implications for Wave 3

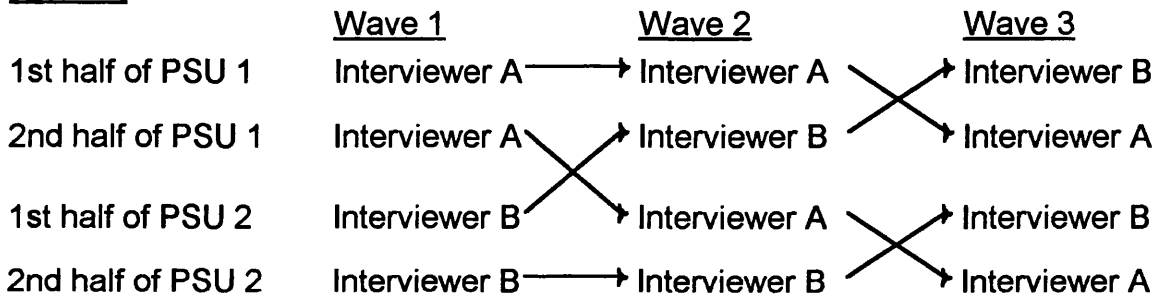
We discussed with BHPS staff the option of implementing an interpenetrated sample design in Wave 3 of the BHPS. First considered was Option 1 (see Figure 4) in which the existing sample was fully crossed at Wave 3. There was also discussion of designing the sample so that it contained half of the Wave 2 interpenetrated areas and half non-interpenetrated areas to bring in a new sample and lessen the number of interviewers who were burdened both years. This option offered an ideal opportunity for replication and an opportunity for longitudinal work.

The proposal was not taken up by the BHPS for two main reasons: (1) There was a concern that switching to a new interviewer for all households in the experiment could lower response

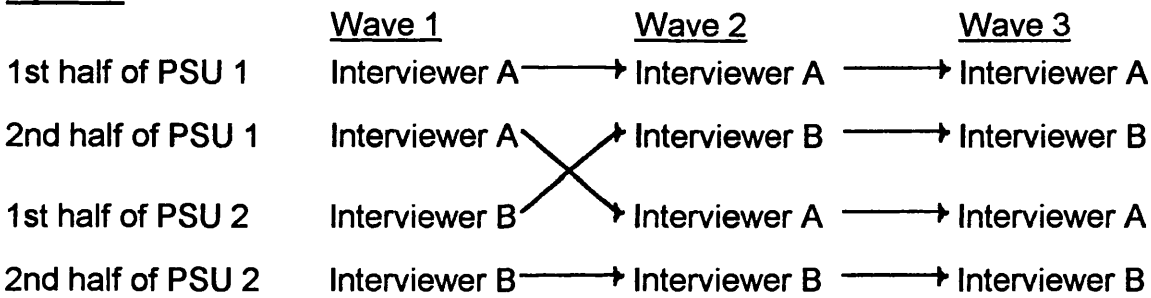
rates and (2) the Wave 2 interpenetrated sample had complicated fieldwork logistics and they were loath to add further complexities.

Figure 4: Interpenetrated Options for Wave 3

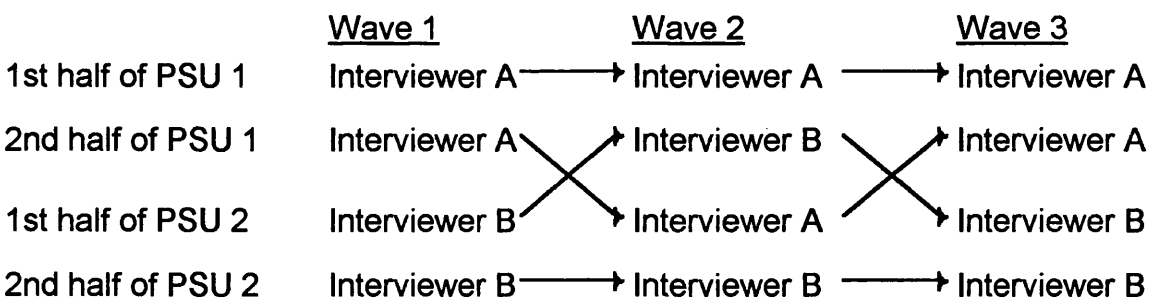
Option 1



Option 2



Option 3



There was still a decision, however, which could be made on how Wave 3 cases could be assigned. These are shown as Options 2 and 3 in Figure 4. Option 2 minimised the two BHPS concerns because the field allocations for Wave 3 replicated Wave 2. In Option 3, the

field allocations for Wave 3 replicated Wave 1. We argued for Option 3. It minimised BHPS concerns but allowed for a randomisation at Wave 3: the cases which were randomly assigned one way in Wave 2 were essentially randomly assigned the other way for Wave 3. The BHPS adopted Option 3.

Although Option 3 again confounds Interviewer and PSU variability it does allow for the continued investigation into interviewer continuity. We can see that a random half of respondents in PSU 1 received Interviewer B at Wave 2 and Interviewer A at Wave 3 and the other random half received Interviewer A on both occasions. Thus at Wave 3 the data can be viewed as a hierarchical cross-classification in which households or individuals are seen to be nested within the cross-classification of interviewer/PSU by interviewer continuity, which are in turn nested within geographic pool.

1.3.4 Auxiliary Data

Auxiliary data to supplement the basic BHPS data were gathered from a number of number of sources. Data on the characteristics of interviewers were generously provided by NOP Social and Political. These included interviewer age, gender, length of service, and interviewer grade level (standard interviewer, supervisor, area manager). We were also able to determine if the same interviewers interviewed the same respondents over time and thus develop an indicator of interviewer continuity.

Data on the characteristics of areas were taken from 1991 Census small area statistics (thanks to special permission of the BHPS who allowed me access to identifiers necessary to conduct the match). Data on Wave 1 and 2 call records were specially keyed by BHPS staff for this project. The BHPS also allowed the inclusion of 5 of our questions on an

interviewer debriefing questionnaire (see Section 8.1.2) so that interviewer's own views on the impact of interviewer continuity could be explored.

1.4 Analysis Considerations for British Household Panel Study Data

1.4.1 Two-way and Three-way Tables

Chi-square tests of independence were used for an initial examination of the categorical predictors of nonresponse in two-way tables. Although the Pearson and the likelihood ratio chi-square statistics are asymptotically equivalent, the Pearson chi-square statistic has better small sample properties. For example, even in tables with expected cell counts between 1 and 4, its "*approximation is 'on average' about right*" (Fienberg, 1985, p. 173). Therefore the Pearson chi-square statistic was preferred over the likelihood ratio chi-square statistic in the two-way tables. The formula is as follows:

$$\text{Pearson chi-square statistic} = \sum \frac{(\text{Observed} - \text{Expected})^2}{(\text{Expected})} \quad (20)$$

with df = (number of rows-1)(number of columns-1)

Loglinear modelling was used in the case of three-way tables (e.g., for an investigation into the randomness of interviewer attrition in Chapter 8). The equation for a saturated model for a three-way table is as follows:

$$\log m_{ijk} = u + u_i + u_j + u_k + u_{ij} + u_{ik} + u_{jk} + u_{ijk} \quad (21)$$

Selecting an appropriate loglinear model for a particular three-way table proceeded by determining the best fitting, yet most parsimonious model. A nested hierarchy of models were considered (e.g., those with just main effects, those with the various two-way interactions). The fit of each model was evaluated with the likelihood ratio goodness-of-fit statistic which is as follows:

$$\text{Likelihood Ratio Goodness-of-Fit Statistic} = 2 \sum (\text{Observed}) \log \left[\frac{(\text{Observed})}{(\text{Expected})} \right] \quad (22)$$

with $df = \text{Number of cells} - \text{Number of parameters estimated}$

Here the simpler models have larger chi-square values, attained levels of significance closer to zero, and thus poorer fit. A fully saturated model, fits the data perfectly. The likelihood ratio chi-square can be partitioned for a comparison of two nested models (model₁ and model₂) as follows:

$$\text{Likelihood Ratio Test Statistic} = 2 \sum (\text{Expected}_1) \log \left[\frac{(\text{Expected})_1}{(\text{Expected})_2} \right] \quad (23)$$

with $df = df \text{ for model}_2 - df \text{ for model}_1$

where model 2 has fewer terms in it than model 1. This allows one “to test whether the difference between the expected values for two nested models is simply due to random variation, given that the true expected values satisfy model 1” (Fienberg, 1985, p. 57). (The Pearson goodness-of-fit statistic does not lend itself to this type of partitioning and therefore is not considered in the loglinear context.) The likelihood ratio test statistic, reflecting the difference between nested models, is reported in the text as the “change χ^2 ”.

A particular problem for loglinear modelling is presented by large sparse multinomial tables as these can contain lots of empty cells. To minimise this problem, categories were collapsed where possible, based on meaningful theoretical distinctions.

1.4.2 Calculation of ρ with Hierarchical Analyses of Variance

Given that the BHPS Wave 2 experimental sub-sample involved interpenetration but not re-enumeration, the ANOVA model for the calculation of ρ was the starting point for this thesis (see Section 1.2.1). The cross-classified, hierarchical nature of the BHPS data, however, meant that this simple approach needed to be modified. Hierarchical analyses of variance were used where the cross-classification of interviewers and PSUs were nested within geographic pools as shown in equation (24). Here the value for the i -th survey element (either individual or household) is $y_{i(jk)l}$. It can be seen that

$$y_{i(jk)l} = \mu + \alpha_l + \beta_{j(l)} + \gamma_{k(l)} + e_{i(jk)l} \quad (24)$$

where μ = the grand mean,

α_l = the effect of the l -th geographic pool, and

$\beta_{j(l)}$ = the effect of the j -th PSU within the l -th geographic pool,

$\gamma_{k(l)}$ = the effect of the k -th interviewer within the l -th geographic pool,

$e_{i(jk)l}$ = an error term representing the i -th survey element within the cross-

classification of PSU by interviewer within the l -th geographic pool.

Note that the subscripts i, j, k, l , have been used here in a manner to keep them consistent with the multilevel model which appears as equation (27). Analysis of variance equations typically assign subscripts in the reverse order, with subscript i assigned to the largest rather than the smallest unit (see, for example, Dunn and Clark, 1974; Winer, 1962). In order to increase comparability with the multilevel models, the decision was made to exclude the interaction of the j -th PSU by k -th interviewer within the l -th geographic pool ($\beta\gamma_{jk(l)}$) from the models used to calculate ρ . (A re-analysis of a test sample of the 820 variables described in Chapter 9 suggested that the prevalence of significant interaction terms was low and that the presence of a significant interaction term introduced only negligible changes in the value of ρ .)

Although a single dependent variable was present in all cases, capturing the nested and cross-classified aspects of the data require implementation through the SPSS MANOVA procedure. For the hierarchical analyses of variance a 'regression approach' was used in which each term is corrected for every other term in the model.¹ A hierarchical analysis of variance approach, although derived from the ANOVA model standardly used in the literature to calculate ρ , has both advantages and disadvantages especially when compared to a cross-classified multilevel regression approach (see Section 1.4.5 for a description of multilevel modelling). First, due to the fact that the vast majority of variables in the BHPS represent categorical or ordinal levels of measurement rather than interval or ratio ones, each category needed to be analysed separately as a proportion. This, however, means that the dependant variables for the hierarchical analyses of variance are dichotomous rather than the needed continuous variables. Despite this violation of assumptions, for proportions between .20 and .80, the model should be fairly robust. Small categories were combined, where possible, to ensure this. Multilevel analyses, on the other hand, allows for the direct use of a dichotomous dependent variable. However in the case of a cross-classified multilevel logistic regression model, the variance components cannot be directly used for the calculation of ρ (see discussion in Section 1.4.7). Second, the ANOVA model traditionally has been used simply to calculate ρ so that the correlated interviewer effect can be estimated (see equation 18) and confidence intervals around means and proportions can be adjusted appropriately. The multilevel analyses, on the other hand, allows one to directly and easily include the hierarchical variance structure in various complex substantive models of interest (see Goldstein, 1995).

There were further advantages and disadvantages with respect to the software programs chosen to implement the models (i.e., SPSS MANOVA for the hierarchical analyses of

1 Note that as our design is not balanced, the sums of squares for the various components of the model will not add up to the total sum of squares.

variance and ML3/MLn/MLwiN for the cross-classified multilevel models - see Section 1.4.5). For example, the SPSS MANOVA program facilitated a quick and efficient exploration of a large number of variables, particularly in comparison with the multilevel software, ML3/MLn/MLwiN). A disadvantage of the SPSS MANOVA program, however, was that it could not simultaneously calculate results from both the 2x2 and 3x3. Therefore the 3x3 geographic pools were eliminated from the analysis resulting in a reduction in sample size of roughly a quarter at both the household level and individual level. In contrast, ML3, MLn, and MLwiN do allow for the simultaneous calculation of both types of geographic pool.

1.4.3 Logistic Regression Models for Data Reduction

As described in Section 1.3.4, numerous auxiliary variables were available to serve as explanatory variables or as control variables, depending on the analysis. As many of the individual, household and area variables were indicators of the same underlying concepts, some type of data reduction step was necessary. We opted for the model selection facilities available in multiple logistic regression in SPSS. As the purpose was to define a good set of control variables, rather than developing the 'best' model, an automated stepwise procedure was deemed adequate.² Models were initially selected via the Wald test. The drawback of the Wald statistic is that if the logistic regression coefficient B is large, its estimated standard error is inflated, resulting in failure to reject the null hypothesis when in fact this should be done (see Menard, 1995). In the models

2 Note that there are typically three major complaints about a stepwise procedure. First, if it is used initially it can lead to mindless predictor selection rather than those based on theory. In the current case, theory had already suggested the variables for inclusion. Second, a stepwise approach produces a single model. However, depending on the interrelationships between variables, there may be a large subset of equally good models. This problem is usually overcome with an 'all possible subsets' regression approach (Neter, Wasserman, and Kutner, 1985). Third, an automated selection procedure attempts to maximise fit, ideally the generalisability of the models should be tested by replicating them on new data. For models, not requiring the cross-classified aspect, the BHPS data offer a solution. There are, for example, the 35 PSUs listed on the initial sampling frame for the interpenetrated sample design experiment which were not selected. These offer a comparable set of new data.

considered, there were no particularly large coefficients (see, for example, Table 5a in Section 5.4.2). For comparison purposes, selection was later repeated using the likelihood ratio test comparing nested models with and without a given explanatory variable. Completely equivalent models were selected.

In terms of evaluating the logistic regression model, SPSS uses -2 times the log likelihood so that tests can be conducted between nested models, such as the full model versus the intercept model. SPSS includes a classification table comparing the predicted probabilities to the observed outcomes. This can be particularly misleading if your dependent variable is less than .1 or greater than .9, because a high level of agreement can be reached simply because of high agreement for the large modal category (e.g., SPSS calculated an agreement rate of 90 percent for the first of the models in Table 5a!), even though the prediction of the rare characteristic of interest may be poor. SPSS also includes a histogram of the predicted probabilities to the observed outcomes which is slightly more helpful than the classification table as it shows exactly where the predicted probabilities fall, not just whether these are greater or less than one-half.

SPSS, however, does not include any of the measures equivalent to R^2 in linear regression to assess the adequacy of the model. Two have been included for this thesis: R_L^2 as suggested by Hosmer and Lemeshow (1989, p. 148) and *Pseudo R*² as suggested by Aldrich and Nelson (1984, p 57).³

$$R_L^2 \approx 100(L_0 - L_p) / L_0 \quad (25)$$

3 Note that there is also some argument for the actual use of R^2 from linear regression (see Menard, 1995).

where L_0 = the likelihood from the intercept model and
 L_p = the likelihood for the full model.

When divided by 100 (as suggested by Menard, 1995, and as used in this thesis), R_L^2 is analogous to R^2 in that it varies between 0 and 1 with 0 indicating that the explanatory variables are making no contribution and 1 indicating perfect prediction. It can be thought of as the proportional reduction in the absolute value of the log-likelihood.

$$\textit{Pseudo } R^2 = c / (N + c) \tag{26}$$

where c = equals the chi-square for comparing the intercept model to the full model, and
 N = the total sample size.

Pseudo R² accounts takes into consideration the sample size and is similar to R_L^2 , but has the disadvantage of never being able to actually reach a value of 1 (see Menard, 1995). Thus it tends to offer lower values than R_L^2 .

1.4.4 Assumptions Behind the Logistic Regression Model

There are many similarities, yet differences between the assumptions of logistic regression analysis and the ordinary least squares approach of linear regression. In both cases, one obviously needs to make sure to that the explanatory variables are continuous or dichotomous, and to the extent possible that these are measured without error and that all relevant explanatory variables are included in the model as well as interactions and that the final model is as parsimonious as possible. In both linear and logistic regression, one

needs to ensure that the explanatory variables are not highly correlated, thus avoiding the problem of multicollinearity.

As suggested by Menard (1995, p. 67-71), in logistic regression one also has to be concerned about zero cell counts and the problem of complete separation. With respect to the first problem, when the odds are 0 or 1 for an entire group of cases for a specific value category of a categorical explanatory variable, *“the result will be a very high estimated standard error for the coefficient associated with that category.”* In contrast, the problem of complete separation occurs if you are too good at predicting the dependent variable with a set of predictors. As Menard points out, this often happens when you have made an error, such as having almost as many explanatory variables as you have cases. It results in extremely large logistic regression coefficients and standard errors.

In linear regression there is the assumption of a linear relationship between the dependent and explanatory variables. In logistic regression the assumption is often referred to as linearity in the logit (see Menard, 1995). To test for violations of this assumption, Hosmer and Lemeshow, (1989, p. 90) suggest the use of a Box-Tidwell transformation which is simply to add the term $x \ln(x)$ to the model. *“If the coefficient for this variable is significant we have evidence for non-linearity in the logit.”*

In linear regression, the residuals are typically checked for a number of concerns, e.g., whether the residuals are normally distributed, whether they have constant variance (homoscedasticity), whether individual errors are related to each other (autocorrelation) or with the dependent variable, and whether there are outliers exerting a strong influence.

Menard (1995, p. 71) suggests that “*the principal purpose for which residual analysis is used in logistic regression is to identify cases for which the model works poorly, or cases that exert more than their share of influence on the estimated parameters of the model.*”

In logistic regression, the distribution of the errors is assumed to be binomial, but thanks to the central limit theorem, these approximate a normal distribution for large samples. The influence of individual cases was considered by searching for cases with large leverage values, ranging from 0 (no influence) to 1 (completely determining the coefficients in the model), large Cook’s distance (an indicator of the overall change in regression estimates attributable to deleting an individual observation), and/or large standardised and studentised residuals, with standardised residuals typically being bigger than the studentised residuals. All three of these indicators don’t necessarily point to the same case as being problematic. For example as described by Hosmer and Lemeshow (1989, p. 154), for probabilities less than 0.1 or greater than 0.9, the leverage decreases and rapidly approaches zero. This suggests that “*the points most extreme in the covariate space may have the smallest leverage. This is the exact opposite of the situation in linear regression.*”

Autocorrelation should be considered in the case of time series data or in other instances where correlation among the error terms is suspected. As Aldrich and Nelson (1984, p. 81) suggest “*when residuals are serially correlated, maximum likelihood estimates remain unbiased in large samples, but they are not efficient.*”

1.4.5 Cross-classified Multilevel Models

The data from the interpenetrated sample design experiment in Wave 2 of the BHPS can be

seen as cross-classified and hierarchical in nature with households/individuals nested within the cross-classification of areas by interviewers nested within geographic pools at Wave 2. In multilevel analysis, the hierarchical partitions and the terms corresponding to them in the model are considered to be random effects. The treatment of the interviewer and PSU effects as random effects rather than as fixed effects postulates a 'superpopulation' of interviewers from which the interviewers used in the study were drawn and an infinitely large population of PSUs. In the case of interviewers we can consider the inference as being made to the population of potential interviewers from whom the survey interviewers were drawn. For the PSUs the assumption involves essentially ignoring the finite population correction (see, for example, Kalton, 1979). For surveys of the size of the BHPS the effect of the finite population correction factor is negligible as it is essentially equal to 1.

It is only recently that cross-classified multilevel analysis has become feasible (see Goldstein, 1995; Rasbash *et al*, 1995; Woodhouse, 1995; Goldstein *et al*, 1998); the design is implemented in the ML3, MLn, and MLwiN software by viewing one member of the cross-classification as an additional level above the other (see Section 1.4.6). Alternative programs for multilevel analysis are VARCL (Longford, 1988) and HLM (Bryk *et al*, 1986).

A basic multilevel model to capture the interviewer by PSU cross-classification within geographic pool can be defined as follows:

$$y_{i(jk)l} = \alpha + \beta x_{i(jk)l} + u_j + u_k + u_l + e_{i(jk)l} \quad (27)$$

for the i -th survey element, within the j -th PSU crossed by the k -th interviewer, within the l -th geographic pool, where $y_{i(jk)l}$ is a function of an appropriate constant (α), an explanatory

variable x and its associated coefficient β , and an individual error term ($e_{i(jk)l}$). Here u_j is a random departure due to PSU j , u_k is a random departure due to interviewer k , and u_l is the random departure due to geographic pool l . Each of these terms and $e_{i(jk)l}$ are random quantities whose means are assumed to be equal to zero. In cases where the dependent variable is a dichotomy we model the log of the odds of the probability of the event (i.e., $\log(\pi_{i(jk)l}/1-\pi_{i(jk)l})$ or $\text{logit}(\pi_{i(jk)l})$). Thus the cross-classified multilevel logistic regression model is:

$$\log(\pi_{i(jk)l}/1-\pi_{i(jk)l}) = \text{logit}(\pi_{i(jk)l}) = \alpha + \beta x_{i(jk)l} + u_j + u_k + u_l \quad (28)$$

Here $\pi_{i(jk)l}$ represents the probability of the event. Alternatively $\pi_{i(jk)l}$ can be expressed as follows:

$$\pi_{i(jk)l} = \frac{\exp(\alpha + \beta x_{i(jk)l} + u_j + u_k + u_l)}{1 + \exp(\alpha + \beta x_{i(jk)l} + u_j + u_k + u_l)} \quad (29)$$

In the case of multinomial models, the dependent variable is now a vector of t proportions. As one is chosen as the base category, $t-1$ equations of the form of equation (28) are necessary.

For example, in Chapter 7 the three category dependent variable for nonresponse is considered (i.e., interviews, refusals, and non-contacts; coded 0, 1, and 2, respectively). Interviews is

chosen as the base category. Thus the two equations are:

$$\log [P(y_{i(jk)l} = 1)/P(y_{i(jk)l} = 0)] = \alpha^1 + \beta x_{i(jk)l}^1 + u_j^1 + u_k^1 + u_l^1 = g_1 \quad (30)$$

$$\log [P(y_{i(jk)l} = 2)/P(y_{i(jk)l} = 0)] = \alpha^2 + \beta x_{i(jk)l}^2 + u_j^2 + u_k^2 + u_l^2 = g_2 \quad (31)$$

Using the shorthand notation g_1 and g_2 to represent equations (30) and (31), the probabilities of each of the three events can be constructed as follows:

$$P(y_{i(j)kl} = 0) = \frac{1}{1 + e^{g_1} + e^{g_2}} \quad (32)$$

$$P(y_{i(j)kl} = 1) = \frac{e^{g_1}}{1 + e^{g_1} + e^{g_2}} \quad (33)$$

$$P(y_{i(j)kl} = 2) = \frac{e^{g_2}}{1 + e^{g_1} + e^{g_2}} \quad (34)$$

Multilevel models have a natural congruence with many important aspects of the survey situation; both the sample design and the fieldwork implementation can be described appropriately as introducing *hierarchical levels* into the data and thus multilevel analysis provides a framework. Of equal importance is the facility to incorporate substantive covariates directly into the analysis, thereby making it possible to include both substantive and design factors in the same analysis. The use of covariates in multilevel modelling with respect to interviewer effects has recently been demonstrated (c.f. Hox, de Leeuw, and Kreft, 1991; Wiggins, Longford and O'Muircheartaigh, 1992; Ecob and Jamieson, 1992). (For this thesis numerous covariates are available including the characteristics of households and individuals (see Section 1.3.1) as well as data on the characteristics of interviewers, the characteristics of areas, and the patterns of respondents co-operation and contactability from call record data (see Section 1.3.4). Despite the focus on multilevel modelling in this thesis, it should be noted that this thesis is not about multilevel modelling per se, but rather about how multilevel modelling offers a valuable tool for exploring the effects of interest.

1.4.6 Implementation of the Cross-Classified Multilevel Models

ML3/MLn/MLwiN requires two levels for the estimation of a cross-classified structure. For example, a model which involves households nested within the cross-classification of

interviewers by PSUs is a two level model conceptually, however, ML3/MLn/MLwiN uses the third level as a device to allow estimation (see, for example Goldstein *et al*, 1998). Thus within ML3, which has a maximum of three levels, level one can only be allocated to households or to individuals. In the individual level analyses, the option to also include the household level is not available. At the start of this research this limitation was accepted. From a theoretical standpoint we expected households and individuals to behave differently and to require different types of explanatory variables. Thus keeping the household and individual level analyses separate was considered reasonable, however, there does remain a concern that household variation should be included in the individual level analyses. The other concern with ML3 was how to include the term for the variance of the geographic pools. In a household level model with the PSU by interviewer cross-classification within geographic pools, the geographic pool variance should ideally be placed at level 4. J. Rasbash of the Multilevel Models Project team suggested incorporating the geographic pool variance through a separate term added to level 3. This works particularly well with the BHPS interpenetrated sample data because the cross-classification is within geographic pools. In this case, ML3 can be used to search for the pools via the XSEA command and the geographic pool codes can then be used to define a block diagonal structure for the variance-covariance matrix at level 3 (see Goldstein *et al*, 1998).

The majority of analyses in this thesis proceeded within these constraints (see also Section 1.4.9). The issue of household variation within individual level models, however, is examined further in Section 6.4 with the use of MLwiN. The issue of the placement of the geographic pool variation was also revisited using MLwiN. These analyses (although not shown) suggested that essentially identical results are achieved whether the geographic pool variation is placed at level 3 or level 4.

1.4.7 Calculation of ρ in the Cross-classified Multilevel Models

When the dependent variable is continuous, ρ can be calculated directly from the variance estimates in a variance components model (e.g., interviewer variance divided by total variance). When the dependent variable is dichotomous, the variance components are given on the logistic scale and a course of action for the calculation of ρ is unclear. In ML3 (MLn and MLwiN) the level one variance is constrained to 1.0 (which is a scale factor rather than a variance). As the variance of a binomial distribution depends on its mean, Hedeker (1993) has suggested that the variance of the binomial distribution can be standardised as $\pi^2 / 3$ (i.e., 3.29). Applying this to multilevel logistic regression, this would suggest standardising the level 1 residual variance at 3.29, assuming of course no over or under-dispersion. If one could combine this standardised level 1 value (which is based on the binomial distribution) with the higher level terms (which have a normal distribution), then the calculation of ρ could then proceed as in the case of the linear model. Further investigation along this line, however, is necessary. In addition to the concerns about the actual calculation of ρ , the accuracy of the standard errors of the random parameters can be improved through a some type of simulation (see, for example, Goldstein *et al*, 1998).

1.4.8 Technical Aspects of the Cross-classified Multilevel Models

The distributional assumption that the Level 1 (individual or household) variance is binomial can be easily tested in MLwiN by relaxing the constraint that $\sigma_e^2 = 1.0$ and comparing the confidence interval around the resulting value to 1.0. As suggested by McCullagh and Nelder (1983), over-dispersion is more common than under-dispersion in practice. Goldstein (1995) comments that over-dispersion typically occurs when a level is

ignored, such as household clustering in a survey sample of individuals and he suggests several ways of modelling such extra binomial variation.

Iterative Generalised Least Squares (IGLS) alternates between random and fixed parameter estimation until the procedure converges. As suggested by Goldstein (1995, p. 23), this “*maximum likelihood procedure produces biased estimates of the random parameters because it takes no account of the sampling variation of the fixed parameters.*” This is clearly important in small samples. MLwiN has a Restricted Iterative Generalised Least Squares (RIGLS) option to produce these modified estimates. The RIGLS option was used throughout.

A binary dependent variable requires a non-linear model. MLwiN offers two estimation options. Marginal quasi-likelihood (MQL) uses just the fixed part of the model in estimating linearisation through the Taylor series expansion. Penalised or predictive quasi-likelihood (PQL) adds in estimated higher level residuals to the linear component of the non-linear function when forming the Taylor series expansion. “*In many applications, the MQL procedure will tend to underestimate the values of both the fixed and random parameters*” (Goldstein, 1995, p. 99).

The MQL and PQL approaches use only the first-order terms from the linear expansion. “*Greater accuracy is to be expected if the second-order approximation is used rather than the first-order upon the first term in the Taylor expansion*” (Goldstein, 1995, p. 99). The PQL and second-order options are particularly needed when the predicted probabilities are extreme, or where there are few level 1 units per level 2 unit or few level 2 units with large numbers of level one units. The first concern appears to be a problem

with the BHPS interpenetrated sub-sample. For example, the proportion of refusals and the proportion of non-contacts are both less than 10%, whereas the data structure is well balanced with 30 geographic pools each containing 2 or 3 interviewers and 33 to 74 households. Therefore PQL and second-order estimation were chosen throughout, where this was feasible.⁴ As Woodhouse (1995, p. 92) points out “*the PQL and second order approximation options, when they converge, produce more statistically efficient, less biased estimates than the MQL and first order approximation options. However, the PQL and second order procedures are less computationally robust than their simpler counterparts.*”

Rough confidence intervals for the random parameters can be constructed using their standard errors. A better option is to use MLwiN’s ‘Intervals and Tests’ window which provides separate and simultaneous Wald tests. For precise inference, one should use Markov Chain Monte Carlo (MCMC) methods and Bootstrapping. In version 1.0 of MLwiN, MCMC methods are not available for multinomial models. It was also impossible to use the MCMC approach with the cross-classified multilevel logistic regression models used in this thesis because the structure of the data as a series of geographic pools created too many structural zeros. The Bootstrapping available in Version 1.0 of MLwiN is still experimental and crashed rather than coping with the cross-classified data.

4 At first it was impossible to use the second-order approximation with the multinomial models. Advice from the Multilevel Models Team at the Institute of Education was to alter the second order macro so that ‘comparative variances’ of the residuals were used rather than ‘adjusted comparative variances’. After this change, second-order estimation for the multinomial models was possible, although convergence in the case of the household level variance components models was not achieved.

The level 1 errors are assumed to follow a binomial distribution which approximates a normal distribution in large samples. Residuals for higher levels are assumed to be normally distributed.

In the case of binary dependent variables we also have the choice of three link functions: the logit link function, a complementary log-log link function, and the probit (normal) link function. McCullagh and Nelder (1983) suggest that logit and probit link functions are very similar for $0.1 < \pi < 0.9$. In general the probit tails are slightly thinner than the logistic tails and the estimates in logistic models are about 1.6 to 1.8 times those in the probit models (Agresti, 1990). For small values of π , the complementary log-log function is close to the logistic, but as π approaches 1, it tends much more slowly to infinity than either the logit or probit transforms (McCullagh and Nelder, 1983).

Throughout we have chosen the logit link function (1) because of its simple interpretation as the logarithm of the odds ratio (in contrast to the log-log link function)

1.4.9 Availability of British Household Panel Study Data and Multilevel Software

This thesis has had a 6 year gestation period. Over this period its growth has been highly dependent on the availability of BHPS data and the development of multilevel software (i.e., ML3, MLn, MLwiN) by the Multilevel Models Project Team at the Institute of Education, London. For example, initial modelling was limited to Wave 2 household level nonresponse variables, because this was the only data that were ready prior to the archive release version. During the summer of 1993, I assisted the BHPS staff with substantial cleaning of the interpenetrated sample design indicator and interviewer identification variables. ML3 was the current version of the multilevel software at that time. The macros for the logistic regression options were a separate feature and J. Rasbash from the Multilevel Models Project team was

just beginning the development of the cross-classified option which at that time was not fully integrated with the logistic regression macros.

Over the years, access to the data and the proper software have increased greatly. Take, for example, the current cross-wave identifier files available for Waves 1 through 6 of the BHPS on CD rom. This made the longitudinal interviewer continuity analyses described in Chapter 8 feasible. Other key developments were the MLn and more recently MLwiN versions of the software. Along with this growth in access has been growth in the sophistication of analysis options for this thesis. For example, the cross-classified multilevel multinomial models described in Chapter 7 represent work from the summer of 1998. Ironically, these models were slightly ahead of the software. J. Rasbash from the Multilevel Models Project Team had to check that the macros for the multinomial models were compatible with the macros for the cross-classified part as he believed that no one had tried this combination of macros before.

1.4.10 Options to Replicate Fellegi (1964)

As discussed in Section 1.2, Fellegi (1964) introduced an element of replication in addition to interpenetrated samples. Given that the BHPS is a panel study with repeated measurements on the same individuals, Colm O'Muircheartaigh and myself discussed the possibility of re-visiting Fellegi's work. In the end, we decided against this route in favour of the others which are outlined in Section 1.5.

The primary reason was the fact that

- There was the year interval between enumerations, an interval long enough for true change to have occurred on most variables, except for certain facts which should be fairly impervious to interviewer effects such as gender.

Other complications were as follows:

- There was respondent attrition between the enumeration and the re-enumeration,
- The BHPS suffered from interviewer attrition between waves,
- The BHPS used a multistage stratified cluster design, not a simple random sample,
- Each geographic pool would have had to be handled separately to bring the BHPS data in line with Fellegi's assumptions, and
- The BHPS subsample size, n , is variable not a constant.

1.4.11 The Use of the Interpenetrated Sub-Sample

The majority of analysis in this thesis is based on this interpenetrated sub-sample, although the full sample has been consulted for the work on interviewer continuity (see Chapter 8).

No weighting has been employed. The complex weighting variables developed for the BHPS are not applicable. For example, the 65 PSUs contained in the final interpenetrated sub-sample generalise to the sub-sample of 153 PSU which were eligible for interpenetration (these form the original 70 geographic pools described in Section 1.3.2). The full BHPS sample represents 250 PSUs. Evidence from a comparison of the interpenetrated sub-sample to the full sample suggested that there was a slight urban bias.

For example, although at least one geographic pool from the interpenetrating sample is found in 16 of the 18 regions, proportionately more interpenetrated sample cases than non-interpenetrated cases are found in inner and outer London, the West Midlands Conurbation, greater Manchester, Merseyside, the rest of the Northwest, south Yorkshire, and Tyne and Wear. Tables comparing the interpenetrated sample to the full BHPS sample are shown in Appendix A for selected individual and household level variables. These included gender, age, marital status, employment status, socio-economic

grade, housing tenure, number of cars, number of pensioners, type of accommodation, and net monthly housing costs. The most striking differences are found with type of accommodation. In keeping with the idea of a population density bias, the interpenetrated sample under-represents detached houses and bungalows and over-represents flats in comparison to the full sample. The remaining differences are small, but again reflective of an urban/rural divide. The interpenetrated sample is seen to have fewer elderly and pensioners, fewer people looking after home and family and more never married individuals, individuals in non-manual occupations and those with higher housing costs.

1.5 Specific Aims of the Thesis

This thesis looks at the influence of the interviewer as a source of response variance and as a contributing factor to nonresponse bias and variance. This investigation has many practical as well as theoretical uses. In the United Kingdom, for example, interviewer-based data collection is the norm on social and market research surveys and is likely to remain so for the foreseeable future.

The specific aims of the study interacted with the options available in the chosen data source as the BHPS dataset offers some unique opportunities for analysis.

1. The interpenetrated design allows the opportunity to separate interviewer effects from area effects. This is a rare opportunity because survey research in the UK typically assigns one interviewer to one PSU and thus the two sets of effects are confounded.
2. Implementation of the interpenetrated design at Wave 2 allows for a study of the correlates of attrition nonresponse because Wave 1 data are available for both Wave 2 respondents and nonrespondents.

3. The random cross-over enacted in the interpenetrated design allows for the experimental testing of the conventional (but rarely tested) survey wisdom that sending back the same interviewer in subsequent waves of a panel study is highly beneficial for maintaining high response rates and good response quality. I'm referring to this as the 'interviewer continuity' issue.
4. The Wave 3 design chosen (i.e., a random cross-over back to the Wave 1 design) allows for a continued investigation into the 'interviewer continuity' issue.

Thus, with respect to nonresponse, the specific aims are:

- to separate interviewer effects from area effects on nonresponse at Wave 2 (see Chapter 6),
- to explore combined interviewer/PSU effects at Wave 3 and see how these differ from Wave 2 (see Chapter 7),
- to assess experimentally the impact of interviewer continuity at Wave 2 (see Chapter 8), and
- to assess experimentally the impact of interviewer continuity at Wave 3 and see how this compares to the full sample results (Chapter 8).

These are facilitated through a general examination of the correlates of nonresponse (see Chapter 5).

With respect to response error, the specific aims are:

- to measure the complex variance due to both interviewers and areas (see Chapter 9), and
- to explore the relationships between interviewer continuity and response quality (see Chapter 10).

As described in Section 1.1, Chapters 2, 3 and 4 are devoted to reviews of the relevant literature on the interviewer, nonresponse, and response error, respectively. In addition, literature reviews for specific sub-topics (such as interviewer continuity) are found in their respective chapters.

The work of the thesis as a whole is then summarised in Chapter 11 and implications for survey research practice are discussed.

1.6 Chapter Summary

This Chapter has introduced the concept of total survey error and explained in detail the statistical measurement of simple and correlated variance due to interviewers as part of total survey error. The data source, the BHPS was described and various design decisions and various options for analysis were discussed, including the use of multilevel statistical models. This chapter ended with an elaboration of the major opportunities for analysis in the BHPS and the resulting specific aims of this thesis.

CHAPTER 2 SURVEY RESEARCH INTERVIEWERS

2.1 The Role of the Interviewer

As Kahn and Cannell (1957, p. vi) suggest, there are many disciplines in which interviewing is either the major professional technique or an important auxiliary skill.

These include everything from journalism and medicine to law and social work. Thus, a common view about interviewing is that it is

“ . . . a sort of battle of wits between respondent and interviewer. The respondent has somewhere inside him information which the interviewer wants, and the interviewing techniques are designed to force, trick, or cajole the respondent into releasing the information.”

The focus of this thesis is the applied area of survey research. The survey research literature as well as the firsthand experience of interviewers show that this common view about interviewing is a misleading and inaccurate conception for survey research interviews. As Kahn and Cannell (1957, p. vi) describe,

“ . . . the [survey] interview is an interaction between the interviewer and respondent in which both participants share . . . The end product of the interview is a result of this interaction.”

But Kahn and Cannell (1957, p. vi) go on to point out that . . .

Therein lie the strengths and weaknesses of the interview as an information-getting technique. If the interaction is handled properly, the interview becomes a powerful technique, capable of developing accurate information and getting access to material otherwise unavailable. Improperly handled, the interaction becomes a serious source of bias, restricting or distorting the flow of communications.”

Thus, we see that the interviewer can indeed be a double-edged sword, with the potential to either minimise or create survey error. Thus the first goal of any interviewer training is to ensure, to the greatest extent possible, that the interviewer is *a neutral collector of accurate data*. The professional survey interviewer has other tasks as well. These include: *to educate the respondent as to his/her role, to conduct the last stages of sampling, to fulfil*

administrative duties, and to gain co-operation with the respondent. Each of these areas will be discussed in turn.

2.1.1 A Neutral Collector of Accurate Data

In survey interviews, the key is standardisation of the measurement process. As Fowler and Mangione (1990, p.14) suggest:

“The goal of standardization is that each respondent be exposed to the same question experience, and that the recording of answer be the same, too, so that any differences in the answers can be correctly interpreted as reflecting differences between respondents rather than differences in the process that produced the answer.”

Thus, a prime area for concern is exactly how the interviewer uses the questionnaire, handles the respondent's answers, and, of course, how the interviewer manages the general interaction with the respondent. Each of these will be discussed in turn.

Handling the Questionnaire. As suggested by various interviewer manuals (see, for example, Survey Research Center, 1976; McCrossan, 1991; Social and Community Planning Research, 1995), interviewers are instructed to:

1. Ask the questions exactly as they are worded in the questionnaire,
2. Read each question very slowly,
3. Ask the questions in the order in which they are presented in the questionnaire,
4. Ask every question specified in the questionnaire,
5. Repeat questions which are misunderstood or misinterpreted, and
6. Probe in a non-directive manner.⁵

5 Despite the case for standardisation (described above) and the numerous possibilities for interviewer error without standardisation (described below), it is worth noting that the survey interview has actually come under criticism because of its standardisation. Viewed as a highly

These points seem obvious when the goal is a standardised measurement instrument, but as suggested by Fowler and Mangione (1990, p. 16), “*it is one thing to say that interviewers should be standardized; it is not so easy to accomplish.*” Let’s look at this more closely. With respect to Task 1, research has found (see, for example, Bradburn and Sudman, 1979; 1991; Schuman and Presser, 1981; Kalton and Schuman, 1982; Converse and Presser, 1986; Fowler, 1995, Sudman, Bradburn, and Schwarz, 1996, among others), that even small changes to question wording can produce different results. For example (see Survey Research Centre, 1976, p. 11), if the question in the questionnaire is “*Where do you get most of your news about current events in this country – from the radio, the newspapers, TV, or talking to people?*” and the interviewer only reads “*Where do you get most of your news about current events?*” because the respondent interrupts the interviewer at that point with his/her answer. The frame of reference for this respondent is clearly different from that of respondents who are given the prompts, “*radio, the newspapers, TV, or talking to people?*”.

Researchers make Task 2 difficult for interviewers when interviewers’ pay or other incentives are contingent on the number of interviews they complete. There are also times when interviewers will feel very uncomfortable complying with Task 4. Take, for

stylised conversation, the critics (see, for example, Briggs, 1986; Mishler, 1986; Suchman and Jordon, 1990) argue that standardisation does not allow for the “full resources of conversational interaction”. For example, in true conversation, meaning is typically negotiated over multiple turns between the participants. The critics argue that the goal of the interview should be standardised question meaning, rather than standardised question wording and that this can only be accomplished through ‘flexible’ interviewing. Schaeffer (1991) and Beatty (1995) suggest that both positions, standardised interviewing and flexible interviewing, have merit. Schober and Conrad (1997) were the first to conduct a systematic comparison of the two. They found that the flexible interviewing approach was only needed for survey items where the “*mapping between the question and the respondent’s situation was complicated*” (p.595). In such cases, the flexible interviewing actually “*increased accuracy by almost 60 percent*” (p. 577). But they also found that “*this accuracy came at a real cost – a more than threefold increase in duration*” (p. 595) although they point out that their “*flexible interviews may be longer than they would be in actual practice*” (p. 595).

example, the first four questions which are asked of respondents (see Figure 5) in the pre-redesign version of the U.S. Current Population Survey (see Campanelli, Martin, and Rothgeb, 1991) and imagine how the interviewer feels asking each subsequent question when the respondent reveals in answer to Question 19 that he is a 92 year old pensioner who hasn't worked in years. (To relieve such uncomfortable situations, interviewers are allowed to preface subsequent questions with statements like, "*I believe you may have answered this, but can I just check . . .*" or "*To ensure that we collect the most accurate information, my office requires me to ask every question.*")

Figure 5: Excerpt from U.S. Current Population Survey, October 1980 Version

19. What was . . . doing most of LAST WEEK -
Working
Keeping house
Going to school
or something else?
 20. Did . . . do any work LAST WEEK, not counting work around the house?
 21. Did . . . have a job or business for which he/she was temporarily absent or on layoff LAST WEEK?
 22. Has . . . been looking for work during the past 4 weeks?
-

Probing in a non-directive manner (Task 6) can also be more elusive than straightforward at times. Take, for example, the same example question used to illustrate Task 1 (from Survey Research Center, 1976, p. 11), but this time the interviewer follows it with a probe so that the full question now becomes, "*Where do you get most of your news about current events in this country – from the radio, the newspapers, TV, or talking to people? That is, which one do you rely on most?*" Although intended simply to add a conversational feel to the interaction, this probe changes the meaning of the question.

Some organisations (see Survey Research Centre, 1976) actually instruct interviewers to record all of the probes which they used so that office coders can determine if the probe influenced the respondent's reply.

Handling Respondents' Answers. In addition to asking the questions, interviewers are also instructed about how to record the answers they receive from the respondents (i.e., writing a verbatim, rather than summarised or paraphrased version, of respondents' answers to open questions; being careful to circle the right code; not assuming knowledge of what respondents' answers will be without asking the question; probing for enough information to ensure accuracy; and so on).

Managing the Respondent (Rapport). The survey interview is a highly stylised form of conversation. Yet it is still a conversation and as such, interviewers need to look at the respondent and to give him/her encouragement and attention if the respondent's interest and concentration are to be maintained during a lengthy interview (see McCrossan, 1991; Survey Research Center, 1976). While maintaining this rapport, interviewers need to reinforce the right kind of respondent behaviour (see Section 2.1.2) and avoid saying anything which could influence respondents' answers (e.g., reserving any comments about themselves until after the interview). Research suggests that up to a certain level such rapport should positively affect survey quality (Harkess and Warren, 1993). The impact of rapport is illustrated by the following comment of a face-to-face panel interviewer:

*"[interview] relationships matter. People are then more patient with you. They remember more and try to be accurate. They are also able to tell you personal details, which otherwise they might avoid."*⁶

6 From the British Household Panel Study Interviewer Debriefing Study implemented as part of this thesis (see Section 7.2.2).

The hypothesised relationship between rapport and validity in interviews, however, is a curvilinear one. Initially, as rapport increases so does validity. There is a point, however, where this levels off. From that point onwards increasing rapport can actually lead to poorer response quality. Examples of bias due to over-rapport can be found in Section 2.2.1. The topic of rapport will also be revisited in Chapter 8 when the effects of interviewer continuity are explored.

2.1.2 Educating the Respondent as to His/Her Role

As Cannell, Miller and Oksenberg (1981, p. 401) suggest,

“The combination of a non-directive approach and rapport is supposed to reduce response error by removing the pressure from the respondent to maintain a totally positive self-presentation and by motivating the respondent to work hard. Overall, the respondent is supposed to get the message that it is all right to reveal himself or herself accurately.”

They go on to say that

“The problem with this notion is that the neutral, or non-directive, interviewer style frequently does not sufficiently motivate or inform respondents . . . The cues to the respondent for deciding when an adequate response has been rendered are obscure; and they are likely to remain so if the interviewer sticks to the simple rules of being non-directive and friendly.”

After getting respondents' feedback on National Health Survey interviews, Cannell, Fowler, and Marquis (1968) were surprised at how many respondents answered the survey but claimed not to know who it was for and didn't know that exact and complete information was needed. In addition, through detailed analysis of tape-recorded interviews, Marquis and Cannell (1969) found that interviewers delivered positive feedback (e.g., *“that's okay”*, *“all right”*, *“you're doing fine”*, etc.) indiscriminately and actually were seen to reinforce some of the worst respondent behaviour (e.g., refusal to answer a question). As described in Cannell, Miller, and Oksenberg (1981, p.412), the authors of the earlier studies investigated a number of techniques to *“deliberately*

manipulate the communication from interviewer to respondent in line with their expectations about response error.” The techniques included the use of a combination of *instructions* to educate the respondent about his/her role, proper contingent *feedback* to reinforce the right kind of respondent behaviour, and *commitment*. *Commitment* was used to increase a respondent’s motivation to perform well. The commitment technique involved having the respondent sign a printed agreement (like a formal contract) in which he/she committed to providing accurate, complete information etc. In a second experiment to evaluate these new techniques, Oksenberg, Vinokur, and Cannell, (1979) found that each of the three techniques led to improvement in respondent reporting. The best improvement was found with a combination of all three: instructions, feedback, and commitment.

2.1.3 Conducting the Last Stages of Sampling

This aspect of the interviewer’s role occurs in both market research and social surveys, but takes different forms. For example, market research surveys (in the United Kingdom and the United States) typically employ sample designs in which the initial stage(s) of sampling is often based on probability sampling, but the final stage which is implemented by the interviewer, is based on quota sampling. In a quota sample interviewers are instructed to follow a predetermined quota to ensure what appears to be a representative sample (e.g., to find so many men, so many women, so many elderly persons, so many unemployed persons, and so on). In social and academic surveys, probability sampling is typically used throughout. In the UK, the face-to-face interviewer conducting a household survey often has two different sampling roles. For example, when the UK Postcode Address File is used and multiple households are found at a selected address, interviewers are given instructions on how to randomly choose a certain proportion of these for interview.

When only one member of the household is required for interview, interviewers are given instructions on how to randomly choose one individual from the household.

2.1.4 Fulfilling Administrative Duties

The specifics of these depend on the given survey organisation but essentially consist of filling in a series of forms to allow the head office to assess the disposition of a particular survey case.

2.1.5 Gaining the Co-operation of the Respondent

In probability surveys, this is a critically important aspect of the survey interviewer's tasks, as only the selected respondent can be interviewed, no other. Substitutions are frowned on in U.S. and U.K. surveys, but are sometimes used in other countries, e.g., Slovenia (see, Vehovar, 1996). Although a case can be argued mathematically for substitutions in probability samples (see Lessler and Kalsbeek, 1992), these arguments typically break down in practice. Interviewers often don't work as hard to obtain a response from a given household the first time around, if they know that substitutions are allowed. Thus a bias can result toward over-sampling the co-operative and the available.

The probability interviewer's role in gaining co-operation can be divided into two main components: countering refusals through persuading potential respondents to take part and countering 'non-contacts' by being persistent in following up on those who can not be contacted initially. In the quota samples of the market research interviewer, the interviewer simply fulfils her or his quota, without the effort to counter refusals or to reach the hard-to-find, thereby potentially allowing and magnifying the bias that can be found with the use of substitutions (see, Lynn, 1995).

Minimising Refusals. The key to countering refusals is the interviewer's 'doorstep' introduction. Although the term 'doorstep' arose from face-to-face household surveys, it refers more generally to the short interaction between interviewer and respondent prior to the interview itself in which the respondent is persuaded to take part. The term is thus equally applicable to telephone, face-to-face home, and face-to-face street interview situations. Although the advice that follows is based on face-to-face home interviews, the key points can be useful in other 'doorstep' situations. (What follows is this sub-section is a general review of the interviewer's role in the two main components of nonresponse, refusals and non-contacts. A more thorough review of nonresponse error is given in Chapter 3.)

Through her study of tape-recorded doorstep introductions, Morton-Williams (1993) offers the following advice to interviewers on their doorstep introduction:

- Prepare in advance
(be familiar with purpose of the survey, survey procedures, feel confident),
- Keep the initial introduction brief
(incorporate prescribed items such as giving one's name, showing one's id card),
- How you present yourself is important
(e.g., smile, make eye contact, have a warm friendly confident manner),
- Accurately observe the contacted person
(look at verbal and non-verbal cues; put yourself in his/her place),
- Pre-empt/answer the respondent's reluctance
(Be reassuring and avoid being threatening),
- Be positive about the benefits of taking part

(Appeal to altruism, indicate the value of the survey, ask for help in carrying it out),

- Address respondents' reluctance with succinct, appropriate information

(Long arguments and pressurising tactics should be avoided),

- Reluctance or refusal is often situation specific

(withdraw and re-approach), and

- There is no one right way to do a doorstep introduction

(be individual and flexible).

In their work with tape-recorded doorstep introductions of face-to-face home interviews, Campanelli, Sturgis, and Purdon (1997) and Sturgis and Campanelli (1998) grouped their advice on good interviewer practice into three categories: 'idiosyncratic strategies that work for the particular interviewer' (e.g., interviewer who believes that being older makes her perfectly convincing), 'idiosyncratic strategies that can be used more broadly' (e.g., shifting the interaction from an 'official' to more a 'personal' kind, 'selling' the survey to other household members, and most importantly avoiding yes/no questions where the respondent can easily answer 'no') and 'general strategies'. The advice given under their general strategies is similar to Morton-Williams (1993). For example, their general suggestions to interviewers were as follows:

- Be prepared with good answers to the most common respondent questions,
- Make sure your response is clear, coherent and to the point,
- Make sure your response is relevant and appropriate to the respondent's specific concerns (i.e., in the language of Groves and Couper (1996), 'tailor' your response to the respondent's concerns),
- Maintain the interaction within and across calls to build rapport and receive clues, and
- Always be prepared to withdraw and try again at a later point in time.

Interestingly, the kind of interviewer who is good at 'doorstep' persuasion is often not the kind of interviewer who is good at being a neutral data collector. As Converse and Schuman (1974, p. 60-61) suggest,

"Interviewers are required to be both technically standardised and interpersonally responsive to many different kinds of individuals - a blend of styles that is often contradictory enough to require special modes of compromise and strain."

Minimising Non-contacts. Reducing the non-contact portion of nonresponse depends on the number of calls which the interviewers make. For example, for large-scale professional surveys in the UK, a rule suggesting a *maximum* of four interviewer call-backs call lead to an 11 percent non-contact rate, whereas a rule suggesting a *minimum* of four interviewer call-backs can lead to a 4 percent non-contact rate (see, for example, Campanelli, Sturgis, and Purdon, 1997; Purdon, Campanelli, and Sturgis, 1998; Lievesley, 1986).

Reducing the non-contact portion of nonresponse also depends on the timing of the calls. Interviewers are typically given some basic instruction about when to call. For example, the Social and Community Planning Research Interviewers' Manual (1995, p. 13) suggests that,

"An address or a selected person cannot be coded as a 'non-contact' until at least four calls have been made at that address. These must be at different times of the day including at least one evening call, and on different days of the week including at least one call at the weekend."

A number of papers have appeared over recent years detailing how the timing of calls to an address influences the probability of contact (see, for example, Weber and Burt, 1972; Weeks *et al*, 1980; Vigerhous, 1981; Swires-Hennessy and Drake, 1992; Campanelli,

Purdon, and Sturgis, 1997; and Purdon, Campanelli, and Sturgis, 1998). Weeks *et al* (1980), for example, found that in a 1975 US survey conducted in the spring, the chances of finding at least one household member at home improved significantly if calls were made in the late afternoon or evening. In the UK, Swires-Hennessy and Drake (1992, p. 72) found in 1986 that “*the highest probability of a successful outcome was between 17.00 and 22.00 hours.*” Purdon and her colleagues in the UK (see, Campanelli, Purdon, and Sturgis, 1997; Purdon, Campanelli, and Sturgis, 1998) confirmed the main findings of Weeks *et al* (1980), and Swires-Hennessy and Drake, (1992), namely that, if the probability of contact at any one call is to be maximised then calls should be made on weekday evenings. Nevertheless Purdon and her colleagues found that it was not obvious, a priori, that instructing interviewers to increase the proportion of calls they make in evenings will lead to gains in efficiency and suggested that

“Restricting calls to evenings reduces the length of the working day and may result in interviewers having to make extra visits to sampling points to complete their assignments. This will increase travel costs and may, in principle at least, increase the number of hours spent working on the survey. Searching for the “most efficient” calling strategy would have to be an organisation specific activity, incorporating information on the organisations’ payment system, interviewer workloads, and travel time and distance to sampling points. Organisations attempting to enforce efficiency in calling strategies would also have to be sensitive to the wishes and concerns of interviewers” (Campanelli, Purdon, and Sturgis, 1997, p. 3-33).

In the case of telephone surveys, optimal call scheduling algorithms have been developed (see, for example, Groves and Robinson, 1982; Weeks, Kulka and Pierson, 1987; and Kulka and Weeks, 1988).

Also Purdon *et al* (1998) suggested that if organisations were to adopt extreme versions of calling strategies (e.g., only during mornings and afternoons), then some consideration would need to be paid to bias as such restriction could effectively exclude certain sub-

groups. Non-contact bias has also been shown to vary with the number of calls with those who are found at home during the first few calls differing in characteristics from those found at later calls (see Dunkelberg and Day, 1973; Lynn, 1995).

2.2 Interviewer Effects

Section 2.1 described the role of the survey research interviewer. According to Sudman and Bradburn (1974), this could be called the 'role demands' of the interviewer (as affected by the skill of the interviewer in actually fulfilling her/her role). This section looks at the literature which describes the mechanisms which can lead to sub-standard 'role behaviour'. It also looks at the 'other aspects' of the interviewer which can have an impact on response quality. As coined by Sudman and Bradburn (1974, p. 15), these 'other aspects' of the interviewer are often called 'extra-role characteristics' and include such things as the interviewer's "*race, educational level or social class, age, and perhaps, religious, ethnic, political, or other affiliation.*"

Different researchers have used slightly different categorisations to describe areas where interviewer effects occur. As noted above, Sudman and Bradburn (1974) made a three-part distinction between 'interviewer role demands', 'interviewer role behaviour' and the 'extra-role characteristics of the interviewer'. Dijkstra and van der Zouwen (1982, p.11) use a slightly different categorisation. They make a distinction between the 'role-restricted' and 'role-independent' characteristics of the interviewer. Under their categorisation, the role-restricted characteristics of the interviewer, include "*style of interviewing, interviewer competence, or expectation about respondents.*" Role-independent characteristics of the interviewer include aspects such as "*gender, race, age, or own opinions.*"

Yet, a different scheme was suggested by Kahn and Cannell (1957). It includes the following three categories:

1. *Interviewer background characteristics* (such as age, education, socio-economic status, race, religion and gender)
2. *Interviewer psychological factors* (such as expectations, attitudes, perceptions, and motives)
3. *Interviewer behavioural factors* (e.g., not correctly implementing the various task rules described in Section 2.1).

The Kahn and Cannell (1957) scheme is particularly useful because it makes a distinction between psychological factors and behavioural factors which Sudman and Bradburn do not. The addition of the psychological category also means that interviewer expectations and attitudes can be grouped together rather than separately as in Dijkstra and van der Zouwen (1982). In addition, the Kahn and Cannell (1957) trichotomy illustrates a natural progression of how interviewer effects can actually occur. For example, the background factors can be seen as the 'soil' in which interviewer's expectations, attitudes, etc. become rooted. In turn, it is only through behaviour that the interviewer's expectations, attitudes, etc. can become an operative component of bias. It is interesting to note that Kahn and Cannell (1957) also apply these same three categories to the respondent as well as the interviewer (see Section 4.2.1).

The structure for the remainder of this section is based on the three part scheme proposed by Kahn and Cannell (1957). Each area of effect is described in turn, with a focus on both response quality and nonresponse.

2.2.1 Background Characteristics

Response Quality. Interviewers' own background characteristics can be seen as a source of some of their expectations, attitudes, motives, and perceptions. Similarly, in the survey interaction itself the background characteristics of the interviewer provide certain cues to the respondent which interact with the respondent's expectations, attitudes, motives, and perceptions. The goal of standardised interviews is to present a consistent stimulus to respondents, yet interviewer background characteristics represent visible differences which cannot be removed through interviewer training. Thus, interviewer background characteristics could be an area of great concern for both interviewer recruitment and assignment practices.

There have been several studies which have reported effects due to interviewers' background characteristics. For example, Hyman *et al* (1954, 1975) reported a study where the race of the interviewer made a difference in black respondents' reporting of their resentment over discrimination, with more information obtained from black rather than white interviewers. Robinson and Rohde (1946) described a similar finding with respect to the Jewish-ness of the interviewer, with fewer anti-Semitic comments made in front of Jewish interviewers. Katz (1942) found more reports of pro-labour opinions were obtained by working class interviewers than middle-class interviewers. Hyman *et al* (1954, 1975) found that female respondents were more likely to agree with a statement about men not respecting women who have had sex before marriage when the interviewer was male rather than female. Freeman and Butler (1976) found that male interviewers rather than female interviewers produced more interviewer variability when interviewing a predominantly female sample about child-rearing practices. Similarly, Erlich

and Reisman (1961) found differences in reporting by adolescents, depending on whether the interviewer was closer to the adolescents' own age or closer to the age of their parents. With the older interviewers, the adolescents gave the more normative types of answers that a parent might want to hear.

As summarised by Fowler and Mangione (1990, p. 105), the general thought behind these findings is that "*when the topic of a survey is very directly related to some interviewer characteristics so that potentially a respondent might think that some of the response alternatives would be directly insulting or offensive or embarrassing to an interviewer*" this is where interviewer effects due to background characteristics are most likely to occur. For example, in a study of racial attitudes shortly after the Detroit race riots, Schuman and Converse (1971) confirmed the findings of Hyman *et al* (1954, 1975) and clearly support Fowler and Mangione's views. Interviewer effects were only noted on the 10 percent of items that dealt with direct feelings and opinions about a racial group. This summary is also consistent with the findings of several general studies that interviewer demographic characteristics have no significant impact on respondents' answers (see, for example, Esbensen and Menard, 1991).

As large differences between respondent and interviewer might produce social distance, some researchers hypothesised that the best reporting would occur when interviewers and respondents were similar in status. Research evidence, however, suggests that this may not always be a wise idea. Weiss (1968) found that among a sample of welfare mothers, better reporting occurred when the interviewer was of a higher socio-economic class or better educated. Similarly, Anderson, Silver, and Abramson (1988a, 1988b) found that black respondents were much more likely to over-report their voting behaviour when they

were interviewed by a black interviewer. These findings could tie back to the idea of rapport discussed in Section 2.1, with very high levels of rapport leading to bias.

Other studies have found effects of interviewers' background characteristics which at first glance do not fit into the conceptual explanations described above. For example, Collins and Butcher (1982, p. 51-52) investigated the explanatory power of several characteristics of interviewers. As they described, "*Of these 98 analyses, 14 yielded an explanatory variable effect significant at the 1 percent level, and a further 10 an effect significant at the 5 percent level. Many of these were, however, isolated results not part of any apparent pattern.*" The strongest evidence for a systematic pattern of effects was with respect to interviewers' age. Sample size limitations, however, restricted further analysis of these age effects. Freeman and Butler (1976) also found age effects. Interestingly, they found that pairing older interviewers with older respondents actually produced more interviewer variance. The combination which minimised interviewer variance was young interviewers with young respondents.

Nonresponse. A few studies have looked at the background characteristics of interviewers and unit nonresponse. It appears that only under certain circumstances, can these make a difference to the interviewers' success in obtaining the interview. For example, Oksenberg, Coleman, and Cannell (1986) found that various aspects of the voice quality of telephone interviewers did make a difference. Fowler and Mangione (1990) found that female interviewers were more likely than males to be perceived as 'friendly'. Morton-Williams (1993) found that several of her respondents said they would feel more wary of a male stranger on the doorstep, but Lessler and Kalsbeek (1992) in their review of the literature found little systematic evidence to support the view that male interviewers have lower response rates than female interviewers. Campanelli, Sturgis, and Purdon

(1997) found no effect of interviewer gender on nonresponse and no effect of an interaction between gender of the interviewer and gender of the respondent.

Morton-Williams (1993) found that age of the interviewer was not important, but Lievesley (1986) reported that middle-aged interviewers had higher response rates than the younger or older ones. In contrast, Singer, Frankel and Glassman (1983) suggest that the highest response rates are found with the older interviewers.

Morton-Williams (1993) found that interviewers had the most success when they dressed to avoid extremes of smartness, casualness or exoticness.

2.2.2 Psychological Factors

Response Quality. There is always the question of experimenter effects caused by the experimenter's expectations (see, for example, Rosenthal, 1966; and a review of the topic by Barber and Silver, 1972). Interviewer effects can clearly be seen as an analogue to experimenter effects. Early work reported by Hyman *et al* (1954, 1975) makes it clear that interviewers frequently do hold expectations about their respondents, which can take one of three forms:

- *Attitude-structure expectations*, where the interviewer might expect certain answers based on the answers that the respondent has already given,
- *Role expectations*, based on the interviewer's initial impression of the respondent, and
- *Probability expectations*, where interviewers have some expectations about the frequency of occurrence of some characteristics among respondents.

Of course, whether or not these expectations actually cause response effects in surveys is another issue. As Kahn and Cannell (1957, p. 185) point out: attitudes can be considered as a “*predisposition to behave in a certain way, or to react characteristically to a given stimulus.*” Attitudes can, in turn, lead to certain expectations about the respondent.

According to Kahn and Cannell (1957) the key is whether the stimulus appears to evoke the interviewers’ attitudes and expectations. Such a stimulus could be the content of the interview, the respondent himself/herself, or something the respondent says.

Standardised survey interviewing as described in Section 2.1 is designed to minimise such possibilities interfering with survey results. For example,

“the specific interviewing procedures prescribed for the interviewer tend to check the arbitrary exercise of his expectations. . . For example, the rule to record the respondent’s words verbatim and to code a reply in the answer box that most nearly corresponds to the actual words reduces the biases arising even when the interviewers hold contrary expectations“ (Hyman et al, 1954, p. 84).

Yet, Hyman *et al* (1954, p. 85) point out that such rules do not preclude effects due to expectations and attitudes. For example,

“under conditions of stress, or difficulty in the interview situation, the rules may be consciously flouted. . . Moreover, . . . the interview situation is not that rigid. There are various choices left to the interviewer. He can continue to probe, or he can accept the answer already given. He can ask the next question, or he may assume that he already knows the answer and that the question is therefore redundant. In addition, the interviewer must apply his judgement in coding an equivocal answer into one of a limited number of prepared answer boxes, and even the most rigid rule to record answers “verbatim” allows the interviewer to omit irrelevancies without defining what an irrelevancy is. At all these points of choice, the interviewer may well let his expectations be his guide.”

Hyman and his colleagues (1954, 1975) report on several studies which showed the presence of effects due to interviewers’ expectations. After the work of Hyman *et al*, however, little is seen in the literature regarding the effect of interviewer expectations until Sudman, Bradburn, Blair and Stocking (1977). Sudman and co-workers found

modest effects suggesting that interviewers who expected the study to be difficult actually obtained lower percentages of the respondents acknowledging sensitive behaviours.

Similarly, those interviewers who expected their respondent to underreport sensitive behaviours obtained lower levels of reporting. Singer and Kohnke-Aquirre (1979) who revisited and extended this work found similar results: that on the sensitive questions that interviewers rated as difficult they actually obtained lower percentages of reported behaviour.

So far, this Section has focused on interviewers' expectations. With respect to interviewers' attitudes, Barr (1957) and Freeman and Butler (1976) both found evidence of correlation between interviewers' attitudes and respondent opinions. But it should be noted that interviewers' attitudes were gathered after the survey and could have been affected by the respondents' attitudes. Collins (1980) who measured his interviewers prior to the survey, found no association between interviewers' and respondents' opinions. Kahn and Cannell (1957) report on several early studies where clear evidence was found that differences between respondents are in the direction of their interviewers' own opinions (see, for example, Cahalan, Tampulonis, and Verner, 1947; Ferber and Wales, 1952; Blankenship, 1940). Hyman and his colleagues (1954, 1975) also found that there was little support for effects due to interviewers' ideological processes.

In summary, it appears that the psychological characteristics of interviewers are indeed potentially biasing, but past research suggests that the size of the bias may be modest (see Sudman *et al*, 1977; Singer and Kohnke-Aquirre, 1979; and Esbensen and Menard, 1991). As Singer and Kohnke-Aquirre point out, however, this does not mean that the issue should be ignored. For example, as Hyman *et al* (1954, 1975) suggest, there may be

hidden effects due to the highly institutionalised patterns and beliefs shared by all interviewers. The greater such homogeneity in belief (what Hyman and his colleagues call the 'universally held expectation'), the more difficult it is to measure because there is nothing to contrast it against.

Nonresponse. Singer and Kohnke-Aguirre (1979) also reported that interviewer beliefs about item sensitivity could significantly predict the likelihood of their obtaining or failing to obtain responses on those items; a point well known by survey field managers, particularly with respect to the income question.

Singer, Frankel and Glassman (1983) confirmed that interviewers' expectations about the ease of persuading respondents to be interviewed was significantly related to their response rates. Campanelli, Sturgis, and Purdon (1997) found that interviewers were able, at a greater than chance level, to predict the likelihood of achieving an interview by simply observing the outside of the selected dwelling before making any contact with the occupants. They point out, however, that it is unclear to what extent this finding simply reflects a self-fulfilling prophecy among interviewers.

Recently, however, researchers have looked at interviewers' attitudes towards persuasion strategies as well as their attitudes about the role of the interviewer. For example, interviewers who believed in strong persuasion strategies and did not believe that the voluntary nature of the survey should be emphasised, had better response rates (see Lehtonen, 1996; de Leeuw *et al.*, 1997). These studies, however, are based on aggregate comparisons. Using individual interviewer level data and the same 5 item scale,

Campanelli, Sturgis, and Purdon (1997) found no relationship between interviewers' attitudes and their response rates.

2.2.3 Behavioural Factors

Response Quality. An interviewer's actual behaviour during the interview is of course a key aspect which differentiates between good quality data or data riddled with response effects. Interviewer behaviour deficiencies can be due to several causes. For example,

- the interviewer may not have been properly trained in his/her role requirements as described in Section 2.1,
- the interviewer may have been properly trained, but
 - may be a novice interviewer who has not yet learned to fully integrate the interviewers' role,
 - may be an experienced interviewer who has not been properly supervised and whose skills have therefore deteriorated (e.g., Fowler and Mangione, 1990),
 - may be an interviewer who has little motivation to perform well, or
 - may be an interviewer whose expectations and attitudes have a tendency to over-ride his/her training and affect how he/she records answers or how he/she interacts with the respondent, thereby alerting the respondent to his/her expectations and attitudes.

All of these deficiencies can ideally be handled through proper interviewer recruitment and training with continued monitoring and supervision.

Nonresponse. As we saw from Section 2.1, interviewers' behaviour has a direct impact on the proportions of non-contacts in a survey, through the guidelines from their office,

but more importantly through their own persistence in trying to make contact with each case. Interviewers' behaviour on the 'doorstep' has a direct impact on the refusal component of nonresponse and depends on both their personality and their skill. Several authors have found relationships between interviewer experience and success in obtaining an interview. As Groves, Cialdini, and Couper, (1992, p. 478-9) suggest experienced interviewers' success could be seen to derive from their "*larger number of combinations of behaviours proven to be effective for one or more types of householders.*"

Durbin and Stuart (1951) found the refusal rate for inexperienced amateur interviewers to be about three times that of experienced professional interviewers. Other studies have also found some evidence to suggest interviewer experience positively influences response rates (see, for example, Colombo, 1983; Groves and Fultz, 1985; Lievesley, 1986; Couper and Groves, 1992; de Leeuw and Hox, 1996). It should be noted, however, that such findings are typically confounded with interviewers' self-selection to remain as interviewers. Singer, Frankel, and Glassman (1983, p. 80) found a curvilinear relationship between response rates and years of experience. "*Some experience is better than none ..., but with longer experience . . . the response rate actually declines.*" They state the caveat, however, that their findings are based on a small number of interviewers with a disproportionate number of young and inexperienced ones.

Although early research has experimented with varying the content of what the interviewer says 'on the doorstep' (see, for example, Dillman *et al*, 1976; O'Neil *et al*, 1980), it is only recently that interviewers' doorstep behaviour has been subjected to detailed research, through the use of taped-recorded doorstep introductions and the collection of doorstep behaviour on special contact description forms (see, for example,

Morton-Williams, 1993; Maynard, Schaeffer, and Cradock, 1993; Groves and Couper, 1994, 1996; and Campanelli, Sturgis, and Purdon, 1997). This research has resulted in various training recommendations (see, for example, Morton-Williams, 1993; and Campanelli, Sturgis, and Purdon, 1997).

A key concept in this research is the idea of *tailoring*. Groves, Cialdini and Couper (1992) employ the term *tailoring* to describe the way in which interviewers deliberately manipulate their dress, behaviour and language to suit the perceived characteristics of each sample unit in order to maximise the probability of obtaining an interview there. Data from doorstep interaction sequences clearly supports the role of tailoring in improving response rates (see Groves and Couper, 1994, 1996; Campanelli, Sturgis, and Purdon, 1997).

In an attempt to measure interviewers' tailoring skills from interviewers own self-assessment, Couper and Groves (1993) adapted a scale by Snyder (1979) on *self-monitoring*. They believed the construct of *self-monitoring* to be closely related to their notion of tailoring. This scale included questions about interviewers' awareness of how they acted with respect to different people in different situations. Later Campanelli, Sturgis, and Purdon (1997) adapted this scale from Couper and Groves (1993) and adapted a scale by Spiro and Weitz (1990) which was based on the theory of *adaptive selling*. The adaptive selling scale included questions about the extent to which interviewers' actually adapted their doorstep approach to different people and different situations. Campanelli, Sturgis, and Purdon (1997) then included both scales as part of a questionnaire that was used for studies at two different organisations. They found that the two scales were moderately correlated to each other ($r = .41$), but generally not

correlated with current or past response rates! The one significant finding with respect to response rates was actually for the adaptive selling scale. But, this was only for one organisation and was in the opposite direction of what would have been expected.

2.3 Types of Questions Prone to Interviewer Effects

Are some types of questions more prone to interviewer effects than others? As Kish (1962) suggests, “*interviewer effects may vary markedly with differences in question form and content*” (p. 96). This section provides a brief review about what is known about this topic.

2.3.1 Effects Based on Question Content

Background Characteristics of the Interviewers. As noted in Section 2.2.1, the influence of an interviewer’s background characteristics are typically felt when the question is sensitive and highly related to that particular background characteristic (see, for example, Hyman *et al*, 1954, 1975; Robinson and Rohde, 1946; Katz, 1942; Erlich and Reisman; 1961; Schuman and Converse, 1971).

Sensitive questions. Bradburn and Sudman (1979) suggest that task variables which deal with social desirability, i.e., sensitive questions in general, should be more prone to interviewer effects. Hanson and Marks (1958), in their study of census variables, found significant interviewer variance of questions for which the interviewer was resistant or hesitant to ask. This is similar to Hansen, Hurwitz and Bershad’s (1961) finding that difficult items such as income were more susceptible to interviewer effects. Fellegi (1964) found that outstandingly high values of intraclass correlation coefficient (see Section 1.2) were found on emotionally charged items. In contrast, Fowler and Mangione (1990)

found no relationship between question sensitivity and the size of intraclass correlation coefficients.

Attitudes Versus Facts. Hyman *et al* (1954, p. 204) found that “*on most of the nonattitudinal questions*” . . . [*interviewer differences were*] *not significant.*” In contrast, “*on the attitudinal questions most differences [were] highly significant.*” Kish (1962) and Groves and Magilavy (1986) explored the attitude/fact hypothesis further and failed to find any systematic pattern of this kind. Similarly, Collins (1980, p. 81) found that that “*high interviewer effects (of 0.05 or more) were as common among factual questions as among attitudinal questions.*” Collins and Butcher (1982, p. 43) note that “*it has been our consistent experience that very high interviewer variance effects can occur as often with behavioural and descriptive questionnaire items as with attitudinal items.*” However, they go on to say “*nevertheless, we do find evidence that attitudinal items tend to be subject to rather more interviewer variability overall.*” Further support for the susceptibility of attitude questions was found by Gray (1956).

2.3.2 Effects Based on Question Format

Questions with Poor Definitions. Hanson and Marks (1958, p.635) found that one of the key factors determining whether or not interviewer effects would occur was whether the survey items had a “*relatively high ambiguity, “subjectivity,” or complexity in the concept or wording.*” In line with this, Gray (1956) found interviewer effects on a question where the term ‘carpeted’ had a wide range for interpretation. He found a similar phenomenon with interviewers’ classification of respondents as ‘living at home’ or ‘living in rooms’. O’Muircheartaigh (1976) found interviewer variance due to one interviewer who had a different interpretation of the classification of the respondent’s job

as skilled, semi-skilled, or unskilled. Collins (1980) describes a comparable situation with the interpretation of 'go for walks in the country' as part of a series of leisure activities.

In addition, to ambiguity of question concepts, Feather (1973) found high interviewer effects on questions where there was ambiguity about the acceptable interviewing procedure. Similarly, Collins (1980) reported an example where interviewers' interpretation of the phrase, 'Can I just check . . .?', at the beginning of one question led to interviewer variability on actually omitting the question. He hypothesised this was because the 'lead in' was similar to standard interviewer 'check items' which are not to be asked of the respondent. (Interviewer check items are typically used in paper and pencil questionnaires to bring forward previously collected information which is then used to determine the proper filtering for subsequent dependent questions.)

Unspecified Categories. Interviewers can also experience ambiguity with respect to the use of certain categories on closed questions. For example, Collins (1980) cites two example: one where interviewers differed in their use of the 'not applicable' category and the other where interviewers differed in the use of a category which was to cover respondents who volunteered the information that they were 'not looking for work'.

Open Questions. As reported by Kish (1962, p. 96), Cannell (1954) "*hypothesised that greater differences among interviewers would occur when the nature of the subject made extensive probing necessary.*" This suggests that open questions could be more vulnerable to interviewer effects than closed questions. This hypothesis has been supported by the work of Gray (1956), Kemsley (1965), Shapiro (1970), and Collins (1980). Although Groves and Magilavy (1986) actually found higher interviewer effects

on the closed questions rather than the open questions, they did find high interviewer variability on the number of items mentioned on open questions.

Repetitive Questions. McKenzie (1977) found the presence of high interviewer variance effects on long repetitive series of questions, suggesting variation in interviewers' use of prompts over the series. With respect to repetitive questions, Collins (1980) found differences in the proportion of strongly-held responses depending on the interviewer. Collins (1980) also found interviewer effects on a filter question where the answer of 'no' absolved the interviewer from supplementary questions.

2.3.3 Interviewer Effects on Self-Completion Documents

Self-completion documents come in many forms. Some, such as postal surveys, completely exclude the use of interviewers. Others may use interviewers for the distribution and collection of the survey instrument. In addition, when answers to sensitive questions are needed, interviewer surveys have embedded self-completion components. Given the presence of the interviewer in these latter cases, there is still the possibility for interviewer effects, although one would hypothesise that these should be smaller than on direct interviewer administered questions. There has been very little research in this area. Kish (1962), for example, found little evidence to suggest the presence of interviewer effects on the written questionnaires he examined.

O'Muirheartaigh and Wiggins (1981), however, did find an effect for a health supplement completed in the presence of an interviewer.

2.4 Chapter Summary

The first section of this chapter has described in detail the role of the interviewer through outlining five different areas of interviewers' responsibilities: (1) being a *neutral collector of accurate data*, (2) *educating the respondent as to his/her role*, (3) *conducting the last stages of sampling*, (4) *fulfilling administrative duties*, and (5) *gaining co-operation with the respondent*. This provides useful background information for this thesis. For example, the impact of deviations in interviewers' behaviour from these standards can clearly be seen to impact on total survey error with Task 1 and 2 being a potential source of both response variance and response bias; Task 3 being a potential source of both sampling variance and sampling bias, and Task 5 being a potential source of nonresponse variance and bias. Errors in Task 4 have less to do with variance and bias and more to do with generally inefficiency and thus cost.

Different classifications of the influence of interviewers were explored in Section 2 of this Chapter. Kahn and Cannell's (1957) breakdown of interviewer effects into those due to background characteristics, those due to psychological factors, and those due to behavioural factors provided a useful classification scheme for reviewing the literature with respect to interviewer effects. Within each of the three areas, literature with respect to response quality and nonresponse error were considered separately. With respect to response quality, the literature suggests that the background characteristics of interviewers are only influential when the characteristics of the interviewers mirror highly sensitive topics in the questionnaire (e.g., race of interviewer and questions on racial discrimination). Psychological factors are less well documented, although effects have been shown to exist. Good interviewer training is often assumed to minimise any

problems in this area. How interviewers behave during the interview, however, is key to the resultant level of response quality.

With respect to nonresponse, there is very little literature on the effects of interviewer background characteristics, although a general conclusion could be that as long as the interviewer does not present extremes of appearance their background characteristics have little impact on response rates. The literature has shown that if interviewers think a survey question or the questionnaire as a whole will be difficult, this often proves to be the case with higher item or unit nonresponse rates. Again, however, the behaviour of the interviewer is key. With respect to nonresponse, this is the behaviour of the interviewer during the ‘doorstep’ interaction where the respondent is either convinced or not convinced to participate.

This section has provided a thorough overview with respect to the effects of interviewer background characteristics and years of experience. These characteristics are specifically examined with respect to BHPS nonresponse in Chapter 5 and with respect to response quality in Chapter 9. Although psychological or behavioural data on interviewers was not available for this thesis, the random effects models considered in Chapters 6 and 8 with respect to nonresponse and Chapter 9 with respect to response quality capture these residual influences of the interviewer.

The last section of this chapter looks at interviewer effects from the perspective of survey questions. The susceptibility of various types of questions to interviewer effects are considered. With respect to question content, mixed results were found for the susceptibility of sensitive questions and attitude questions. Stronger support for

interviewer effects was found with respect to question format. Susceptible questions included those with poor definitions, those with unspecified categories, those with an open format, and repetitious questions. The literature also suggests that even self-completion documents completed in the presence of an interviewer may be prone to a mild interviewer effect. Many of these themes will be tested by the research presented in Chapter 9.

CHAPTER 3 EXPLORING NONRESPONSE ERROR

A serious hazard to drawing inferences from sample surveys is that a substantial proportion of the sample of the population originally selected may fail to respond. Quite apart from the waste of fieldwork resources which this entails, the non-responding units are likely to differ from responding units in terms of characteristics which are directly or indirectly related to the variables which the survey is intended to measure. Where this is so, nonresponse can seriously bias estimates of means and proportions as well as the strength of relationships between variables. As described in this Chapter, the literature contains many instances of demonstrated or inferred nonresponse bias of this kind. Before proceeding, several frequently used concepts in this area must be defined. These include the term 'nonresponse' itself, 'response rates', 'nonresponse bias', and 'nonresponse variance'.

3.1 Definitions of Nonresponse

There are two types of nonresponse. Unit nonresponse refers to the failure of a sample unit to participate in the survey. Depending on the survey, units can be individuals, households, establishments, schools, etc. Item nonresponse refers to failure to obtain an answer to a particular questionnaire item from an otherwise participating sample unit. It can also include cases where the initial response recorded during the data collection is found later to be unusable, such as an impossible age value (Lessler and Kalsbeek, 1992).

The main focus in this thesis is on individual and household unit nonresponse in probability surveys. The main categories of unit nonresponse in a one-off probability survey are refusals (where sampled units refuse to participate) and non-contacts (where

sampled units can not be contacted during the fieldwork period). There is typically also a small category labelled as 'unable to participate'. This includes individuals such as those who are ill or in hospital during the whole fieldwork period, those who are senile, and those who are unable to speak English when no interpreter is available. In addition to these three, an 'other' category sometimes exists although it is rarely mentioned. It includes the unusual cases where data are lost (such as completed paper and pencil questionnaires being lost in the post or completed computer-assisted questionnaires being lost during electronic transmission to headquarters).

In longitudinal studies, sample loss – attrition – includes all of the categories mentioned above. It also typically includes sample units who can not be traced, those who moved out of scope, and former respondents who have died.

3.1.1 Response Rate Calculation

From the perspective of the sponsors of survey research, a survey's *response rate* (which is often seen as the complement of the amount of nonresponse) is often seen as a survey's key indicator of survey quality. This is unfortunate for several reasons. First is the lack of attention to the many other sources of survey error described in Section 1.1. Second is the fact that a high response rate is not necessarily synonymous with a low amount of nonresponse bias (see Section 3.2). Third, is the fact that response rates typically vary by many other factors independent of survey quality (e.g., the type of population surveyed, and the topic of the survey, etc.). For example, much higher response rates can be expected of a survey of mothers when the topic is their children than a survey of households when the topic is their detailed financial situation.

Given the importance of a survey response rate, it is interesting to note the great variation in its calculation (see, for example, a discussion by Groves, 1989; Lessler and Kalsbeek, 1992). For example, Wiseman and McDonald (1980) asked 40 survey research organisations in the United States to compute a response rate as they would report it. Using the same set of outcome data from a typical telephone survey, 29 different rates ranging from 12 to 90 percent were produced by these organisations. A similar research exercise was conducted by Spaeth (1992) of survey research organisations in the United States and Canada. Her findings were comparable to Wiseman and McDonald (1980). As she suggests, *“for the 38 organisations that reported how they calculate response rates, there are none in which the wording is completely identical. Some of the calculations are close, but the wording varies just enough to make it difficult to tell if they are really the same”* (p. 19). For example, the denominator which is simply stated as ‘eligibles’ hides all the differing opinion of what cases should be included as ‘eligibles’.

Because of this variation, there has been a push within the International Workshop for Household Survey Nonresponse (see Hole, 1993; Foster, 1994), to standardise response rate calculation. Groves (1989, p.141) reports what he believes to be the most universally accepted rate. The core definitions which follow are based on the work of Groves and are extended to include the categorisation of nonresponse described in Section 3.1 (refusals, non-contacts, those unable to be interviewed, and other). The definition has also been expanded to account for the fact that in some instances, the eligibility of a unit must be estimated.

$$\text{Response Rate} = \frac{\text{Total Completed Interviews}}{\text{Total Completed Interviews} + \text{Nonresponse} + \text{Estimated Eligibles}} \quad (35)$$

or

$$= \frac{\text{Total Completed Interviews}}{\text{Total Sample - Ineligibles - Estimated Ineligibles}} \quad (36)$$

where nonresponse = refusals,
non-contacts,
those unable to be interviewed, and
other

Ineligibles (for a face-to-face household survey using an address-based sampling frame like the BHPS) =

sampled addresses that don't exist, are vacant,
are temporary or vacation homes, are businesses,
etc.

Although this formula clarifies the numerator and the meaning of nonresponse, there is still room for interpretation with respect to the eligible categories and even more so when a unit remains *unresolved*. The complexities vary depending on the type of survey situation. For example:

- Is a partial interview to be treated as response or nonresponse?
- In face-to-face household surveys: is a non-contact a non-contact (i.e., nonresponse) or a vacant unit (i.e., ineligible)?
- In random digit dialling telephone surveys: is a non-contact a non-contact (i.e., nonresponse) or a non-working number (i.e., ineligible)?
- In postal surveys: is a non-return a refusal (i.e., nonresponse) or an indicator that the letter has been sent to an ineligible address (e.g., a business or a vacant unit)? (See

Lynn and Purdon, 1994, and note that not all ineligibles are returned by the Post Office.)

- In surveys with a screening questionnaire: Are households which failed to be screened to be treated as eligible or ineligible, or more realistically, what proportion should be treated as eligible and what proportion should be treated as ineligible?
- In surveys where all members of the household are to be interviewed: At what level should the response rate be reported? (Distinctions are often made between all individuals in a household as completely co-operating, some co-operating and some via proxy, and some co-operating and some nonrespondents - see, for example the BHPS response rates in Section 5.3.1)
- In surveys with differential probabilities of selection: Should a weighted or unweighted response rate be used? (A weighted response rate gives an approximation of what the response rate would be like if the sample were selection with equal probability. For a discussion of this issue see Groves, 1989 and Lessler and Kalsbeek, 1992.)

Other rates are also in use in survey organisations and share the proliferation of definitions and enhance the general confusion in this area. For example, in addition to the response rate, there is

- the refusal rate (Lessler and Kalsbeek, 1992, identify two),
- the refusal conversion rate (see Groves, 1989),
- the completion rate (CASRO, 1982, identifies 8),
- the co-operation rate (see Sudman, 1976), and
- the contact rate (see Groves, 1989)

to name just a few. The bottom line for interpreting these rates is knowing exactly how they were calculated.

3.2 Nonresponse Bias

As described in Section 3.1.1, survey response rates are often interpreted as the key indicator of survey quality. However, bias in survey estimates is not simply due to response rates, but also to differences between respondents and nonrespondents on relevant survey variables. This is summarised by Groves (1989, p. 133-134)⁷ as follows

$$y_r = y_t + \left(\frac{n_{nr}}{n_t} \right) (y_r - y_{nr}) \quad (37)$$

where n_t = the total sample eligible for the survey,

n_r = the group of persons who responded to the survey,

n_{nr} = the group who did not respond,

y_r = a statistic estimated from the n_r respondent cases,

y_t = a statistic estimated from the n_t total sample cases, and

y_{nr} = a statistic estimated from the n_{nr} nonrespondent cases,

Here we can see that the value we obtain from the survey (y_r) is a function of the value we should have obtained given 100 percent response (y_t) and a nonresponse error term. This term shows that the extent of non-response bias is a function of the size of the proportion of non-respondents (n_{nr}/n_t) and the presence and magnitude of the difference in the estimates ($y_r - y_{nr}$) between respondents and non-respondents.

⁷ Groves' original notation has been modified slightly to improve clarity. For example, n_{nr} was originally specified as nr which could be confused with $n \times r$.

It is the nature of nonresponse, however, that the latter term, $(y_r - y_{nr})$, is rarely available, because nonrespondents are just that, non-responding units. Special study designs and access to auxiliary information are needed in order to estimate the latter term. (These are described in Section 5.1). Thus, we often come full circle to depending on response rates as a measure of survey quality. Fortunately, there is evidence to suggest that there is a rough correspondence between the two aspects with nonresponse bias tending to be greater on surveys with higher nonresponse rates (see Foster, 1996).

As described in Section 3.4.2, there is considerable evidence to suggest that the types of people who tend to refuse to participate in a survey differ from those who can not be contacted (see, for example, Bebbington, 1970; Campanelli, Sturgis, and Purdon, 1997; Foster, 1996; Goudy, 1976; Gray *et al*, 1996; Groves and Couper, 1992; 1995; Kalton *et al*, 1990; Lievesley, 1986; Pavalko and Lutterman, 1973; Waterton and Lievesley, 1987; Weaver, Holmes and Glenn, 1975; and Wilcox, 1977). For this reason, Groves (1989, p. 134) extends his model given in equation (37) to include separate components for refusal cases, non-contact cases, and other non-interview cases. This issue is investigated in Chapter 5.

There is also some evidence to suggest that the bias from these two sources of nonresponse are compensating rather than additive (see, in particular Wilcox, 1977 and Lievesley, 1986). This issue will also be considered in Chapter 5. Groves (1989) extrapolates from this evidence and runs a simulation in which the typical survey fieldwork strategy of focusing of the reduction of the non-contact rate rather than the refusal rate (because this is much easier to manipulate), may actually increase the total nonresponse bias. This is because the bias of the non-contact cases are no longer present to

compensate for the bias of the refusal cases. Thus Groves (1989, p. 208) points out that often efforts have by and large had “*the sole goal of increasing the proportion of the sample measured, without formal concern about the nature of the people added to the sample or about the response errors they commit.*” The latter point brings up the equally important issue of response quality. If respondents are coerced to participate simply to raise the response rate figures, response quality can potentially suffer (see, for example, Triplett *et al*, 1996).

3.3 Nonresponse Variance

Theoretically, nonresponse can be viewed as a bias or a variance, but is typically viewed as a bias. This bias is often measured through comparison of the achieved sample for a particular survey to a more complete source of data, e.g., the Census. These types of comparison exercises often show that various variables on the survey are biased and the underrepresented types of people are described. Consideration often stops there. But does this say anything about how we should view nonresponse, and thus learn to minimise it?

In order to investigate this issue more fully, one needs to introduce the concept of replication. What would happen if we were to repeat the survey under the same essential survey conditions over all possible samples. It is likely that many of these samples would contain bias and that the magnitudes and types of bias would vary over the samples.

Under such a design, any variation would be due to the respondents chosen for a particular sample. But what is the true nature of this variation. A deterministic approach would assume that for all members of the population, the probability of responding p_i is either equal to 1 or 0 (i.e., some people participate in surveys, some people never do). A review of the literature with respect to the characteristics of nonrespondents *could* be

interpreted as providing support for this hypothesis as there is a pattern to the results across surveys suggesting that some types of people are always harder to include in a survey (see Section 3.4.2 for a review of past research and findings.)

But a person's propensity to respond might not be constant over time. To test this second proposition one would need to replicate the survey over the same sample.

Evidence from most panel surveys is that in moving from one wave to the next, respondents tend to turn into nonrespondents. Those panel surveys which re-visit their nonrespondents, however, often find that nonrespondents also turn into respondents (see, for example, Kalton *et al*, 1987). The former is more pervasive than the latter, probably due to the burden of the multiple interview experience. If respondents' memories of previous visits could be erased, we would hypothesise a much more equal distribution. This is supported by the finding of organisations who conduct re-issues and find that initial refusals are often situationally specific, i.e., highly influenced by what was happening in the respondent's life at the time of the initial calls (see, also Groves and Cialdini, 1991). Under this second model, nonresponse with respect to the individual is seen as a variable process. This stochastic view point would thus model the a respondent's propensity to respond (p_i) as $0 < p_i < 1$. As Lessler and Kalsbeek (1992) suggest, real populations display both types of processes. Yet, despite this, models of survey error rarely include nonresponse variance due to respondents.

Another important source of nonresponse variance is the interviewer. Despite the same training and supervision, and controlling for the effect of the workload area, interviewers do achieve different response rates. Groves (1989, p.159) describes this source of error as "*One of those that is well known by field managers but rarely selected in statistical*

models of nonresponse error.” Another way to put it is that the amount of nonresponse bias in a one-off survey is a function of both the chosen respondents and the chosen interviewers. As Groves (1989, p. 159) goes on to suggest, “*if a survey analyst is interested in estimating the variability in error in survey statistics over replications of the survey, then the interviewer as a source of variability in nonresponse error should be considered.*”

In addition to considering the interviewer as a source of variability in nonresponse error estimates as described above, it is useful to investigate the impact of specific interviewers on actual response rates. Understanding the variability in interviewers’ persuasion skills and persistence in contacting the hard-to-reach respondents at home could allow survey organisations to improve interviewer training and/or interviewer recruitment criteria, thus training/selecting interviewers to be at the top end of the response rate scale rather than the bottom end. This issue is studied in Chapters 6 and 8.

3.4 Theoretical Perspectives

In this section, the sociological and psychological perspectives underlying survey nonresponse are described and a theoretical model for use in this thesis is developed.

3.4.1 Social Context Variation

At a macro level, over the last 40 years there have been various changes in what might be called the ‘social fabric’ of society. Take, for example,

- Aggregate level increases in the number of requests to provide information to outside sources (surveys being just one source) resulting in a general reduction in public

willingness to co-operate and an increase in the value of privacy (Bergman, Hanve, and Rapp, 1978; Cialdini, Braver, and Wolf, 1993; Schleiffer, 1986).

- Changes in population distribution and increasing urbanicity impacting negatively on response rates through changes in social value systems and fear of crime, etc. (House and Wolf, 1978).
- Restructuring of the employment market leading to an increased likelihood of finding no one at home and of people having less time and therefore being less willing to participate if contacted (Weber and Burt, 1972; Weeks *et al*, 1980).

3.4.2 Respondent Level Variation

Although, by definition, good data on those who refuse to participate is hard to obtain, there is substantial evidence to suggest that it is more than just chance factors that differentiate those who do and those who do not agree to participate.

Evidence from panel attrition and survey follow-up studies suggests that non-responders differ from responders in terms of various socio-demographic and economic variables (see Bebbington, 1970; Bergman, Hanve, and Rapp, 1978; DeMaio, 1980; Foster, 1996; Gray *et al*, 1996; Groves and Couper, 1992; 1995; House and Wolf, 1978; Kalton *et al*, 1990; Lindstrom, 1983; O'Neil, 1979; Pavalko and Lutterman, 1973; Silberstein, 1994; Vehovar and Zaletel, 1995; Waterton and Lievesley, 1987; Weaver, Holmes and Glenn, 1975; and Wilcox, 1977 and also Groves, 1989 for a good review). For example, nonrespondents as opposed to respondents are much more likely to be categorised as single family households, individuals with a low educational background or with a low intelligence quotient, the young and the elderly, men, non-white individuals, the non-married, the self-employed and in some instances the unemployed, renters, those less well off economically

but also sometimes the very well off, the highly mobile, those who do not feel part of the community, those not committed to a particular political party, those with strict conservative views, those living in urban areas or areas with a high population density or fear of crime, households of unrelated sharers, and those more likely to answer 'Don't know' in previous surveys. In addition, Cialdini, Braver and Wolf (1991, 1993), Mathiowetz (1992), and Hox, de Leeuw, and Voorst (1995) suggest that attitudes toward surveys, perceived survey participation of friends, and the value placed on privacy are consistently important factors in survey participation.

As suggested in Section 3.2, the pattern of nonresponse bias may be complex because the demographic characteristics of those people who refuse to participate in a survey differ from those who can not be found (see, for example, Bebbington, 1970; Campanelli, Sturgis, and Purdon, 1997; Foster, 1996; Goudy, 1976; Gray *et al*, 1996; Groves and Couper, 1992; 1995; Kalton *et al*, 1990; Lievesley, 1986; Pavalko and Lutterman, 1973; Waterton and Lievesley, 1987; Weaver, Holmes, and Glenn, 1975; and Wilcox, 1977).

These 'social group membership' factors outlined above must be mediated in some way to be manifested as individual respondent decisions. Respondent's perceptions, attitudes, expectations, and motives need to be translated into behaviour. One approach to this process is the Theory of Reasoned Action (Ajzen and Fishbein, 1974; Fishbein and Ajzen, 1975; Van der Putte, 1991) which hypothesises that 'the behavioural intention' is the only direct predictor of behaviour. Using a quasi-experimental design, Cialdini *et al* (1991, 1993) and later Hox, de Leeuw, and Vorst (1995) showed that attitudes and behavioural intentions toward survey participation significantly predicted subsequent response. But

more work is needed in this area as both studies still had a substantial amount of unexplained error variance and may have missed the 'hard-core' of non-responders.

3.4.3 Interviewer Level Variation

The general effect of the interviewer was reviewed with respect to background characteristics, psychological factors (such as their attitudes and expectations), and behavioural factors (i.e., their persistence in pursuing non-contacts and their skill at their doorstep introduction) in Section 2.2. The results with respect to nonresponse were rather sparse in comparison with the data on their effect on response quality. It could be argued that good training is the key to minimising effects from all three of these sources. For example, good training can prevent interviewers' negative expectations from becoming self-fulfilling prophecies and good training can clearly enhance their skill on the 'doorstep'. Interestingly, with these two obstacles overcome, any effects due to interviewers' background characteristics can also be minimised.

The most promising research on the interviewers' role in nonresponse has been on the dyadic interaction between interviewer and respondent. For example, Morton-Williams (1993) draws on a tradition established by Kahn and Cannell (1957) in applying Kurt Lewin's Field Theory of Motivation in the context of the survey doorstep interaction. This conceptualises the doorstep interaction as a goal-directed activity in which both participants are aiming to achieve certain specific (and often conflicting) goals. These goals and factors affecting their achievement form the 'field' of motivations over which the interaction takes place. The achievement of goals by participants in the interaction is dependent upon the outcome and relative importance of a range of social psychological factors in each specific interaction. The task of the interviewer then becomes that of

recognising and mastering these factors and manipulating them in such a way as to maximise the chances that the respondent will agree to be interviewed.

Morton-Williams (1993) and Groves, Cialdini and Couper (1992) provide comprehensive reviews of the array of different psychological and social-psychological principles and behaviour relevant to the outcome of the request for survey participation. Both draw on interviewer discussion sessions to illustrate how these principles are manifested in real interviewer persuasion techniques. From this perspective the best interviewers (in terms of response rate) are those that are the most able to identify and manipulate to their advantage the psychological and social-psychological principles most relevant to each particular sample unit. (See the discussion of “tailoring” in Section 2.2.3.)

3.4.4 An Integrated Approach

Groves, Cialdini and Couper have formulated a model of survey response which attempts to incorporate all the various factors affecting the outcome of a request for survey participation (Groves and Cialdini, 1991; Groves, Couper and Cialdini, 1992; Groves and Couper, 1994; 1995). In this model (see Figure 6), the larger ‘Social Context’ in which the survey is embedded can be seen to influence and the decisions made by survey staff with respect to the ‘Survey Design’ are seen to influence both the respondent and the interviewer. In turn the ‘Respondent’ and ‘Interviewer’ both bring with them various background characteristics and psychological factors with them which affect how each behaves during the ‘Respondent-Interviewer Interaction’ which takes place on the doorstep, thereby leading to the respondent’s decision to participate or not. This theoretical model provides a basis from which researchers can investigate error in survey measurement from an integrated and holistic perspective. (Although this thesis mainly

focuses on the respondent and interviewer cells, this theoretical model nonetheless provides a useful context for the research. It is revisited in the context of the actual data available for this thesis in Chapter 5, see Figures 14a and 14b.)

3.5 Nonresponse Trends

The final important point to note about nonresponse is the fact that response rates to general population surveys have tended to fall over recent decades (c.f., Steeth, 1981; Lievesley, 1986; Goyder, 1987; Bradburn, 1992). Trends in living patterns and attitudes have increased the problems faced by research agencies in obtaining acceptable levels of response in two main ways. It has become more difficult to find people at home (Weber and Burt, 1972; Weeks *et al*, 1980); and it has become more difficult to persuade people, once contacted, to take part in interview surveys (Steeth, 1981). Some government continuous surveys in both the States (see Groves, 1989) and the UK (see Lievesley, 1986) have maintained their overall rates of response, but even in these cases refusal rates have tended to rise, being compensated by lower non-contact rates achieved at higher cost in time and money than many other surveys can afford. This point is illustrated by Figure 7 and Table 3. Figure 7 is taken from Groves (1989, p. 147) and presents data from the US National Health Interview Survey from 1967 to 1985. This figure suggests that total nonresponse has actually gone down over time. Looking at the detailed lines for refusals and other non-interviews (mainly non-contacts), clearly shows that rising refusal rate and the decreasing other non-interview rate.

Table 3 presents data from the UK General Household Survey as reported in Lievesley (1986). Here the fairly consistent rates conceal the measures taken to reduce nonresponse

such as raising the minimum number of calls per address, relaxing placement pattern controls and holding interviewer retraining and refresher schools.

3.6 Chapter Summary

This chapter has reviewed the literature with respect to survey nonresponse, looking at changes in society as well as evidence from the level of the respondent and the interviewer and culminating with the presentation of an integrated model of survey nonresponse which offers a sociological/psychological motivation for Chapter 5 through 8 of this thesis. In addition, this Chapter has

- stressed the importance of research into survey nonresponse by describing recent trends in its growth (see Section 3.5)
- clarified the definition of nonresponse and explored various issues in response rate calculation to ensure development of accurate dependent variables in Chapters 5 through 8 of this thesis, and
- looked specifically at what is meant by nonresponse bias (for the investigation in Chapter 5) and nonresponse variance (for the investigations in Chapters 6 and 8).

Figure 6: Factors Affecting Survey Participation: Theoretical Model from the Work of Groves, Cialdini, and Couper

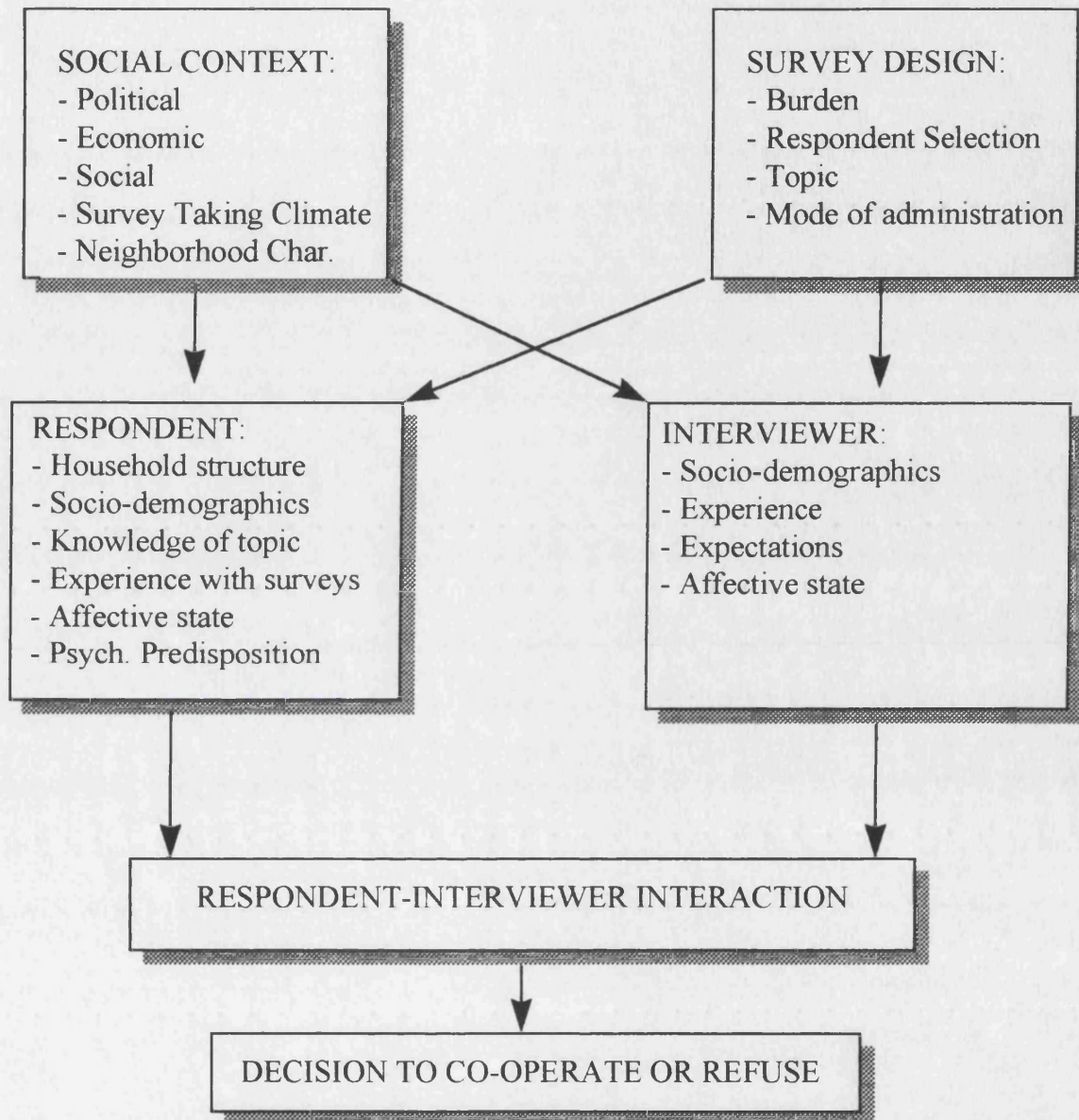
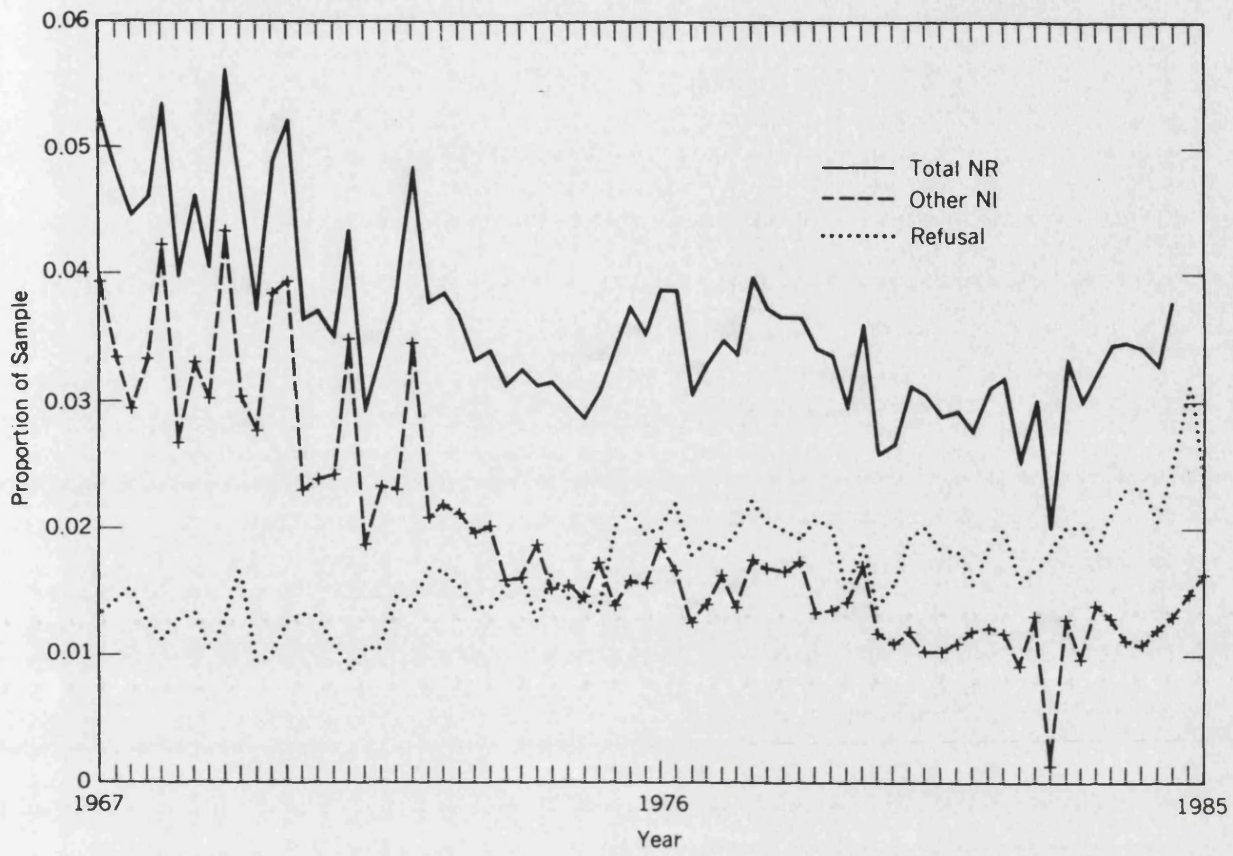


Figure 7: Nonresponse by Type for the National Health Interview Survey, 1967-1985



Source: Groves, R. (1989), *Survey Errors and Survey Costs*, New York: John Wiley & Sons, p. 147.

Table 3: UK General Household Survey Annual Response Rates

Year	Non-contacts (%)	Refusals(%)	Base (total effective sample)
1971	2.7	11.9	15432
1972	2.6	13.5	15307
1973	2.9	13.5	15360
1974	2.3	11.6	14232
1975	2.2	12.0	15327
1976	2.3	11.2	15310
1977	2.1	12.0	15315
1978	2.5	12.8	13957
1979	2.7	11.8	13437
1980	2.4	13.5	13943
1981	2.2	11.6	13939
1982	2.2	11.7	11970
1983	2.6	12.5	11862
1984	3.7	13.7	11867

Source: Lievesley, D. (1986), *Unit Nonresponse in Interview Surveys*, JCSM unpublished paper, London: SCPR (Data originally supplied by what was then OPCS.)

CHAPTER 4 EXPLORING RESPONSE ERROR

4.1 Background

Schrödinger's Cat Experiment

"According to the rules of quantum mechanics, the act of OBSERVATION would cause the cat to jump into either its dead or alive state" (Wilson, 1977, p.34)

As discussed in Section 1.1, the term 'response error' is often used synonymously with 'measurement error', i.e., error that is derived from the survey observation process itself. As suggested by the *Schrödinger's Cat* phenomenon (the 'Heisenberg Effect'), the mere act of observation can change the activities we wish to measure. As suggested by Hyman *et al* (1954, 1975), Rosenthal (1966) and others, this is a problem that has always plagued social scientists as well as natural scientists.

As described in Section 1.1, survey response error is typically defined as originating from four sources: Effects due to the interviewer, due to the respondent, due to the questionnaire, and due the mode of data collection. An extensive literature exists with respect to each of these four sources. As the focus of this thesis is on the response effects due to the interviewer, it could be argued that only an investigation into this one source is needed. For example, the impact of the interviewer from a social perspective has already been treated in Chapter 2, with the impact from a mathematical perspective being treated in Chapter 1. Rather than retrace this ground, this Chapter seeks to provide a larger context for the material in Chapters 1 and 2.

From the social perspective, it can be seen that the four sources of response error in a survey are not independent. The interviewer does not work in isolation; but is in constant interaction with the respondent and both interact with the questionnaire. In addition, the

mode of data collection can mark the absence of the interviewer (e.g., through postal and self-completion methods) as well as mark the limitations of the interviewer's channels of communication (e.g., in telephone as opposed to face-to-face interviewing).

Section 4.2.1 explores the interaction between the interviewer, and the respondent and/or questionnaire through reviewing the literature on models of the survey process which have been suggested over the years. The goal is to illustrate the complexities and subtleties of where response error can arise. This is followed by a brief discussion of the interaction between interviewer and mode of data collection (see Section 4.2.2).

From a mathematical perspective, response error due to interviewers was discussed in Chapter 1. Section 4.3 revisits this topic looking at response error more generally by comparing and contrasting the work of psychometricians and statisticians (see Section 4.3.1) and by discussing the effects of response error on estimates (see Section 4.3.2).

4.2 The Conceptual Perspective

4.2.1 Interviewers, Respondents and the Survey Task

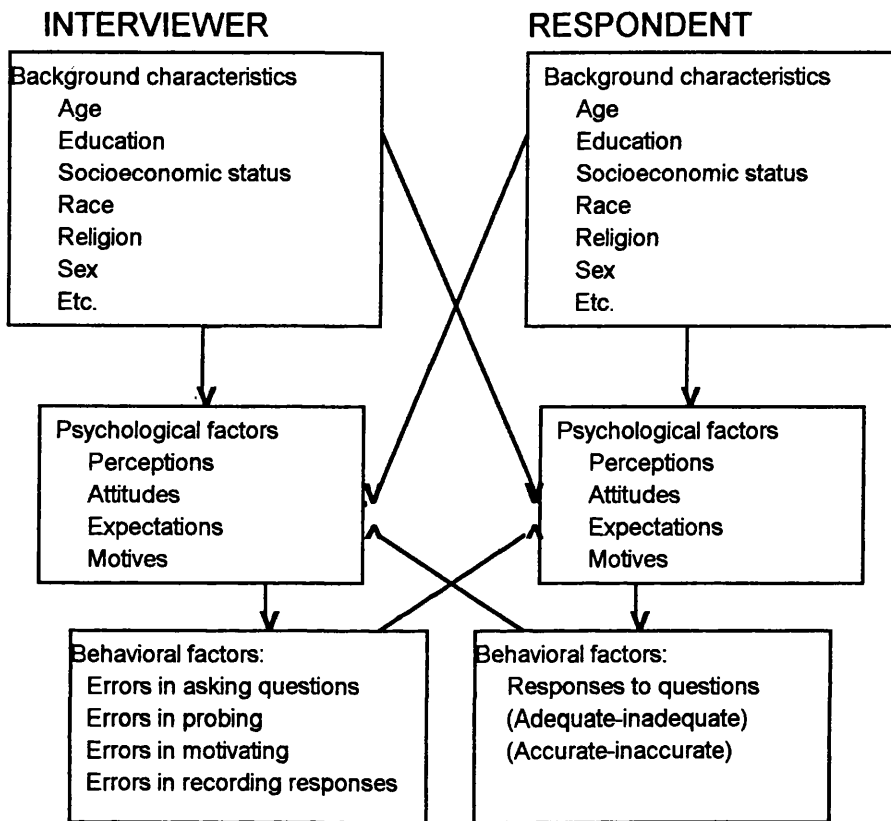
Through the years, various researchers have proposed conceptual models of the survey interaction process (see, for example, Kahn and Cannell, 1957; Sudman and Bradburn, 1974; Cannell, Marquis, and Laurent, 1977; Oksenberg and Cannell, 1977; Cannell, Miller, and Oksenberg, 1981; Dijkstra and van der Zouwen, 1982; Esposito and Jobe, 1991)⁸. As suggested by Dijkstra and van der Zouwen (1982, p.9), no one model has yet to be able to incorporate the “*diverse results of research on response effects*” and that

8 Note there was also a model proposed by Sceuch, 1967, which is not included in this review as the text is only available in German. See Scheuch, E.K. (1967), *Das Interview in der Sozialforschung*, in: R. König (ed), *Handbuch der Empirischen Sozialforschung*, Vol. 1, Stuttgart: F. Enke.

given the complexity of the survey interview interaction, “*it might very well be impossible to incorporate all variables of interest in a comprehensive interview theory.*” As Dijkstra and van der Zouwen (1982) go on to point out, these statements do not invalidate the attempts to construct an overall model, but do point to the enormity of the task. It is thus useful to review similarities and dissimilarities in what these various models have suggested.

Kahn and Cannell (1957, p. 194) describe “*a model of bias in the interview*” (see Figure 8). In particular they look at three levels of factors: Background characteristics, psychological factors, and behavioural factors. These are examined for both the interviewer and the respondent as well as the role of the interaction between these factors for these two players. The effects of these characteristics with respect to interviewers are discussed in Section 2.2. The Kahn and Cannell model proposes a similar array of background characteristics and psychological factors for respondents, and by design a different set of behavioural factors. Respondents’ main behaviour is seen with respect to how they answer the survey questions, with their response falling along the ‘Adequate to Inadequate’ dimension as well as the ‘Accurate to Inaccurate’ dimension. As indicated by the central arrows, it is through inferences generated from the background characteristics and behaviour of the other player “*that the interviewer and respondent form attitudes toward and expectations of each other*” (Kahn and Cannell, 1957, p. 194). Background characteristics and behaviours can be easily misinterpreted leading to “*inferences that are incorrect and dysfunctional for the interview process*” (p. 195). Examples of this are clearly seen in the results of sensitive surveys where the topic of the survey is clearly associated with given background characteristics of the participants (see Section 2.2.1).

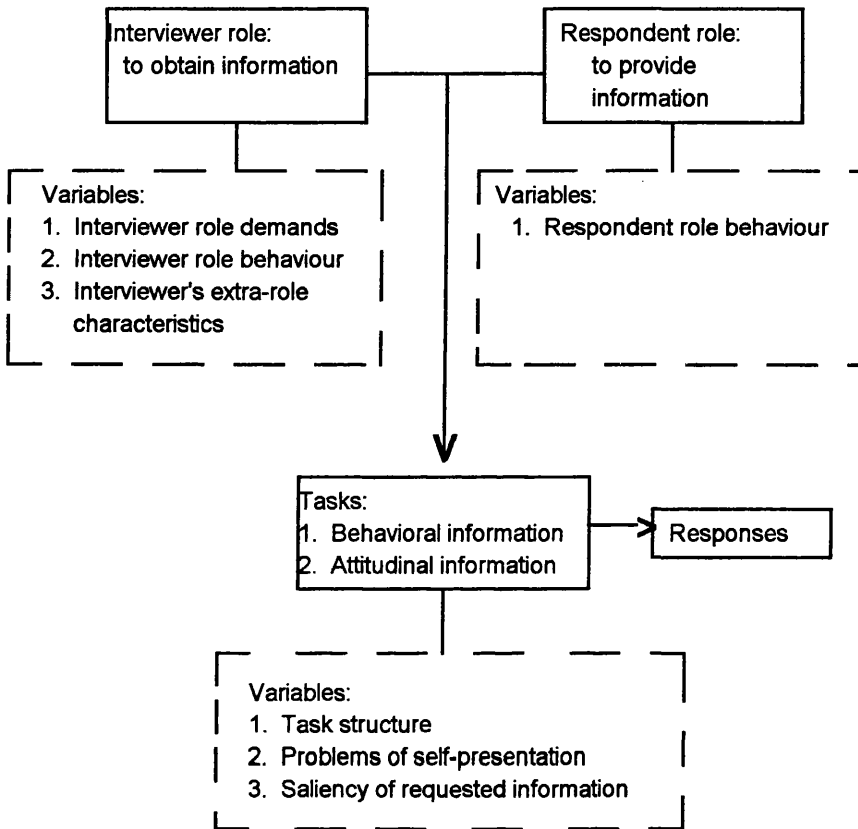
Figure 8: A Model of Bias in the Interview



Source: Kahn, R.L., and Cannell, C.F. (1957), *The Dynamics of Interviewing*, New York: John Wiley & Sons, p. 194.

The model proposed by Sudman and Bradburn (1974) (see Figure 9) draws on the work of Kahn and Cannell (1957) as well as Hyman *et al* (1954). As with the Kahn and Cannell model described above, it focuses on the *role of the interviewer* and the *role of the respondent*, yet there are also several differences. According to Sudman and Bradburn (1974, p. 13-14), the role of the interviewer includes role demands (“*the rules of behaviour which the interviewer is expected to follow*”), role behaviour (“*the degree of competence with which she carries out these role demands*”), and extra-role characteristics (such as “*background characteristics*”). Thus Sudman and Bradburn

Figure 9: A Simple Model of the Research Interview



Source: Sudman, S. and Bradburn, N. (1974), *Response Effects in Surveys*, Chicago: Aldine Publishing Co., p. 17.

(1974) focus on the distinction between the interview rules and the interviewer's behaviour and do not specifically mention the psychological side. In contrast while Kahn and Cannell emphasise the psychological side, they do not make a separate distinction for interviewer role demands. In addition, Sudman and Bradburn (1974) emphasise the role behaviour of the respondent and in particular the respondent's motivation for accurately participating in the survey, but exclude background characteristics of the respondent, as well as other psychological factors. The main distinction, however, between the two models is Sudman and Bradburn's addition of the survey 'task', a very useful addition which highlights the fact that the accuracy of respondents' answers do indeed depend on

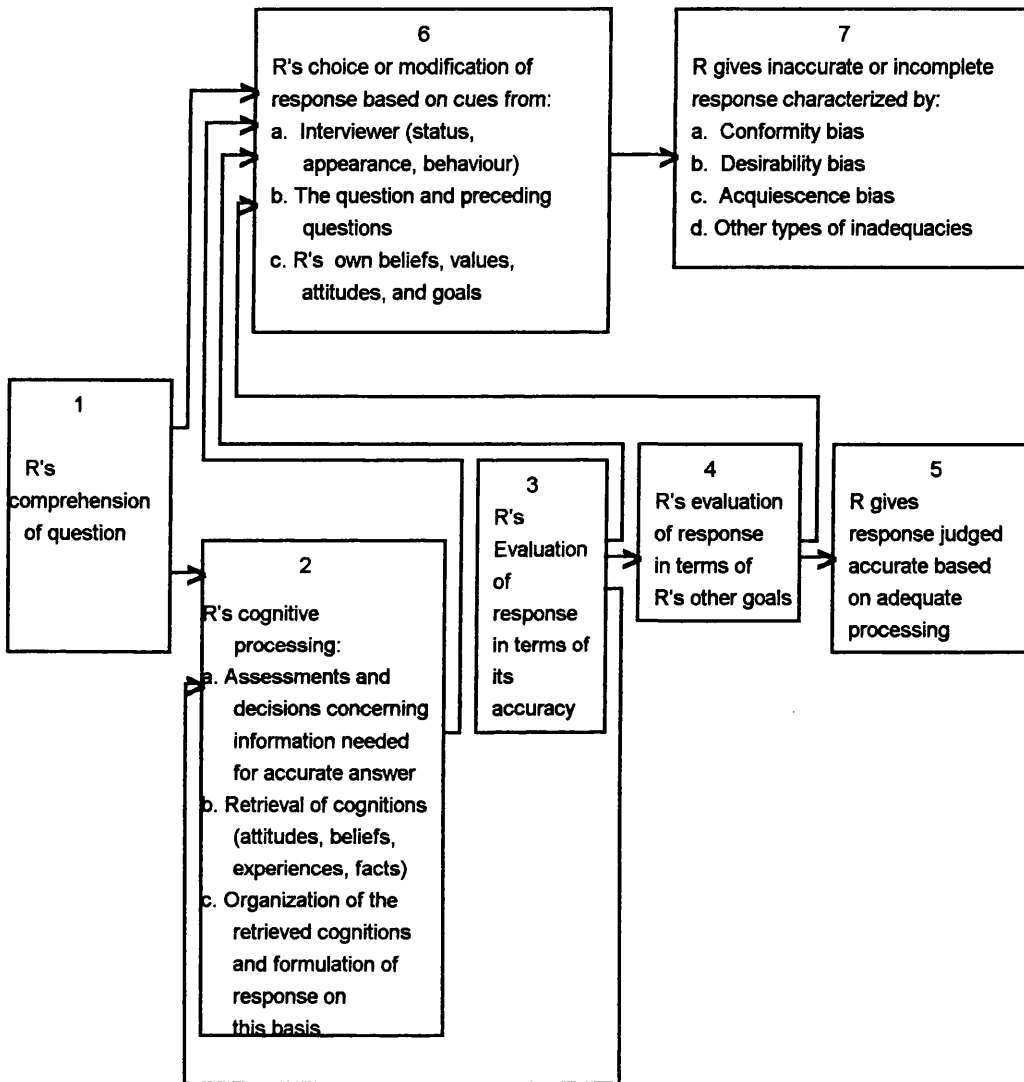
the survey question they are asked to answer. The task variables include the task structure which covers all the various aspects of the questionnaire from its mode of administration, length, format of the questions (open versus closed), location of the question within the questionnaire, etc. The task variables also include such issues as the problems of self-presentation which occur with sensitive topics (see also, Lee, 1993; Fowler, 1995) and the saliency of the requested information which can improve the recall of the desired information (see also, Mathiowetz and Duncan, 1988; Eisenhower, Mathiowetz, and Morganstein, 1991).

The model of the question-answer process originally proposed by Cannell, Marquis, and Laurent, 1977 was later modified by Oksenberg and Cannell, 1977 and Cannell, Miller, and Oksenberg, 1981. The latter version is shown in Figure 10. As Cannell, Miller, and Oksenberg (1981, p. 395) suggest,

“Once the respondent departs from the appropriate answering process (steps 1 to 5) and relies on other situational cues (steps 6 and 7), the response will exhibit some kind of invalidity (step 7). Researchers have labelled the effects on response of such situational cues as social desirability bias, acquiescence bias, and the like. It is sometimes argued that these biases result from the respondent’s personality characteristics, such as an “acquiescence trait,” a “social desirability trait,” or a need for approval – but we assume that the response process is more likely to be shaped by situational cues in the interview itself: from the interviewer, the questionnaires, or the organisation for which the research is being conducted.”

As opposed to the previous two models discussed, the Cannell, Miller, and Oksenberg model is clearly focused on the respondent and the cognitive processing involved. The ‘task’ has the potential to direct the respondent to Step 6 through cues from the question and preceding questions and thus negatively affect the response quality by inducing the respondent into any of the inaccuracies listed in Step 7. In a similar way, the impact of the interviewer is felt.

Figure 10: Diagram of Respondent's (R) Question-Answering Process



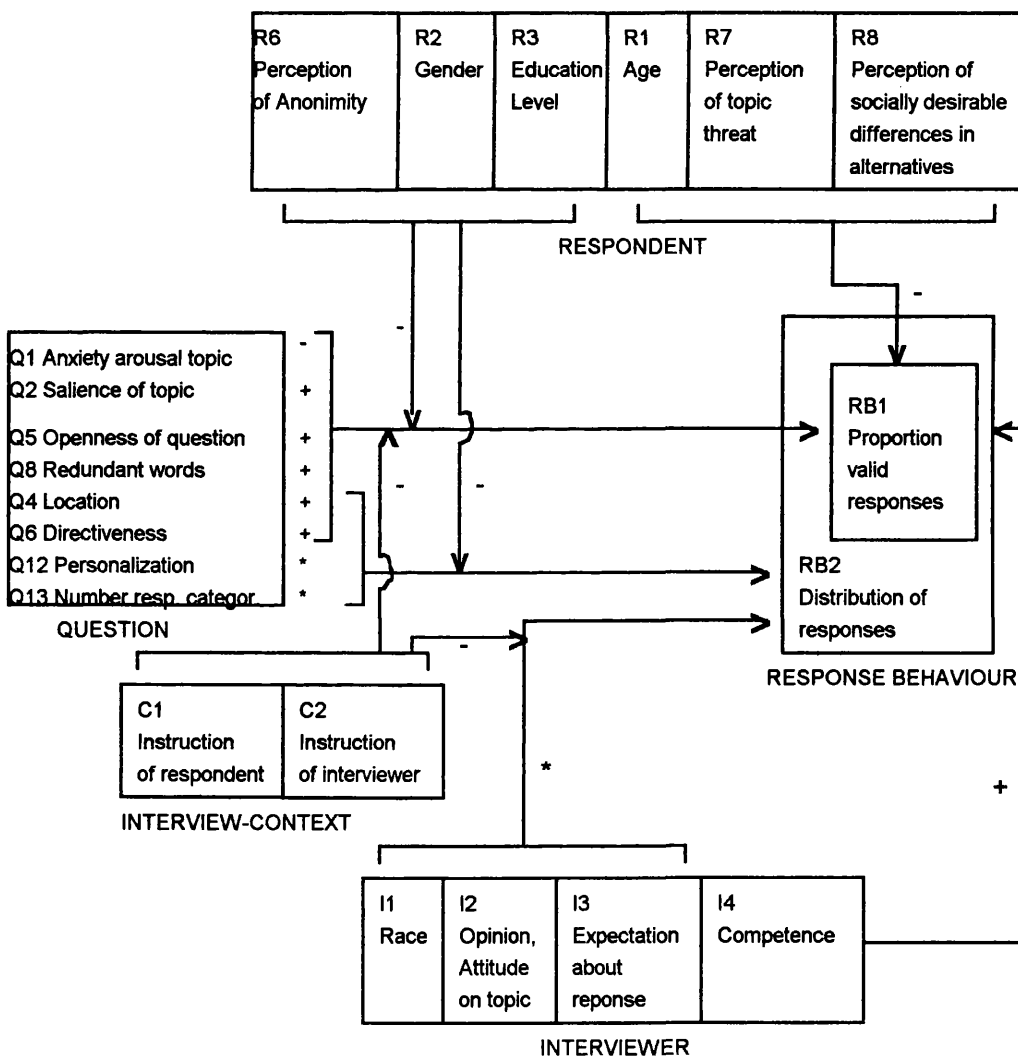
Source: Cannell, C.F., Miller, P.V., and Oksenberg, L. (1981), Research on Interviewing Techniques, in: S. Leinhardt (ed), *Sociological Methodology*, San Francisco: Jossey-Bass, p.393.

Dijkstra and van der Zouwen (1982, p. 9) refer to their model as a “*modest attempt*”.

Interestingly, they took a different approach to the others in that their model is constructed from the existing empirical evidence upward rather than from theory downwards! Their model is based on a summary of the research contained in Chapters 2 through 7 of their book and translates into “72 propositions that state the presence of a

relationship between a dependent and independent variable or the presence of an interaction effect” (p. 210). These propositions translated into a large number of variables, “26 characterise the question, 14 concern the interviewer, 10 the respondent, and 6 other aspects of the interview” (p. 210).

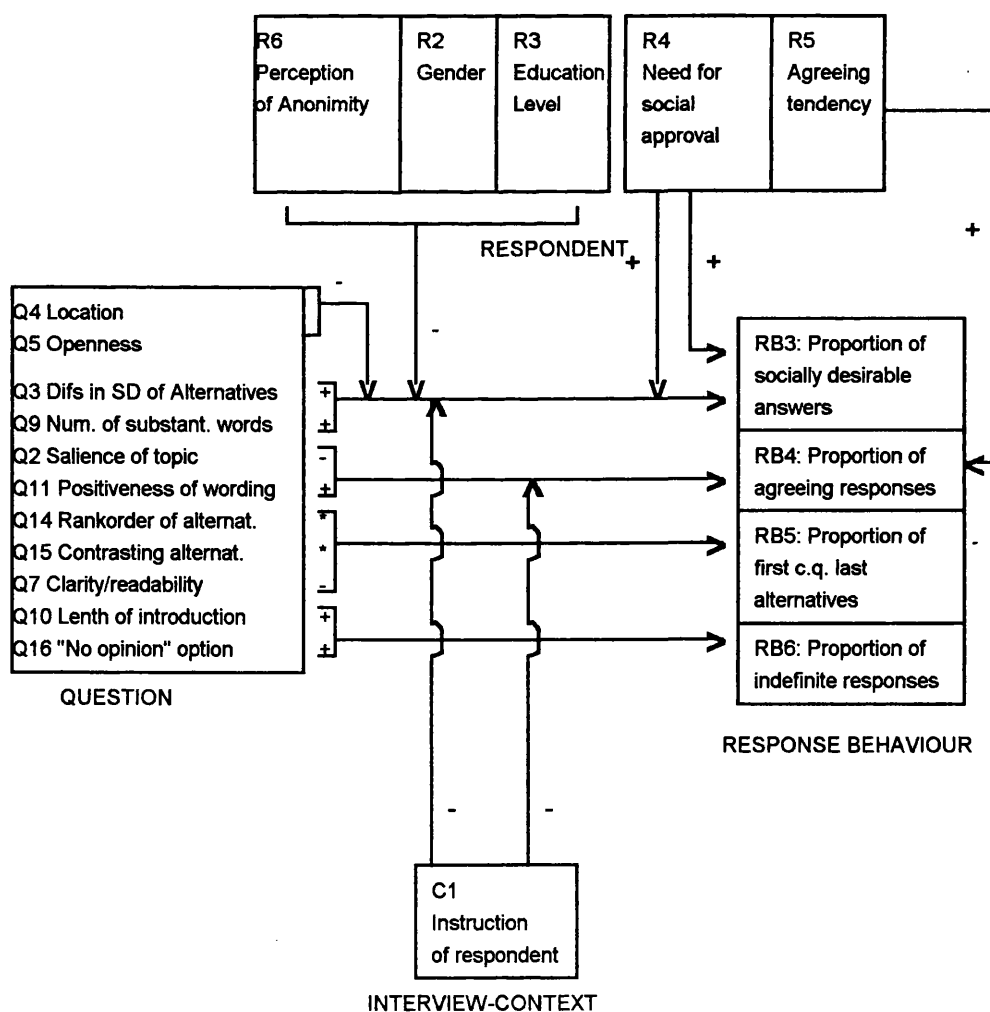
Figure 11a: Relations Among Characteristics of Questions, Interviewer, Respondent, and Interview-Context on the One Hand, and Proportion of Valid Responses and Distribution of Responses on the Other Hand



Source: Dijkstra, W. and Van der Zouwen, J. (1982), *Response Behaviour in the Survey-Interview*, London: Academic Press, p. 215.

Focusing on what they considered the key aspects of response behaviour, the list was reduced to “16 question characteristics (Q_1 - Q_{16}), four interviewer characteristics (I_1 - I_4), eight respondent characteristics (R_1 - R_8), and two variables characterising the context of the interview (C_1 and C_2)” (p.211) all of which had been shown to be related to the six dependent variables (RB_1 - RB_6). This was then translated into a diagram. Due to its complexity, they have shown their diagram in two parts: the first part covering the first two dependent variables, the “proportion of valid responses” (RB_1) and the

Figure 11b: Relations Among Characteristics of Questions, Interviewer, Respondent, and Interview-Context on the One Hand, and Some Characteristics of Response Behaviour on the Other Hand



Source: Dijkstra, W. and Van der Zouwen, J. (1982), *Response Behaviour in the Survey-Interview*, London: Academic Press, p. 216.

“distribution of responses” (RB₂); and the second covering the remaining four dependent variables, the *“proportion of socially desirable answers”* (RB₃), the *“proportion of agreeing responses”* (RB₄), the *“proportion of first or last mentioned response alternatives”* (RB₅), and the *“proportion of indefinite responses”* (RB₆) such as don't know or no opinion. These are displayed in Figures 11a and 11b, respectively. In the figures, an arrow pointing to a box indicates a direct effect and an arrow pointing to a line indicates an interaction effect. Note that the original diagrams given in Dijkstra and van der Zouwen (1982) also contained code numbers which indicated the propositions represented by each of the relationships. For the aim of simplicity, these are not shown in Figures 11a and 11b.

As can be seen by Figures 11a and 11b, the Dijkstra and van der Zouwen (1982) model bears the most similarity to the Sudman and Bradburn (1974) model in their focus on the interviewer, the respondent and the task. (It should be noted here that 'interview-context' does not refer to mode of data collection.) As the Dijkstra and van der Zouwen (1982) model was generated from empirical research, it has the considerable advantage of allowing much greater specificity of precisely what aspects of interviewers, respondents, and questions, lead to response error.

The model suggested by Esposito and Jobe (1991) is shown in Figure 12. This schematic perspective traces all of the steps in the survey interaction process in comparison to the Cannell, Miller, and Oksenberg (1981) model which focuses on an individual question-answer process. The Esposito and Jobe model is backed up by what they call *“a taxonomy of contextual variables to be considered in the analysis of the survey*

interaction process” (Esposito and Jobe, 1991,p. 540). This taxonomy represents single-space text covering two A4 sheets. The key points, however, are summarised in Figure

Figure 12: A Model of the Survey Interaction Process

Phase Sequence:

1. Interviewer and Respondent Orient Themselves Within Survey Context
2. Interviewer Asks Question
3. Respondent Processes Question and Provides Answer
4. Interviewer Processes and Records Respondent’s Answer
5. Interviewer and Respondent Reorient Themselves and Proceed to Next Questions [Recycle to Phase 2, or proceed to Phase 6]
6. Interview Is Concluded
7. Interviewer Reviews and Adjusts Questionnaire Protocol

Schematic Diagram:

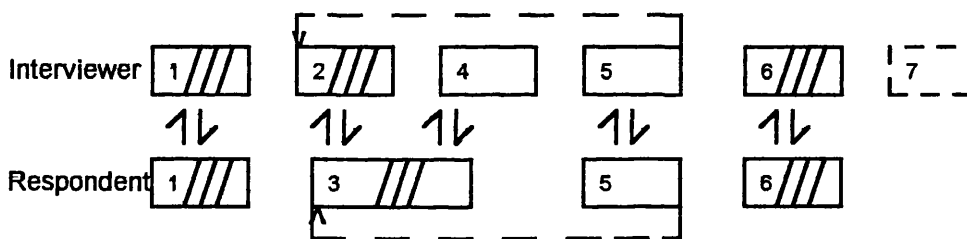



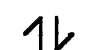


Diagram Key:

-  Phase with Directed Verbal Communication
-  Information-Processing Phase
-  Optional Phase
-  Bidirectional Verbal and/or Nonverbal Communication

Source: Esposito, J. and Jobe, J. (1991), A General Model of the Survey Interaction Process, in: Proceedings of the 1991 Annual Research Conference, Washington DC: U.S. Bureau of the Census, p. 548.

13 below. These contextual variables cover background and psychological characteristics of the interviewer and the respondent as seen in previous models. These contextual variables allow for the Esposito and Jobe (1991) model to be the broadest of the models by introducing many influences previously unconsidered (such as interview setting, timing

of contact, survey publicity method, survey sponsorship, mode of administration, and incentives) but curiously omitting explicit reference to the possible variable levels of

Figure 13: Highlights of the Esposito and Jobe (1991) Taxonomy

Contextual Variables to be Considered

- **Interview Setting:**
Location of interview; characteristics of location
 - **Timing of Contact**
Events affecting person (state/national/global; local/community; personal)
 - **Respondent/Interviewer Characteristics**
Demographic, socio-economic, organismic characteristics (biophysical status, psychological, other), and experiential characteristics
 - **Survey Publicity Method**
Advance letter, 'cold contact', use of media, word of mouth
 - **Survey Sponsorship**
Government/University/Private
 - **Survey/Interview Administration Model**
Face-to-face/telephone/self-administered
 - **Response Security Level**
Complete anonymity, confidentiality, restricted availability, unrestricted availability
 - **Attributes of the Survey Instrument**
General questionnaire characteristics (content, length, pace, homogeneity, amenability to self vs. Proxy) and item characteristics (target, focus, content and salience, desirability and sensitivity, response format, length and complexity)
 - **Incentives**
Monetary incentives, tangible non-monetary incentives, verbal incentives
-

competence in how interviewers carry out their roles. The list in Figure 13 is a useful summary of what is known in the survey literature about the various influences on response error aside from the interviewer. However, it could be improved by more

explicit reference to the cognitive stages of question comprehension, recall and judgement (Tourangeau, 1984) with respect to question characteristics. For example, item characteristics could specifically mention the ambiguity of question terms and concepts which lead to poor respondent comprehension (see Belson, 1981), the length of the recall period required by the question, and the complexity of the task posed for the respondent. Of more importance is to somehow communicate the interaction between question characteristics and respondent cognitive processes and past experiences. For example, determining the number of doctor visits in the last year may be easy for respondents with little or no contact with the doctor, but extremely difficult for those with chronic illnesses, at which point it is useful to know the respondents' retrieval and judgement strategies in order to evaluate the accuracy of their answers.

Such a broad taxonomy such as the Esposito and Jobe one has its place even though it is in complete contrast to the specificity introduced by the Dijkstra and van der Zouwen (1982) model. Ideally, what is needed is a model which incorporates the 'specific' relationships for each of the points from a complete taxonomy.

In summary, these models of the survey interaction process demonstrate the complexity of effects which are occurring in the interview setting and how the effects of the interviewer are tempered by the effects of the respondent and the survey task itself.

4.2.2 Interviewers and the Mode of Data Collection

Reference to the mode of data collection was only given fleeting mention in the Section 4.2.1 through the models of Sudman and Bradburn (1974) and Esposito and Jobe (1991). This section explores some of the issues more deeply.

The three main modes of survey data collection are face-to-face interviewing, telephone interviewing, and the self-completion mode (typically postal surveys). One of the main disadvantages of the latter is actually the absence of the interviewer who controls who responds, question order and how the respondent answers through motivating the respondent and probing to ensure good quality answers.

Telephone and face-to-face surveys (as summarised by de Leeuw, 1992), differ with respect to several media related factors. For example, people tend to be differentially *acquainted* with providing information in face-to-face settings (e.g., to doctors, teachers, supervisors) as opposed to the telephone setting. The second media factor is *locus of control*. Interviewers are more in control in the telephone situation as “*traditional rules of behaviour dictate that the initiator of a telephone conversation . . . controls the channel and regulation of the communication*” (p. 15) in the telephone interview. In contrast, *silences* in the telephone mode can lead to uncomfortable moments which are not felt in the face-to-face situation and *sincerity* of purpose is harder to communicate over the telephone as opposed to in person.

Technical aspects of information transmission also differ between the two modes. In the face-to-face setting, interviewers and respondents have access to verbal, paralinguistic and non-verbal behaviour cues. (Paralinguistic communication is concerned with non-verbal but auditory signals which can be picked up from the voice such as emotional tone, timing, emphasis, as well as utterances such as ‘ahhhh’ and ‘mhhmm’ (c.f. Argyle, 1973). Telephone interviews prevent the use of non-verbal cues (which clarify the participants’ intentions and meanings as well as signalling turn taking behaviour) and also prevent the

use of visual stimuli. Each of these limitations make the telephone interviewers' role more challenging than that of the face-to-face interviewer.

4.3 Measurement of Response Variance

The literature on response error can be seen as coming from two streams. The first being sampling theory and the second, psychometric theory. This section looks at the measurement of response variance from these two perspectives.

The terms response bias and response variance come from the sampling statistician's perspective (see Section 1.1). The closest comparable terms from the psychologist's side would be invalidity and unreliability, respectively, where validity refers to the extent to which the measurement process is measuring the theoretical construct it is intended to be measuring⁹ and reliability refers to the extent to which the measurement process provides consistent results (Bohrnstedt, 1983). Thus, *broadly speaking*, unreliability is to variance as invalidity is to bias (see, Kish, 1965, p. 510).

From the sampling statistician's perspective, the presence and magnitude of response bias can only be assessed by comparing the survey data to external validation data. As will be seen from the discussions in Section 4.3.1, the measurement of bias (invalidity) takes on a very different meaning from the psychometrician's perspective. From the sampling statistician's perspective, the impact of response variance can be studied with the

9 This is the definition of *theoretical (construct) validity*. Psychometricians recognise several other categories of validity as well. For example, *empirical validity (criterion validity, predictive validity, concurrent validity)* which reflects the correlation of a particular measure with another variable. *Content validity* refers to the degree that one has representatively sampled from the domain of meaning that a particular construct is intended to measure (Bohrnstedt, 1983). In their multitrait multimethod approach, Campbell and Fiske (1959) define *convergent validity* and *discriminant validity*. Bailey (1978) discusses *face validity* which is the investigator's subjective judgement that the instrument is measuring what it is supposed to be measuring.

approaches of *interpenetration* and *re-enumeration* (see fuller discussion of these terms with respect to interviewer variance in Section 1.2). Psychologists employ *re-enumeration* in their measurement of reliability (c.f. a test-retest design: Carmines and Zeller, 1979; or a parallel measures design: Bohrnstedt, 1983).

4.3.1 *Mathematical Treatment*

Response Errors at the Level of the Individual From both the sampling statistician's and the psychometrician's perspective, the simplest response error model consists of a single observation y on a randomly selected respondent j which is measured with error (Biemer and Stokes, 1991). Subtracting the error from the observed score implies that there should be a true value for person j .

$$y_j - e_j = \tau_j \quad (38)$$

where $\tau_j =$ the actual 'true value' of that observation and

$e_j =$ the error of measurement.

The concept of 'true' values, however, has received a certain amount of debate. There are two extremes. On one end of the continuum, one may consider that a true value is always present. It is what one aims for despite the inevitable fallibility of measurement. At the other end there is "*a cluster of relativistic or solipsistic positions that hold that no true value exists out there*" (Turner and Martin, 1984, p. 98). The pragmatic path searches for a highly regarded mode of measurement which can be an operational substitute for the true value. Alternatively, multiple methods are sometimes used to try to

triangulate on the true value, an approach which can fall considerably short if all methods are subject to the same systematic error.

From the psychometrician's point of view (see, for example, Lord and Novick, 1968; Bohrnstedt, 1983) there is the consideration of platonic versus classical true values. The platonic approach hypothesises a "*unique naturally defined*" true value for each person (Lord and Novick, 1968, p. 28). The conception is reasonable for the natural sciences and for many factual survey questions, but some consider that it breaks down in the case of subjective phenomena such as attitude questions and quasi-factual concepts such as looking for work (Bailar and Rothwell, 1984). The classical approach avoids the problem of a non-observable true value by assuming that if someone could be measured an infinite number of times (assuming identical measurement conditions and of course no recall bias) then the average, or expected value of this distribution would be equal to that person's true score. In practice given the limited ability of survey researchers to actually be able to verify the responses of their surveys with an external source, it is most useful to think in terms of classical rather than platonic true scores.

From both the sampling statistician's and the platonic true score perspective, the simplest response error model consists of $y_j = \tau_j + e_j$ as shown in equation (38). Under hypothetical repetitions (trials t) of the interview process under the same essential survey conditions, we have

$$y_{jt} = \tau_j + e_{jt} \tag{39}$$

Here we can see that the true value, τ_j , is assumed to be a constant as it is not dependent on t -th trial, however, person j 's observed response may still vary due to error in the measurement process. Under the idea of infinite repetitions the variability in the measured response will have a well-defined, though quite likely unknown, probability distribution. Similarly, the e_j 's will have an error distribution. As suggested by Biemer and Stokes (1991), it is assumed that the $E(e_j|j) = 0$, $Var(e_j|j) = \sigma_j^2$, $Cov(\tau_j, e_j) = 0$, and $Cov(e_j, e_{j'}) = 0$ for $j \neq j'$. From these assumptions it follows that

$$\sigma_y^2 = \sigma_\tau^2 + \sigma_e^2 \tag{40}$$

From the *classical* true score perspective we start with

$$y_j = \tau_j^* + e_j \tag{41}$$

where $\tau_j^* = E(y_{jt})$, that is, is the mean of person j 's response distribution over an infinite number of trails (psychometricians call this variation an individual's 'propensity distribution')

$e_j =$ the error of measurement.

Thus by definition we have $E(e_j|j) = 0$. As Biemer and Stokes (1991) point out, if the other assumptions hold, then equation (41) is equivalent to equation (38).

Under the *sampling statistician's* and *platonic* approach one can test whether

$E(e_j|j) = 0$ is true and as suggested by Hansen, Hurwitz, and Bershada (1961) one can

define the response bias B as

$$\beta = E(e_j|j) = E(\tau_j) - E(y_j) \quad (42)$$

As noted above under the *classical* true score perspective, $E(e_j|j) = 0$ is true by definition. At first glance this seems to suggest that bias is ignored from this approach. But Biemer and Stokes (1991) point out that under the classical true score approach, response bias is modelled by specifying that

$$y_{jt} = \tau_j'' + e_{jt} \quad (43)$$

where $\tau_j'' = \tau_j + M_j$ and

$M_j =$ some type of method effect.

Thus response biases are modelled as components of variance.

Interviewer Effects.¹⁰ An interviewer may influence the answers to the survey in a number of different ways (see Chapter 2). Some of these may result in random noise (simple response variance, see Section 1.2.3) which is captured by the e_{jt} term. Other interviewer influences, however, may be systematic across respondents. This is captured by expanding the simple model to include an additional error term (b_i) which represents

10 This effect is not unique to interviewers as other survey personnel such as coders, supervisors, and data processors can also induce correlation among the reported values for the units they handle.

the systematic effects of the interviewers and does not vary over trials. This results in an individual level bias term because over infinite replications under the same survey conditions the expected value of the y_{jt} is no longer equal to the true value y_j , but instead $y_j + b_i$. The second error term (e_{jt}) still represents random fluctuations from trial to trial and results in an individual level variance term. This model is essentially relaxing the assumption that the covariance between error terms is zero. Thus we have

$$y_{ijt} = \tau_j + b_i + e_{jt} \quad (44)$$

where $\tau_j =$ the true value for the j -th individual,

$b_i =$ an individual bias term introduced by the interviewer(i), and

$e_{jt} =$ a random error term which can fluctuate from trial to trial.

By including various respondents (j 's) randomly assigned to various interviewers (i 's), this expanded model can be interpreted from the ANOVA perspective (see Section 1.2.1). This assumes the b_i represent the systematic effect of interviewer i to push the responses in a particular direction. Any effect which is common across all interviewers would result in an overall bias. It is generally assumed, however, that the $E(b_i) = 0$. That is, individual interviewer biases are assumed to balance out in the aggregate and can be considered compensating biases. These, in turn, are measured by variance statistics (see Section 1.2.1).

Interviewer influence can also be viewed from the perspective of the correlation model (see Section 1.2.2). The systematic component of error can be seen to introduce a correlation between the error terms for different respondents of the interviewer thereby

violating the assumption that $Cov(e_j, e_{j'}) = 0$, for $j \neq j'$. This assumption can be maintained by separating out the interviewer component as shown in equation (44).

Response Errors at the Population Level. When thinking at the population level, rather than at the individual level, it is helpful to consider the measurement on a given individual as coming from two stages of random sampling: the sampling of errors within individuals from an infinite population of errors as well as the sampling of an individual from a finite population of individuals (Biemer and Stokes, 1991). This is a useful distinction as the individual errors have nothing to do with the probabilities of selection assigned to the sampling units by the survey design.

4.3.2 The Effects of Response Error on Estimates

The effects of simple response variance and correlated response variance due to interviewers was shown in Section 1.2.3. For example, the effect of simple response variance is seen to inflate the variance of \bar{y} . (Note that when \bar{y} is a proportion, however, both the sampling and simple response variance can not exceed the variance formula, pq/n -- see Hansen, Hurwitz and Bershad, 1961.) The correlated response variance can lead to a sizeable inflation of the variance of \bar{y} (for both continuous variables and proportions) by the factor $[1 + \rho(k - 1)]$, where ρ is the intra-interviewer correlation coefficient and k is the interviewer's workload size.

The purpose of this section is to broaden this perspective. First we need to consider how sources of response variance (other than interviewers) affect the variance of \bar{y} . The effects due to simple response variance (unreliability) hold true for the variance due to the respondent,

the question, and the mode of data collection. Correlated response variance can be observed among the answers given by a particular respondent.¹¹ In the presence of all of these sources of simple and correlated response variance, a 95 percent confidence interval around \bar{y} is no longer

$$\bar{y} \pm 1.96 \cdot \sqrt{\text{Sampling var}}, \quad (45)$$

but rather

$$\bar{y} \pm 1.96 \cdot \sqrt{x \frac{(\text{Sampling Var} + \text{Simple Interviewer Var} + \text{Simple R Var} + \text{Simple Question Var} + \text{Simple Mode Var})}{(\text{Correlated Interviewer Effect})(\text{Correlated Response Effect})}} \quad (46)$$

It should be remembered as noted in Section 1.2.3 that re-enumeration over the type of element studied is necessary in order to estimate each of the simple response variance components and interpenetration over the elements studied is necessary to estimate the correlated response variance components.

Second, is the need to explore the effects of response variance on other estimates than \bar{y} . As suggested by Bohrnstedt (1983) and O’Muircheartaigh (1977) among others, the effects of simple response variance (unreliability) can be summarised as follows:

- In two-variable cases, the observed correlation and regression coefficient are smaller than the true coefficients (resulting in attenuated estimates of the relationship),
- In the k-variable case, the observed partial correlation coefficients and regression coefficients are usually, but not always, smaller than the true coefficients, and
- Unreliability of the explanatory variables can lead to the breakdown of the regression model, because the distance between units may no longer be equal (i.e., no longer reflective of an interval or ratio scale).

11 Coders and other survey data processing personnel can introduce correlated variance, but these are typically classified as correlated processing variance (as opposed to correlated response variance).

Second, it is also useful to note that

- Simple response variance can in principle be estimated from the survey observations themselves (given the presence of re-enumeration, e.g., a test-retest situation or on a single occasion, the presence of multiple questions on the same construct), and
- The effect of simple and correlated response variance can be changed by sampling a larger number of the units involved.

Third, is the need to consider response bias. Although under both the psychometric and sampling perspective biases can be modelled as components of variance (see Section 4.3.1), it is worth pointing out some further properties of response bias as typically seen from the sampling statistician's perspective:

- Response bias is a constant which cannot be measured from within the survey,
- The effect of response bias is fixed regardless of the number of observations taken, and
- Response bias affects means and totals (for the comparison of subclass means and for measures of correlation and association, the effect may be slight).

4.4 Chapter Summary

As described in Section 1.1, response error, often called measurement error, has four major sources: the interviewer, the questionnaire, the respondent, and the mode of data collection. In reality these four effects are not separate but interactive. Thus the effects of the interviewer can not be studied in isolation without an understanding of the interconnectedness and complexity of a typical survey interview. This Chapter has provided insight into this interconnectedness by exploring the psychological and sociological literature with respect to models of the survey interaction process. Five such

models were compared and contrasted. This larger perspective will help with the interpretation of the findings from Chapters 9 and 10.

Response error was then revisited from a mathematical perspective, exploring both the approaches of sampling statisticians and psychometricians. This has provided a larger context for the interviewer effects models discussed in Section 1.2 and interviewer effects are seen in reference to the other sources of survey response error.

PART 2 EMPIRICAL INVESTIGATIONS INTO INTERVIEWERS AND NONRESPONSE ERROR

CHAPTER 5 CORRELATES OF NONRESPONSE

5.1 Background

The purpose of this Chapter is to use BHPS data to explore the various factors that predict household and individual level nonresponse.

5.1.1 Characteristics of Households and Individuals

Groves (1989) notes that the extent of nonresponse bias is a function of both the proportion of nonrespondents and the difference in the estimate of interest that would be obtained from the respondents and from the nonrespondents (see Section 3.2).

Determining differences between respondents and nonrespondents, however, is not straightforward, because by definition nonrespondents are those units which are not measured. Groves (1989, p. 186-187) describes six approaches which have been used to gain information about the attributes of nonrespondents, all of which have disadvantages. These include (1) special studies of nonrespondents (which typically can suffer from nonresponse), (2) using information from sampling frames (which depends on the amount of information available on the frame), (3) asking others about nonrespondents or having the interviewer provide information (which depends on the accuracy of the information/observations), (4) comparison of the characteristics of those interviewed by the n -th call with those found at subsequent calls (which assumes that nonrespondents still not reached at the end of the survey resemble nonrespondents reached late on in the fieldwork process and is less useful for describing refusers), (5) comparison with more complete data such as the Census (which only provides information at the aggregate level and assumes that the 'complete' data are errorless and use the same definitions and question wordings), and (6) studying persons who drop out of a panel survey (which is

excellent for studying attrition nonrespondents, but may or may not generalise to nonrespondents from one-off surveys).

Studies of these kind have amassed a certain amount of knowledge about the likely characteristics of nonrespondents. As described in Section 3.4.2, for example, nonresponding units are often single family households, individuals with a low educational background or with a low intelligence quotient, the young and the elderly, men, non-white individuals, the non-married, the self-employed and in some instances the unemployed, renters, those less well off economically but also sometimes the very well off, the highly mobile, those who do not feel part of the community, those not committed to a particular political party, those with strict conservative views, those living in urban areas or areas with a high population density or fear of crime, households of unrelated sharers, and those more likely to answer “don’t know” in previous surveys. Yet, more clarity in this area would be useful especially given that some studies have shown that the demographic characteristics of those people who refuse to participate in a survey differ from those who can not be found (see Section 3.4.2 for references).

5.1.2 The Data: Moving Beyond Characteristics of Households and Individuals

As a panel study, the BHPS data offer the ideal opportunity to model the characteristics of Wave 1 households and individuals who turned into nonrespondents at Wave 2. Thus the data are well placed to contribute to both practical and theoretical discussions of nonresponse. (As discussed in Section 5.1.1, there is a concern that attrition nonresponse may not generalise to initial nonresponse, but as suggested by Section 5.5.2 this concern may not be problematic.) More importantly, the auxiliary data sources (see Section 1.3.4) offer a rare opportunity to extend the traditional analysis of the characteristics of

households and individuals to also include the characteristics of interviewers, the characteristics of areas, and the characteristics of the call process itself. Each of these auxiliary sources will be discussed in turn.

As described in Section 2.2, it is useful to divide the influence of the interviewer into three categories: background characteristics, psychological factors, and behavioural factors (which includes years of experience). The auxiliary information provide data on two key aspects of interviewer's background characteristics (gender and age) and data on their years of experience and grade level. The effect of interviewer continuity can also be tested, but this is described in Chapter 8. The other effects of the interviewer on survey nonresponse, however, will be detectable through random effects terms in cross-classified multilevel models (see Chapter 6).

As described in Section 3.4.1, past research on area effects has focused on the impact of urban areas with high population density which in turn could be connected to a number of factors concurrent with such environments such as fear of crime, greater alienation from other people, etc. (see, for example, Marquis, 1977; Bergman, Hanve, and Rapp, 1978; House and Wolf, 1978;). The 1991 Census small area statistics file provides the opportunity to look at a wide range of socio-demographic and economic characteristics of areas.

Our final source of auxiliary data is data from the call-records themselves as well as interviewer observations about respondent co-operation. There is little previous research which has used call record patterns from a previous wave to predict later wave nonresponse. This is probably to do with the fact that most surveys do not key this

information. A few studies, however, have looked at respondent co-operation as measured by the interviewer (see, for example, Kalton, *et al*, 1990; Laurie *et al*, 1997). (Interviewers' judgements about respondents' co-operation has also been shown to be a good predictor of problems with response quality. See, for example, O'Muircheartaigh, 1984a; 1984b).

5.2 Returning to the Theoretical Model of Groves, Cialdini, and Couper

Given the many disparate factors described in Section 5.1.2, it would be useful to have a unified theoretical model which integrated them. The best example of such a model in the nonresponse literature is the one developed by Groves, Cialdini, and Couper (see Groves and Cialdini, 1991; Groves, Couper and Cialdini, 1992; Groves and Couper, 1994; 1995) which was presented as Figure 6 in Section 3.4. In a broad sense, it covers all of the factors discussed in Section 5.1.2. For example, it includes the three main influences of interviewers: background characteristics, psychological factors, and behavioural factors, with the latter being captured by the *Respondent-Interviewer Interaction* cell and it mentions neighbourhood characteristics underneath the heading of the *Social Context* in addition to the characteristics of the *Respondent*. In a specific sense, however, it needs to be adapted in order to make it directly useful for this thesis. This 'adaptation' process is shown in Figures 14a and 14b. Note that new elements within cells are indicated with italic text, new cells are indicated with a double line and no shadow, and new causal lines are indicated with dashed lines.

First, as shown in Figure 14a, it was necessary to consider a model which included non-contacts as an outcome in addition to refusals. Both *Interviewer* and *Respondent* factors were seen to determine the *Likelihood of Finding Someone at Home*, which in turn results in either a *Complete Non-contact* or ensuring that a *Respondent-Interviewer Interaction*

will take place. Second, ‘interviewer continuity’ (see Section 1.3.2 and 1.3.3) was added to the *Survey Design* cell. Third, the various geographic areas characteristics taken from 1991 Census small area data were added under Neighbourhood Characteristics in the *Survey Context* cell. Fourth, an additional cell was added to explicitly include *Household Factors* (in addition to individual factors). The household level factors were seen to be influenced by the general *Social Context* and in turn to influence individual *Respondent* reactions. Fifth, in the *Respondent* cell, ‘socio-demographics’ was extended to cover economic characteristics as well. Sixth, a new cell reflecting *Past Respondent-Interviewer Interactions* was added as respondents’ decisions to participate can be influenced by the interaction with the interviewer for the current survey request as well as by interactions with past interviewers for past survey requests. These *Past Respondent-Interviewer Interactions* were also seen to influence the *Likelihood of Finding Someone at Home* as interviewers can utilise information about their last contact with the respondent to facilitate their current contact. Similarly, respondents could deliberately try to avoid or try to make contact with the interviewer depending on their previous interactions with interviewers. Thus, Figure 14a contains useful extensions to the original theoretical model proposed by Groves, Cialdini, and Couper.

Ideally, one would like to have data to estimate *all* of the elements and *all* of the cells considered in Figure 14a. Ironically, although the BHPS data and auxiliary sources have suggested extensions to model proposed by Groves, Cialdini, and Couper, the BHPS data and auxiliary sources fail to provide data on numerous elements. For example, the impact of *Social Context* is limited to the data on geographic areas from the 1991 census and the impact of *Survey Design* is limited to interviewer continuity.¹²

12 The study of *Social Context* and *Survey Design* ideally requires the replication of the same survey under the same essential survey conditions, while systematically varying the factor (or

With respect to *Respondents*, the analysis focused on the socio-demographic and economic characteristics of respondents, although some behavioural and psychological factors were considered such as the number of organisations they belong to and the number of psychiatric symptoms they report.

As stated in Section 5.1.2, we only have information on *Interviewers'* age, gender, years of experience and grade level. Given the three part classification used in Chapter 2, we would ideally want indicators of their attitudes and expectations and as shown in the *Respondent-Interviewer Interaction* cell, data their actual behaviour on the "doorstep".

With respect to the *Respondent-Interviewer Interaction* cell and the *Past Respondent-Interviewer Interaction* cells, we only have information from the call-record data at Wave 1 and 2 and some indicators of respondent co-operation. Ideally we would require explicit doorstep data showing how both the respondent and interviewer behaved at both occasions.

These data limitations are summarised in Figure 14b. Although substantially reduced, this theoretical model still represents a useful framework from which to consider the general

factors) of interest. Implementing such an experiment to assess *Survey Design* factors can be accomplished fairly easily through dividing the survey into random portions with each receiving a different aspect of the factor under consideration. Measuring the impact of *Social Context*, however, is much more challenging as the aspects to be tested usually can't be modified. Thus one is left with the possibility of replicating the survey across various locations or time periods in which the factor of interest varies and then trying to disentangle confounding effects. For example, the difference in findings between any two time periods always contains period effects (general societal changes which affect people of all ages and in all cohorts), age effects (those influences associated with chronological ageing), and cohort effects (influences associated with being one of a group of individuals within a specified population who experienced a common specified event during a specified period of time) (see, for example, O'Malley, Bachman, and Johnston, 1984, and Glenn, 1981, among others).

Figure 14a: Extending the Theoretical Model of Groves, Cialdini, and Couper

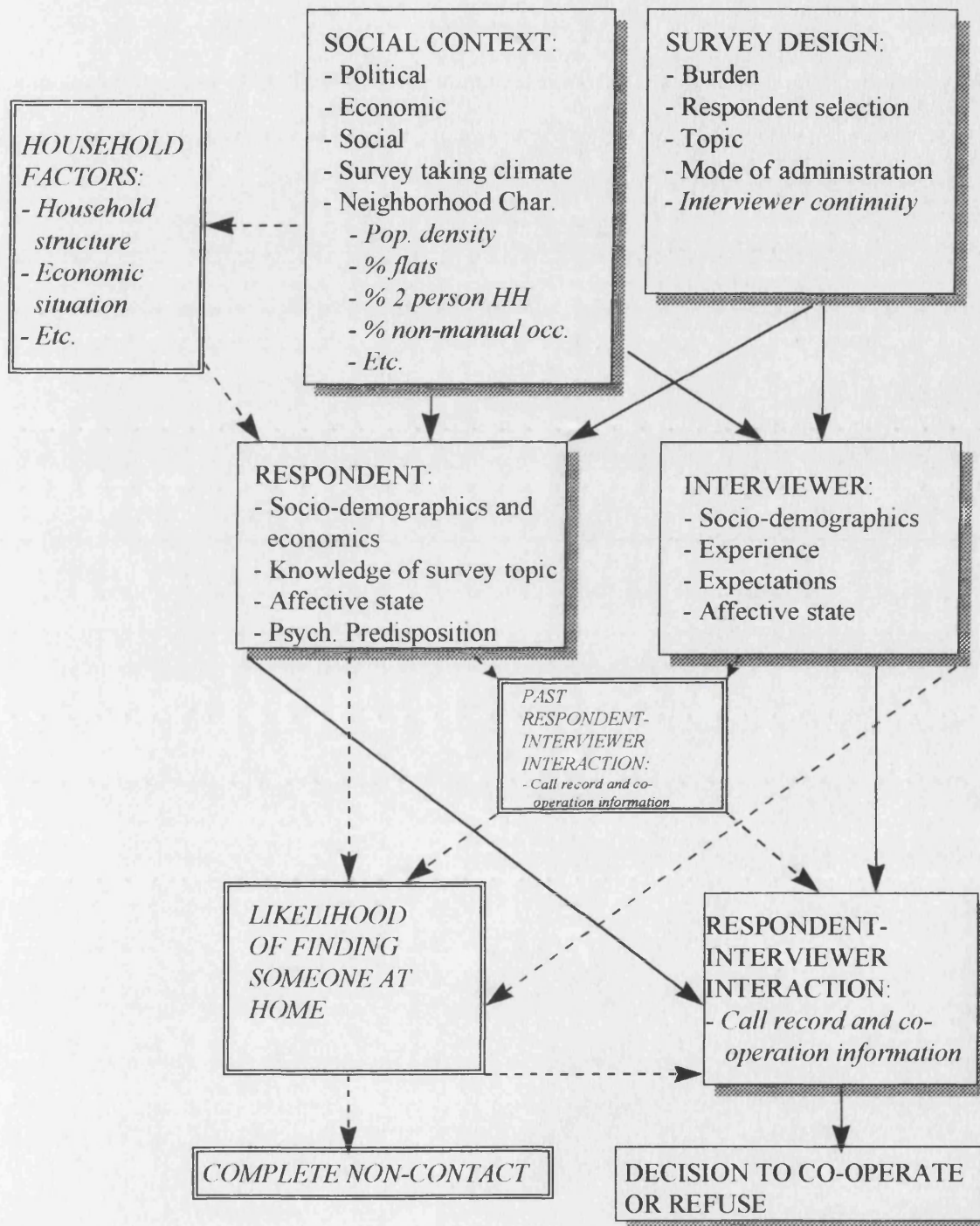
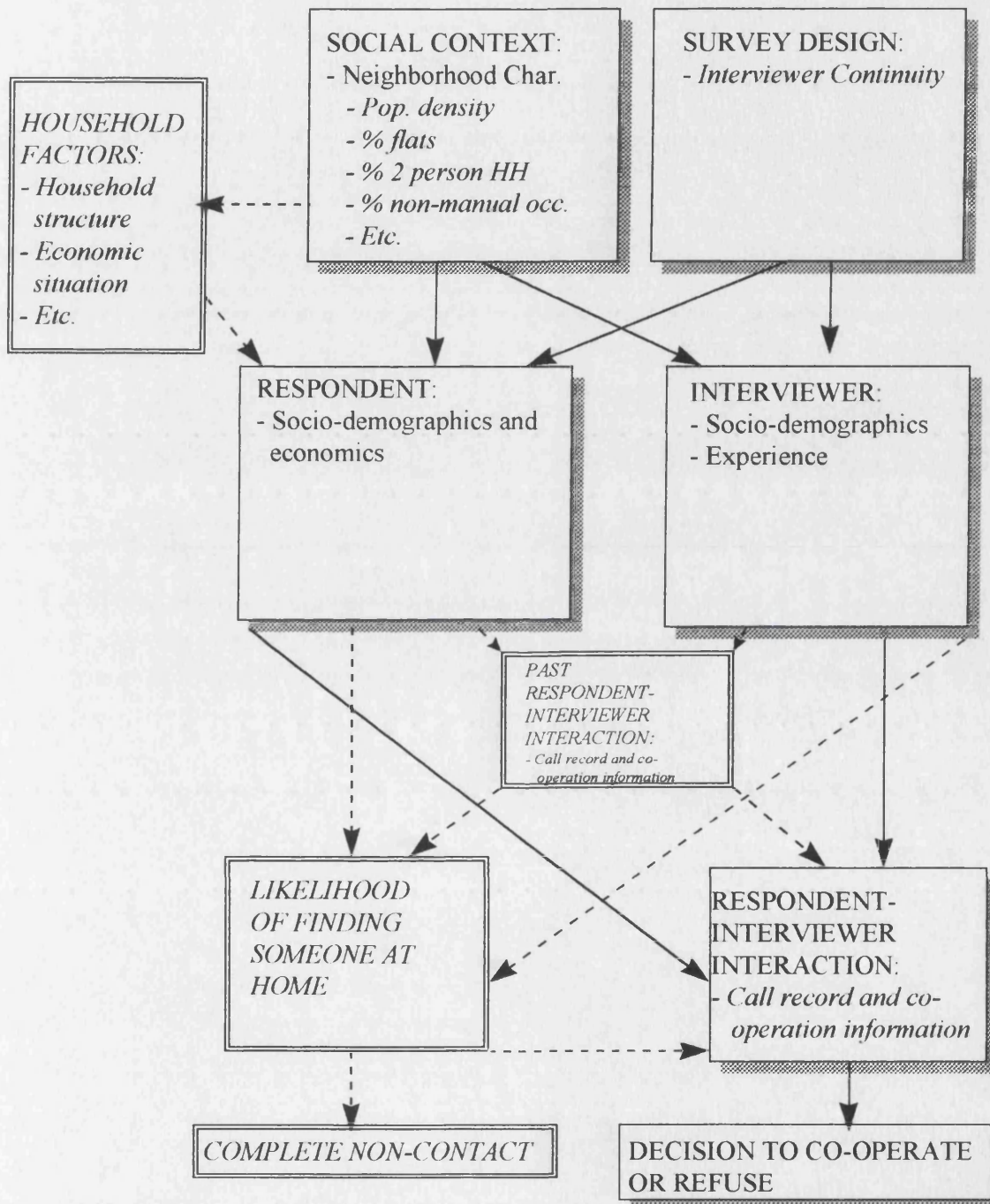


Figure 14b: Elements of the Theoretical Model for which Data are Available



empirical results in this Chapter and the results with respect to interviewer continuity in Chapter 8.

5.3 Methods

5.3.1 Indicators of Survey Nonresponse

At the second wave of a panel study, several types of attrition are possible. In addition to the main nonresponse categories of refusal, non-contact, and incapable of participating, there need to be categories for the untraced and the dead. It is also possible that although

Table 4: British Household Panel Study Response Rates for Waves 1 and 2

<u>Wave 1 Household Response</u>		<u>Wave 2 Household Response</u>	
Original sample addresses	8167	Issued households	5958
		(includes original and split-off households)	
- ineligibles	1033	Ineligible (e.g., whole household deceased, moved out of scope)	84
+ extra households in multi-household addresses	357		
Effective sample size	7491	Effective sample size	5874
All individuals interviewed	4862 (65%)	All individuals interviewed	4556 (78%)
All individuals either interviewed or proxied	5143 (69%)	All individuals either interviewed or proxied	4890 (83%)
Some individuals not interviewed/proxied	5538 (74%)	Some individuals not interviewed/proxied	5219 (89%)
Refusals	1543 (21%)	Refusals	440 (7%)
Non-contacts	288 (4%)	Non-contacts (includes not traced)	215 (4%)
Other non-interview (language problem/infirmity)	122 (2%)		

some individuals will be traced they may not be followed, e.g., because they have moved out of the defined study area, or for reasons of cost. Table 4 shows the BHPS's breakdown of nonresponse categories for Wave 1 and Wave 2.

As described in Section 3.4.2, several studies have noted that people who refuse differ from those who can not be found in terms of their demographic characteristics. We therefore decided to conduct separate analyses for overall nonresponse and for the refusal and non-contact sub-components. In addition, as we hypothesised that different factors would be important at the household and individual levels, we have decided to conduct separate analyses at each level. Thus there are 6 dependent variables. Nonresponse at the household and individual level, refusals at the household and individual level, and non-contacts at the household and individual level.

For the analyses, the household response rate was calculated following the advice of Groves (1989) (see Section 3.1.1). First, ineligible categories were excluded. These included instances where the whole household was deceased, where the whole household had moved to an institution or had moved out of Great Britain. These accounted for only 1.4 percent of the 1,493 households in the experimental sub-sample. The main categories of nonresponse in the sub-sample were refusals (7.1 percent), non-contacts (5.4 percent), and other non-interviews (0.3 percent). Thus the response rate for the second wave of the panel study was calculated as 87.2 percent.¹³ The household level nonresponse variables (household level nonresponse, complete household refusals, and complete household non-contacts) were calculated based on this definition. The individual level nonresponse variables classified as ineligible children under the age of 16 as well as those individuals

13 This is actually the rate for partial household coverage (some individuals not interviewed/proxied) and can be compared to the value of 89 percent for the full sample as

who had moved out of scope or to an institution or who had died. At the individual level refusals included individual refusals as well as individuals in complete refusal households and non-contacts included proxy interviews, those who were absent and no proxy was taken, individuals who had moved as well as those in complete non-contact households.

5.3.2 The Explanatory Variables

The explanatory variables are motivated by the theoretical model shown in Figure 14b. It should be noted that with respect to the household and individual level characteristics of potential respondents, known correlates of nonresponse from previous studies were considered as well as variables hypothesised to be related to nonresponse.

5.3.3 Overview of the Analysis Process

As a first step in the analysis process, bivariate relationships between each of the 6 dependent variables and each of the explanatory variables were examined with chi-square tests of independence (see Section 1.4.1). The next step in the analysis was to look at the impact of each of these characteristics while controlling for all of the other explanatory variables. This was handled through a series of multiple logistic regression models. Details of the specific method used are given in Sections 1.4.3 and 1.4.4. The final step was a series of cross-classified multilevel models where the true structure of the data could be accounted for (see Section 1.4.5).

5.3.4 Considerations with Respect to the Multiple Logistic Regression Models

Interactions. A consideration in any multiple regression model is the extent to which interaction terms should be added. It is interesting to note that the majority of the literature which has examined the correlates of nonresponse has only looked at main

effects. Ideally, theory should motivate the choice of interactions to be examined for inclusion in the model. Several such theory-based interactions were explored. The first was the interaction between respondent gender and respondent age with the prediction being that young men are harder to contact and elderly women are harder to persuade to participate. The second was a common concern among field staff of survey organisations that male interviewers might achieve poorer response rates, particular with female respondents, although there is very little research evidence to confirm this (see Section 2.2.1). The third had to do with the issue of interviewer continuity which will be discussed in Chapter 8. For example Rendtel (1990) found that having a different interviewer return in the second wave of a panel study had a much larger negative impact on older respondents. None of these interaction terms proved significant.

Aside from theory-based derivation, interactions can be found by applying some type of automated binary segmentation program such as CHAID (Chi-squared Automatic Interaction Detector) available through SPSS. A CHAID model was run for each of the six dependent variables including as explanatory variables all of the items listed in Appendix B and Footnote 15. CHAID produces a very complex interaction tree. None of the CHAID interactions were obvious ones that would have been predicted by theory. Therefore, these have not been added to the models which follow. (Note that the addition of CHAID interactions help to maximise the fit of the model to the data. If such interactions were included, ideally the resulting model should be tested with new data.)

Variables Included. The full list of explanatory variables which were included in the multiple logistic regression analyses included all of the variables which were significant bivariate predictors of nonresponse, plus some of the other variables known to be

predictors of nonresponse from other studies which were not significant in the bivariate analyses, e.g., annual household income, household size (see Footnote 15). These later variables were added because it is possible that when other variables are controlled for, the relationships between these explanatory variables and nonresponse will be revealed. The final additions to the model were the variables 'geographic pool' and 'same interviewer'. These variables were 'forced' to enter at step one of the process to act as general control variables. Geographic pool was chosen rather than region, because geographic pool will be used as a random effects term in the multilevel analyses which follow. In a standard logistic regression model, geographic pool and region would be considered linear combinations of each other and could not be entered simultaneously.

Missing Data. Another complication of any model which employs multiple explanatory variables is the differing amounts of missing data that can occur. One approach is to use pairwise deletion of missing data. However, this can lead to inconsistencies in the variance/covariance matrix. In addition, it is not available with the SPSS logistic regression procedure. In most cases, this concern was negated by actually adding 'missing' in as a category to be analysed.

Fortunately, for the vast majority of BHPS variables the level of missing data was very small (i.e., less than 1 percent). Missing data, however, was more prevalent (3-5%) on some of the interviewer observation items such as type of accommodation, respondent cooperation, respondent problems with seeing, etc. For any of the explanatory variables with more than 3 percent missing data, a dummy variable to indicate the missing data category was entered in addition to the dummy variables which reflected the substantive categories.

Unfortunately constructing a separate category to capture the missing data from the Wave 1 proxy cases (5 percent of the total individual sample) did not prove feasible. Proxy data were available for some but not all of the individual level Wave 1 predictors. Naturally, the category 'missing proxy data' was identical on several variables. SPSS will reduce the degrees of freedom of all but one of the variables involved when this happens. The model can still be estimated as long as only one of these variables is entered. Given that this was not always feasible, the 115 cases which only had proxy data for Wave 1 were excluded from the analyses. To be included in the models which follow, values would need to be imputed for the missing cases. (Although not shown, a separate analysis suggested that Wave 1 proxy cases were clearly predictive of Wave 2 individual non-contacts and thus overall individual nonresponse. They were not, however, related to household level nonresponse.)

Assumptions Behind the Model. In Section 1.4.4 the model assumptions with respect to logistic regression are discussed in detail. These points were kept in mind throughout the analysis. For example, to the extent possible all relevant variables have been included, but parsimony has also been taken into consideration. Also, no suspect variables were found with respect to the problem of zero cell counts and the problem of complete separation.

With respect to linearity in the logit, only the four continuous variables in the models (AIVNC – Total number of calls at Wave 1; PROTNCW1 – Proportion of calls that were non-contacts at Wave 1; BIVLNC – Number of calls at last known address for Wave 2; and TOTCNTW2 – Number of calls until first contact at Wave 2) needed to be checked. AIVNC and PROTNCW1 were only used in one model each and neither showed any evidence of non-linearity when the Box-Tidwell test was applied. TOTCNTW2 was used

in all six of the models, with non-linearity in the logit appearing for three of the six (whole household nonresponse, household level non-contacts and individual level refusals). BIVLNC was used in three of the six models and was shown to be highly non-linear in all three (whole household nonresponse, household level non-contacts, and individual level non-contacts). In test models, the two problematic variables were collapsed to four and six categories, respectively, and entered as a series of dummy variables. In its dummy variable form, TOTCNTW2 showed a significant monotonic relationship with nonresponse and BIVLNC showed a significant, essentially monotonic relationship with nonresponse. The explanatory power of the three models using the dummy version of BIVLNC was greatly improved by a factor of 0.10 in R_L^2 , whereas there was little difference in the model which only involved the dummy variable for TOTCNTW2. In the models with BIVLNC as a series of dummy variables, the coefficients for other variables that had p values just under .05 in Table 5a were no longer significant, whereas there was no change in the conclusions about the other variables for the model with the dummy version of TOTCNTW2. These findings suggest that there is little difference in using TOTCNTW2 as a continuous or series of dummy variables, but that BIVLNC is probably best handled as a series of dummy variables if the main concern is the explanatory power of the model. As discussed below, examination of the residuals paints a slightly different picture.

The models in Section 5.4.2 have been checked for both normality of the residuals and for influential cases. Results suggest that the distribution of the errors for all of the models deviate substantially from normality. In contrast, the search for outliers was more encouraging. For each model, there were only a few studentised residuals < -3 or $> +3$, a

few leverage values greater than $k+1/N$ (where k is the number of explanatory variables), and a few values of Cook's distance greater than 0.6 (an appropriate cut-off as suggested by a scatter plot of Cook's distance versus the predicted values). The best behaved models were those for whole household nonresponse and individual level nonresponse, these were followed by the models for whole household refusals, individual level refusals and individual level non-contacts. The model with the most problematic residuals based on all three indicators, was the one for whole household non-contacts. When a model was rerun excluding four highly problematic cases, the resulting model had slightly better explanatory power, but identical conclusions about the effects of the various covariates was reached. Given that none of the unusual values appear to be mistakes in the dataset and that they don't seem to alter conclusions, these values have been left in all models.

One concern about the variable BIVLNC (discussed above) is that its non-linearity may be distorting residuals. Residuals from models with the dummy version of BIVLNC were compared to those with the continuous version. For whole household nonresponse, the dummy variable model led to an increase in the number of big residuals, leverage values and values of Cook's distance. In contrast for household level non-contacts, the dummy variable model led to clear attenuation of the magnitude of the indicators for the problematic cases. In the third case of individual level non-contacts, some cases were no longer problematic, but new problematic ones took their place. These findings make the advantage of BIVLNC in its dummy variable version less clear.¹⁴

14 The models shown in Section 5.4.2 which form the basis of all later nonresponse work, were constructed at an early point in this thesis. I have only recently learned about the Box-Tidwell test. Given these latter findings with respect to the residuals and the focus on individual covariates rather than overall fit, I have decided to leave the models in Section 5.4.2 as they are. Had I originally known about the Box-Tidwell test, I would have looked more fully into the use of BIVLNC as a series of dummy variables.

5.4 Results

5.4.1 Bivariate Analyses

Appendix B shows the bivariate relationship between each of the explanatory variables considered and the various nonresponse indicators, at both the household and individual levels. Note that when individual level variables are analysed at the household level, they refer to the characteristics of the 'household reference person', the person in whose name the property is owned or rented. Variables which proved to be non-significant or essentially redundant with other variables have been excluded from Appendix B.¹⁵ Each grouping will be discussed in turn.¹⁶

Interviewer Level Variables. All of the interviewer characteristics studied (i.e., interviewer gender, age, years of experience, and grade level) were bivariately related to at least one of the six dependent nonresponse variables. Area managers and supervisors (INTTYPE) were seen to have higher nonresponse figures at both the household and individual level. Clearly this is in contradiction to past research (see Section 2.2.3). The exception can come if area managers and supervisors are given the difficult task of refusal conversion. In the current case, these higher nonresponse rates remained even after controlling for the refusal conversion cases. A partial explanation for the results ended up coming from the structure of the data themselves. For example, only three supervisors

15 For example, the following variables were examined but were not found to have a significant relationship with any of the dependent variables: annual household income, net monthly housing costs, number of sources of income, possession of microwave, number of employed adults in the household, individual industry code, number of single parents in the household, number of persons in household, number of married people in the household, and individual health status. We also examined a multi-category tenure variable which had additional levels beyond the owner/renter distinction, but these extra categories were not analytically useful. Several other variables were dropped as they were close correlates of ones included.

16 Usually in instances where a large number of tests are conducted, some type of multiple comparison procedure is recommended. Another approach is to try and replicate the findings with new data. What is reassuring for this analysis is the considerable resemblance of the

received full PSU interviewing assignments. These happened to be in inner London, outer London, and the West Midlands conurbation. Looking just within these three areas, the supervisors had far better response rates than the ordinary interviewers. But because these are difficult areas, the supervisors have higher nonresponse rates than ordinary interviewers for the country as a whole, a classic example of Simpson's paradox. The situation is less clear in the case of area managers who had PSU assignments, some of whom appear to have higher response rates than the regular interviewer within the same areas and some of whom appear to have lower response rates. It is unclear whether there is something unique about the area they were assigned or something unique about these particular individuals involved in this sample of interviewers.

Bivariate analyses also suggested that years of experience (YRSWNOP2), ranging from 0 to 20 years, was related to nonresponse, with those interviewers with 6-7 or 11-13 years of experience having the highest nonresponse rates. As the 'years of experience' variable was not simply a proxy for being an area manager or supervisor, it is again a bit unclear how to interpret the findings.

Interviewer age (AGE2) and gender (GENDER) were related to overall nonresponse at the individual level, but only significantly for one of the six dependent variables. The finding with respect to age is not monotonic. With interviewers ranging from age 26 to 72, it was the interviewers aged 26 to 38 and aged 51-60 who had the higher individual level nonresponse rates. This is partially in keeping with Lievesley (1986) who found that middle-aged interviewers had better response rates than the younger and older ones. But it should be remembered that past research does not paint a consistent picture with respect to interviewer age and nonresponse (see Section 2.2.1) The finding for gender is

in line with typical field organisation beliefs that male interviewers may have more difficulty in gaining co-operation. As described in Section 5.3.3, however, the possible two-way interaction between respondent gender and interviewer gender was not significant. This was true when modelling interviewer gender, respondent gender, and their interaction alone or after controlling for the other explanatory variables. Only the interaction for the individual level non-contact model approached significance ($p = .09$). In the case of individual level non-contacts, female interviewers did better than male interviewers. There was a slight tendency for female interviewers to do better at contacting females within a household than males, but the gender of the respondent made no difference for male interviewers.

An indicator on interviewer continuity (SAMEINT) is included in Appendix B, but is not discussed until Chapter 8.

Indicators of Respondent Co-operation and Contactability. Several of these indicators were created from the BHPS keyed call record files. First examined are those indicating difficulty in gaining co-operation at Wave 1 among households who nonetheless co-operated. It can be seen that the higher the number of calls that the interviewer had to make at Wave 1 (AIVNC), the more likely that individuals at Wave 2 are to remain uncontacted or to refuse at Wave 2 and at the whole household level, to remain uncontacted. A similar pattern with respect to non-contacts was found with the variable PROTNCW1 (the proportion of calls at Wave 1 that were non-contacts), the higher the values of PROTNCW1, the higher the likelihood that Wave 2 will be a complete household non-contact or certain individuals within the household will be missed.

Also examined was the number of visits in which contact was made (PROCONT). The main purpose of this variable was to look at how elusive people are after the initial contact. The variable was divided by the number of persons the interviewer needed to interview. This controls for the effects of household size, a necessary step as all persons in the household have to be interviewed and a high number of contact calls can result from this. The results suggested that the lower the proportion of contact visits that occurred at Wave 1, the higher the likelihood of an individual or whole household refusal at Wave 2. These results are contrary to hypothesis and suggest that this variable may be tapping into aspects other than the ones intended as there are many processes occurring on the doorstep. As it was not significant in the multiple logistic regression analyses which follow it was dropped and not included in Appendix B.

The last Wave 1 variable examined was whether or not someone broke an appointment with the interviewer at Wave 1. Although none of the results for this variable are significant, they are in the direction suggesting that individuals in such households are more likely to refuse or not be found at Wave 2.

The first Wave 2 call level variable considered is the total number of calls at last known address (BIVLNC). As shown in Appendix B, the more calls that an interviewer makes at Wave 2, the less likely that the household will be a non-contact. Also, the larger the number of calls, the more likely that a given individual will refuse (probably because cooperative people can usually be handled in a small number of calls). The next variable considered is the number of calls until first contact at Wave 2 (TOTCNTW2). Here as the number of calls rises and the interviewer doesn't make contact, the more likely that

the whole household, or individuals within that household, will be a non-contact. There is a similar pattern with whole household refusals.

There is an inherent difficulty in looking at the number of broken appointments at Wave 2, because an appointment can only be broken if an appointment was made and an appointment can only be made if a contact occurs. If a contact occurs, then by definition the case can not be a whole household non-contact. Putting aside the issue of contact for a moment, however, it can be seen that the level of whole household and individual refusals go down monotonically as the interviewer proceeds from making contact to making an appointment to having an unbroken appointment (not shown in Appendix B). This is the hypothesised 'foot-in-the-door' phenomenon (see Groves and Magilavy, 1981) (i.e., having an individual agree to a small task so that he/she will be more likely to agree to a larger one).

Other indicators of reluctance and contactability were the presence of a refusal within an otherwise co-operating household at Wave 1 (AIVFHO3), the response rate for the area at Wave 1 (RATENR)¹⁷ and whether or not a face-to-face refusal conversion was attempted at Wave 2 by a different interviewer (REFCON)¹⁸. The first of these proved to be a highly significant predictor of all types of nonresponse at Wave 2. The previous response rate was only weakly predictive of household nonresponse and the refusal conversion indicator was a good predictor of total nonresponse and more specifically refusals at both the household and individual levels at Wave 2.

17 The hypothesis was that a low response rate in the first wave would indicate a difficult geographic area. This hypothesis was confirmed. However, it is also possible in principle that a low first wave response rate could lead to high second wave response rate if all of the difficult households are lost at the first wave and only the more co-operative ones remain.

18 Some telephone conversions of nonrespondents were also tried and then the original interviewers were resent to the household. Unfortunately there is no indicator on the file for these telephone conversions.

Also considered were indicators of respondent co-operation/difficulty as observed by the interviewer at Wave 1. These included an overall rating of respondent co-operation (AIV42), an indicator of whether or not respondents agreed to answer the ‘tracking questions’ to help locate them next year (AIV52B), and several indicators of respondent problems affecting the interview such as eyesight (AIV6A2), hearing (AIV6B2), reading ability (AIV6C2), and English speaking (AIV6D2). With the exception of reading ability, each of these was generally a good indicator of refusals and overall nonresponse at both the household and individual levels at Wave 2. In addition, the interviewer’s rating of respondent co-operation at Wave 1 and whether there were interview difficulties due to language were also good predictors of household and individual non-contact at Wave 2.

Household and Individual Level Variables. As can be seen in Appendix B the household and individual level variables are similar to the standard correlates of nonresponse cited in other studies discussed in Section 3.4.2. For example, Appendix B shows a number of economic and consumer variables from Wave 1 which were related to overall nonresponse and contactability at both the household and individual levels. For example, these include people in ‘other accommodation’ (e.g., blocks of flats) (AHSTYPE2), those who don’t have a phone (APHONE2), those who are renters (AHSOWN3), and those who are on income support (AF1322). Several are also related to Wave 2 refusals as well as non-contacts: those who only have 1 or 2 rooms as part of their accommodation (AHSROOM4), those having no interest from dividends last year (AFIYRDI2), and not having a freezer (ACD3USE), washing machine (ACD4USE), or home computer (ACD8USE). Several are also mainly related to refusals at Wave 2: having net monthly housing costs between £1-125 (AXPHSN3), a total income last

month of £0-333 (AFIMN2), and not having a dish washer (ACD6USE) or CD player (ACD9USE). In addition, having no access to a car (ANCARS2) was related to overall nonresponse at both the household and individual level, having no VCR in the accommodation (ACD2USE) was related to overall household nonresponse, problems with damp (AHSPRBB2) was related to individual level non-contacts and having no tumble dryer (ACD5USE) was related to whole household non-contacts. Generally these all indicate that respondents who are less well-off economically are harder to include in the survey. As described in Footnote 15, an overall measure of annual household income was also considered. Although it did not prove to be significant, it showed a similar pattern.

Related to these direct economic indicators are the circumstances that bring about low income. For example, at both the individual and household level (AHGEST2 and ANUE2, respectively), the unemployed were harder to contact. Individuals within a household with 1 or more unemployed people were also more likely to refuse. Those with low educational qualifications (i.e., highest qualifications A level or below; AQFEDHI3) were more likely to refuse personally and at the whole household level. Similarly, higher rates of individual refusals were found among those in craft and related occupations, plant and machine operatives, sales occupations, and the general category of 'other' occupations (in the UK Standard Occupational Classification, SOC – see OPCS, 1990) as well as those who were not working (AJBSOC2). The highest rates of non-contact were found among those in associate professional and technical occupations and in craft and related occupations. Skilled, semi-skilled, and unskilled manual workers and unsurprisingly, small proprietors were also harder to contact (AJBGOLD2).

In addition to their economic circumstances, ownership of cars (ANCARS) and phones (APHONE2) can also be seen to influence a person's connectiveness with the outside world or lack thereof. It is interesting to note that phone ownership was found to be an important predictor of response/nonresponse with respect to individuals' completion of a self-administered supplement in a UK face-to-face survey (see, Farrant and O'Muircheartaigh, 1991). Supporting the connectiveness hypothesis, the number of organisations to which a respondent belonged (ANORGM2) was also related to individual level nonresponse and non-contact and overall household nonresponse. This is similar to findings by Gray and her colleagues (1996). Individuals with no religion or those who were not Church of England were also harder to contact (AOPRLG13). In terms of overall nonresponse, individuals with higher scores on a GHQ scale (i.e., expressing more psychiatric symptoms) were also more likely to be nonrespondents.

Type of accommodation (AHSTYPE2) and car ownership (ANCARS) can also be related to the urbancity of the area, with more flats in inner city areas and less need for car ownership. Higher proportions of ethnic groups can also be found in inner city areas. Generally, the non-white were harder to contact (ARACE3) and show a similar pattern to those who have problems with English (described above as AIV6D2) except that having problems with English is also related to individual refusals.

In terms of type of household, households with no children present (AHHTYPE3 and ANKID2) were more likely to refuse and thus be nonrespondents at both the household and individual level. Also households with a large number of persons of working age, including unrelated adults (ANWAGE2), were related to all six dependent variables with the exception of whole household refusals. Similarly, the never married (AMASTAT2)

were harder to contact. Curiously, having a couple in the household (ANCOUPL2) was related to a higher rate of individual level refusals. This is contrary to what would be expected (see Section 3.4.2), unless the key aspect for co-operation of a couple is the presence of dependent children. Although one person households had higher nonresponse rates than larger households as shown in other research, the 4 percent difference found in this study was not statistically significant (see Footnote 15).

In terms of other characteristics of the occupants, Appendix B shows the well known pattern that the young (AAGE2) are much harder to contact but generally co-operative, whereas the elderly are easy to contact, but generally less co-operative. A similar pattern was found with the household level variable: number of pensioners (ANPENS). Also men were harder to contact than women (ASEX).

Geographic Level Variables. The geographic level variables taken from the 1991 Census small area statistics file show a similar pattern to the individual and household level variables. Several characteristics predict higher non-contact rates and thus higher overall nonresponse at both the individual and whole household level. These will be discussed in terms of the 5 dimensions used for household and individual characteristics: economics, characteristics of the household related to economics, urbanicity, household structure, and other characteristics of the occupants. For example non-contacts were more likely to occur for the following economically related variables: areas which include a higher percentage of accommodation with less than 5 rooms (PERRM12; PERRM34) and a lower percentage of accommodation with 5+ rooms (PERRM5P), areas with a higher percentage of no car ownership (PER0CAR) and a lower percentage of car ownership (PER1CAR; PER2CAR), and areas with a higher percentage of local authority (and in

some cases, other) renters (PERCLA; PERCORNT) and a lower percentage of owner-occupiers (PERCOO). Areas with characteristics of households related to low socioeconomic status were also related to non-contacts. These included areas with a higher percentage of manual social grades (PERCC2DE) and a lower percentage of the professional social grades (PERCCL1) and areas with a higher percentage of unemployed (PERCUNEM). Areas with a lower percentage of higher qualifications (PERCEDHQ) were related to overall nonresponse at the individual level. Along the urbanicity dimension, non-contacts were most likely to occur in areas with higher population density (POPDENS), areas with a higher percentage of flats (PERCFLAT) and a lower percentage of detached houses (PERCDET), areas with a higher percentage in other service industries (PERO_SER) and a lower percentage in agriculture (PERAGRI) or energy and water supplies (PERENWAT) or mining (PERMINE), areas with a higher percentage of non-white (PERCNWT), and areas with a higher percentage of those who migrate (PERCMIG). Along the household structure lines, non-contacts were more likely to be found in areas with a higher percentage of one person households (PERC1P) and generally a lower percentage of households with two or more members (PERC2P; PERC3P; PERC4P), and areas with a higher percentage of households with no families (PERCF0) or a lone parent with dependent children (PERCFLDC) and a lower percentage of households with couples with children (PERCODC) or other types of family structure (PERCOFAM). In terms of characteristics of the occupants, non-contacts were more likely to be found in areas with a higher percentage of those with long term illness (PERCILL), and areas with a lower percentage in the 35-54 age group (PERC35P; PERC45P) and a higher percentage in the 15-24 and 25-34 ages groups (PERC15P; PERC25P).

In addition to increased difficulty in contacting households and individuals, a few of the geographic variables also suggest an increased level of individual refusals. These include areas with a higher percentage of renters (PERCLA; PERCORNT), areas with a higher percentage of lone parents with dependent children (PERCFLDC), areas with a higher percentage of non-white residents (PERCNWT), areas with a higher percentage of persons aged 55-64 (PERC55P), areas with a higher percentage of migrants (PERCMIG), and areas with a higher percentage of industries focusing on distribution, hotel and catering (PERCDISC) and other services (PERO_SER).

5.4.2 Multiple Logistic Regression Analyses

Discussion of Table 5a. The results of the multiple logistic regression modelling process are shown in Table 5a. In reviewing the results, it should be kept in mind, that the models were selected based on an automated selection procedure (see caveats in Section 1.4.3). Some of the significant relationships found with bivariate analysis remained significant when other factors were controlled for in the multiple regression analysis, others did not. Given the inter-relationships between the various explanatory variables it is possible that other equally plausible models exist that would suggest a slightly different subset of the bivariate relationships. Also sample size can play a role. With a larger sample, more of the bivariate results may have been supported. Thus, it is perhaps best to generalise from these results in terms of thinking of classes of variables rather than individual ones. For example as shown below, individual and household level economic variables are clearly linked to nonresponse. This is the key finding. What the particular economic measures are, is less important.¹⁹

19 Although for the vast majority of variables we had preconceived hypotheses based on prior research about which direction the relationships with nonresponse would take in the bivariate case, significance of the individual explanatory variables was assessed using two-tailed tests, a more conservative approach.

The values of R_L^2 (see Section 1.4.3) suggest a modest contribution of the explanatory variables to the explanation of nonresponse, ranging from 0.17 to 0.40, and with the best explanation being of whole household non-contacts. The values of Pseudo R^2 , which can not obtain the value of 1.0, are considerably smaller.

The discussion of Table 5a focuses on non-contacts and refusals rather than overall nonresponse, as overall nonresponse is almost entirely made up of these two main components. As can be seen, the interviewer, at least in terms of the characteristics that are measured in this study, had very little impact on nonresponse. Of the 30 possible combinations of interviewer characteristics and dependent variables, only two are present in the final models. For example, while controlling for the individual, household, and area characteristics of the sample, it is the younger interviewers and those aged 51-60 who have the highest refusal rates within households (AGE2). This non-monotonic finding is similar to that found in the bivariate analyses for individual level nonresponse. Years of experience (YRSWNOP2) also shows up in the multiple regression models. This time it is those interviewers with the least experience at NOP who have the best rates of contact within the household, contrary to expectation. Given that (1) two significant findings among 30 tests for interviewer characteristics are what one would expect to find by chance alone, (2) the fact that there is no strong evidence in the literature for the influence of interviewer background characteristics on response rates, and (3) the fact that the finding for years of experience is contrary to expectation, one could conclude that the interviewer does not appear to play much of a role in household and individual level nonresponse. The overall impact of the interviewer will be re-visited at the multilevel stage where interviewer effects are entered as a random effects term. If significant

Table 5b: Predicted Probabilities of Nonresponse Depending on Respondent Co-operation and Contactability, Characteristics of the Respondent, and Location

		Co-operation and Contactability			
		Good		Poor	
Household Level Nonresponse					
Demographic and Economic Profile	Easy to survey	London	Scotland	London	Scotland
	Difficult to survey				
Household Level Refusals					
Demographic and Economic Profile	Easy to survey	London	Scotland	London	Scotland
	Difficult to survey				
Household Level Non-contacts					
Demographic and Economic Profile	Easy to survey	London	Scotland	London	Scotland
	Difficult to survey				
Individual Level Nonresponse					
Demographic and Economic Profile	Easy to survey	London	Scotland	London	Scotland
	Difficult to survey				
Individual Level Refusals					
Demographic and Economic Profile	Easy to survey	London	Scotland	London	Scotland
	Difficult to survey				
Individual Level Non-contacts					
Demographic and Economic Profile	Easy to survey	London	Scotland	London	Scotland
	Difficult to survey				

⊕ Background details about how these predicted probabilities were calculated are given in Appendix C.

random effects are found, this would suggest that interviewers are indeed an important factor in the nonresponse process and this influence is not due to their gender, age, experience, or grade level/status or maintaining the same interviewer over time in a panel study.

The most productive explanatory variables in Table 5a are those from the grouping labelled 'Indicators of Respondent Co-operation and Contactability'. As discussed in Section 5.3.2, these include variables which were developed from Wave 1 and 2 call record data and outcomes as well as interviewer observations. These describe individuals and households who are potentially difficult to gain co-operation from even though they did participate in the first Wave. As can be seen, the total number of calls at Wave 1 (AIVNC) and the proportion of calls that were non-contacts at Wave 1 (PROTNCW1) were good indicators of Wave 2 non-contacts in these multiple regression models. Similarly, the more calls an interviewer makes at Wave 2 (BIVLNC), the more likely contact will be achieved. In contrast, it can be seen that the higher the number of calls that are made until first contact at Wave 2 (TOTCNTW2), the more likely that the final outcome will be an individual or whole household non-contact. This is consistent with the findings of Purdon, Campanelli, and Sturgis (1997) and suggests that if an address resident has failed to be in on previous occasions, then s/he is also likely to be out on subsequent occasions. Interestingly, this variable is also a good predictor of whole household and individual refusals at Wave 2.

The co-operation of the respondent at Wave 1 (AIV42), as rated by the interviewer, is another key variable which was a consistently good indicator of both contact and co-operation difficulties at Wave 2.

Whether or not there was an individual refusal in an otherwise co-operating household (AIVFHO3) was a robust predictor of continuing individual refusals and importantly, whole household refusals at Wave 2. Whether or not the occupants had problems speaking English (AIV6D2) at Wave 1 was a good indicator of whole household non-contact at Wave 2.

Several of the household and individual characteristics of respondents remained as good predictors of nonresponse in the multiple regression framework. In contrast, only a small handful of the geographic area variables remained significant in the multiple logistic regression context, probably because the household and individual characteristics are much better predictors. But it also should be noted that there could be other important aspects of areas which are not captured by the Census variables (e.g., the presence of security devices). As was the case for interviewer characteristics, this may suggest that the relationship between area and nonresponse is due to factors other than those measured by the Census variables. These other factors could include the household and individual characteristics themselves, but also other aspects of housing such as the number of floors in the buildings, entry phones, visible security devices, etc. Given the conceptual similarity of the individual, household, and area variables, the results for these will be discussed together.

Non-contacts were most likely amongst those who did not have a phone (APHONE2), or those who were renters (AHSOWN3), or male (ASEX), or young (AAGE2), or non-white (ARACE3), or those who lived in areas characterised by low car access (PEROCAR) and with a lower proportion of couples with dependent children

(PERCODC). Surprisingly, once these other variables are controlled for, it is an area with a lower percentage rather than higher percentage of flats which led to whole household non-contacts (PERCFLAT). These three area findings with respect to non-contacts were not sustained in the comparable multilevel models which follow.

Refusals are best described by those households which didn't have a home computer (ACD8USE), or those who had lower monthly housing costs (AXPHSN3), or those with a couple present (ANCOUPL2), or those who had at least one pensioner in the household (ANPENS2), as well as those who lived in an area with a higher proportion of distribution, hotel and catering industries (PERDISC). In the multiple regression case, it is those who had a medium amount of income from dividends rather than those with little or no dividends (AFIYRDI2), those who are working rather than the unemployed (AHGEST2) and those who lived in an area with a lower rather than higher percentage of non-white residents (PERCNWT), who were more likely to refuse. This unusual finding for non-white residents is not sustained in the comparable multilevel models which follow.

Several variables were only significantly related to overall nonresponse, rather than its components. These suggest higher nonresponse for those individuals in flats (AHSTYPE2), or those with only 1-2 rooms in the accommodation (AHSROOM4), or those with fewer than two cars (ANCARS2), or those made up of four or more adults of working age (ANWAGE2), or those in manual occupations as well as small proprietors (AJBGOLD2).

Thus the multiple logistic regression models show a very similar picture to the bivariate ones in that nonrespondents tend to be those who are less well off economically or who

possess characteristics which could lead to economic hardship. There are, however, some exceptions to this pattern. For example, those who are working should have more economic security than the unemployed, but are more likely to refuse (probably because the work keeps them very busy). Similarly, those who have a *medium* amount of income from dividends should have more economic security than those who have little or none. This is a curious finding as common survey wisdom often views the relationship between income and nonresponse to be curvilinear, with those who are the least and most well-off being the most difficult to include in a survey.

In the multiple logistic regression context, support for the social ‘connectiveness’ and urbanicity themes are less clear, but it is still the case that particular kinds of households and individuals are going to be more problematic with young people, men, and the non-whites being more difficult to contact and the elderly, and the working, being more likely to refuse. There are also the difficulties with proprietors of small businesses and households with large numbers of adults of working age which include households of unrelated sharers. The curious finding that households with couples are more likely to refuse also persists from the bivariate analyses.

Discussion of Table 5b. A table of predicted probabilities based on the coefficients in Table 5a has been included to make Table 5a more accessible. From the discussion about Table 5a it is clear that the best explanatory variables are the indicators of co-operation and contactability and the various characteristics of households and individuals. In selecting profiles to report, the decision was made to compare and contrast the impact of the co-operation and contactability factors with the respondent characteristics factors while holding constant the fixed effects of interviewers and areas. The other

consideration was the variability in response rates known to occur by geographic pool and captured in the models in Table 5 with 29 dummy variables. For the purpose of illustration, predicted probabilities are shown for a geographic pool with a good response rate and one with a poor response rate (i.e., a geographic pool from Scotland and one from Inner/Outer London, respectively). Details of the specific values chosen for each profile are included in Appendix C. An examination of Table 5b shows strong effects of the co-operation and contactability factors and the respondent characteristic factors. Generally, respondents with good co-operation and contactability histories and characteristics showing an easy to survey profile have extremely low predicted probabilities of nonresponse (i.e. $< .03$). Conversely, respondents with poor co-operation and contactability histories and characteristics showing a difficult to survey profile have extremely high predicted probabilities of nonresponse (i.e., $< .86$). Although the demographic and economic characteristics of respondents are clearly important, the co-operation and contactability factor appears to dominate. For example, those with good co-operation and contactability histories generally have low predicted probabilities of nonresponse ($< .50$) even if they have a difficult to survey profile (the exceptions being household level nonresponse in London, and individual level nonresponse in London and Scotland) and those with poor co-operation and contactability histories generally have high predicted probabilities of nonresponse ($> .50$) even if they have an easy to survey profile (the exceptions being household level refusals in Scotland, individual refusals in London and Scotland and individual non-contacts in London and Scotland). Location also makes a difference, but this is less pronounced. For example, note the small variation by location across the 'good co-operation and contactability' by 'easy to survey' cells and across the 'poor co-operation and contactability' by 'difficult to survey' cells. There are, however, some marked differences in the other cells. For example, having a good co-

operation and contactability history and a difficult to survey profile, you are much more likely to end up as a household level non-contact in London than in Scotland (0.40 versus 0.09, respectively).

5.4.3 Cross-Classified Multilevel Models

The next step was to rerun the final models shown in Table 5a under the multilevel framework, including the random effects terms for households/individuals, interviewers, PSUs, and the larger geographic pools. These are shown in Table 6. (See Section 1.4.5 - 1.4.9 for a fuller description of the rationale behind multilevel modelling and Section 6.4 and 7.4 for some specific technical concerns.)

Comparing Table 5a to Table 6, it can be seen that exactly the same conclusions about the pattern of substantive results is found in these multilevel models as was found for the standard logistic regression models, with the exception that the few area and interviewer characteristic categories which were significant in the standard multiple logistic regression are generally no longer significant.

The random effects terms for interviewers, PSUs, and geographic pools will be discussed in Chapter 6.

5.5 Chapter Summary and Discussion

5.5.1 Summary of the Findings about the Correlates of Nonresponse

This Chapter explored the factors which are predictive of overall nonresponse, refusals, and non-contacts at both the household and individual levels at Wave 2 of the BHPS.

Table 6: Cross-Classified Multilevel Logistic Regression Models: Fixed and Random Effects θ , ϕ

Household Level						Individual Level					
Non-respondents		Refusals		Non-contacts		Non-respondents		Refusals		Non-contacts	
Variable	β (se)	Variable	β (se)	Variable	β (se)	Variable	β (se)	Variable	β (se)	Variable	β (se)
SAMEINT	-0.06(.20)	SAMEINT	-0.17(.25)	SAMEINT	-0.01(.31)	SAMEINT	-0.14(0.15)	SAMEINT	-0.13(0.19)	SAMEINT	-0.04(0.20)
								AGE2(1)	0.26(0.51)	YRSWNOP(2)	0.40(0.29)
								AGE2(2)	-0.13(0.41)	YRSWNOP(3)	0.49(0.33)
								AGE2(3)	0.52(0.41)	YRSWNOP(4)	0.74(0.39)
										YRSWNOP(5)	1.19(0.40)*
										YRSWNOP(6)	0.23(0.44)
BIVLNC	-0.22(.06)*			AIVNC	0.17(.06)*					BIVLNC	-0.12(0.05)*
TOTCNTW2	0.39(.07)*	TOTCNTW2	0.18(0.7)*	TOTCNTW2	0.69(.12)*	TOTCNTW2	0.21(0.04)*	TOTCNTW2	0.17(0.05)*	TOTCNTW2	0.29(0.06)*
AIVFHO3	1.46(.33)*	AIVFHO3	1.88(.34)*	BIVLNC	-0.60(.12)*	AIVFHO3	1.51(0.22)*	AIVFHO3	1.99(0.23)*	PROTNCW1	0.97(0.34)*
AIV42(1)	0.34(.54)	AIV42(1)	-0.49(1.05)	AIV42(1)	3.89(1.37)*	AIV42(1)	1.65(0.69)*	AIV42(1)	-0.72(0.73)	AIV42(1)	0.60(0.53)
AIV42(3)	1.04(.21)*	AIV42(3)	1.03(.26)*	AIV42(3)	0.70(.33)*	AIV42(3)	0.74(0.15)*	AIV42(3)	0.75(0.20)*	AIV42(3)	0.77(0.22)*
AIV42(4)	2.10(.59)*	AIV42(4)	1.63(.66)*	AIV42(4)	0.60(1.15)	AIV42(4)	2.38(0.61)*	AIV42(4)	2.00(0.59)*	AIV42(4)	2.29(0.74)*
				AIV6D2(1)	-3.74(1.62)*	AIV6D2(1)	-2.25(0.80)*				
				AIV6D2(2)	2.01(.68)*	AIV6D2(2)	0.69(0.39)				
AHSROOM4	0.71(.24)*			APHONED1	0.79(.32)*	AHSTYP(1)	0.36(0.39)			AHSONWD3	0.68(0.20)*
						AHSTYP(2)	-0.05(0.26)				
						AHSTYP(3)	-0.32(0.19)				
						AHSTYP(4)	-0.57(0.21)*				
						ANCARS(0)	0.57(0.23)*	ANCOUPL2	-0.61(0.21)*		
						ANCARS(1)	0.47(0.19)*				
						ACD8USE	0.49(0.18)*	ACD8USE	0.71(0.25)*		
						AXPHSN(1)	0.11(0.21)	AXPHSN(1)	0.65(0.31)*		
						AXPHSN(2)	0.61(0.20)*	AXPHSN(2)	1.40(0.29)*		
						AXPHSN(3)	0.24(0.19)	AXPHSN(3)	0.72(0.30)*		
						ANWAGE2	0.55(0.21)*				
		ANPENS2	1.14(.25)*			ANPENS2	0.74(0.26)*	ANPENS2	0.96(0.23)*		
AAGE2(1)	0.87(.26)*			AAGE2(1)	1.29(.33)*	AAGE2(1)	0.86(0.17)*			AAGE2(1)	1.23(0.20)*
AAGE2(3)	0.32(.27)			AAGE2(3)	-1.48(.60)*	AAGE2(3)	-0.04(0.29)			AAGE2(3)	-0.79(0.35)*
						ASEX	0.37(0.13)*			ASEX	0.65(0.19)*
AJBGLD2(1)	0.24(.27)					AHGEST(2)	0.09(0.23)	AHGEST(2)	0.07(0.32)	ARACE3	0.91(0.30)*
AJBGLD2(3)	0.85(.39)*					AHGEST(3)	-0.54(0.19)*	AHGEST(3)	-0.58(0.23)*		
AJBGLD2(4)	0.79(.28)*										
AFIYRDI2(1)	0.69(.27)*	AFIYRDI2(1)	0.32(.33)								
AFIYRDI2(3)	0.70(.30)*	AFIYRDI2(3)	0.79(.35)*								
AFIYRDI2(4)	-0.11(.49)	AFIYRDI2(4)	-0.34(.59)								
		PERDISC	0.06(0.3)	PERCFLAT	-0.01(.01)	PERDISC	0.03(0.02)	PERDISC	0.06(0.03)*		
				PERCOCAR	0.01(.01)	PERCNWT	0.00(0.01)	PERCNWT	-0.01(0.01)		
				PERCODC	-0.04(.04)						
GEO POOL	0.08(.15)		0		0.24(.28)		0		0		0
PSU	0		0		0		0.04(.07)		0		0.20(.13)
VIEWER	0.29(.20)		0.23(.19)		0.08(.31)		0.31(.11)*		0.40(.15)*		0
H'HOLD/INDIV'L	1		1		1		1		1		1
CONSTANT	-3.28	CONSTANT	-5.32	CONSTANT	-2.40	CONSTANT	-4.09	CONSTANT	-5.70	CONSTANT	-4.56

θ Random parameters are based on the logistic scale so that they can not be directly compared to Table 7. ϕ Note that when individual level variables are analysed at the household level, they refer to the characteristics of the 'household reference person', the person in whose name the property is owned or rented. * Significant at $p < .05$ as assessed by MLn Wald test ($p < .01$ and $p < .001$ not shown).

Key to variable names (*omitted category in italics*):

SAMEINT - Same interviewer as Wave 1: 0=*different*, 1=same
AGE2 - Interviewer age: 1=26-38 yrs, 2=39-50 yrs, 3=51-60 yrs, 4=*61-72 yrs*
YRSWNOP - Years working for NOP: 1=*0-2 yrs*, 2=3-5 yrs, 3=6-7 yrs, 4=8-10 yrs, 5=11-13 yrs, 6=14-20 yrs
BIVLNC - Number of calls at last known address (Wave 2)
AIVNC - Total number of calls (Wave 1)
TOTCNTW2 - Number of calls until first contact (Wave 2)
AIVFHO3 - Wave 1 final outcome: 0=no internal refusal, 1=internal HH refusal
PROTNCW1 - Proportion of calls that were non-contacts (Wave 1)
AIV42 - Co-operation of respondent: 1=missing, 2=*very good*, 3=good/fair, 4=poor/very poor
AIV6D2 - Problems affecting interview: English: 1=missing, 2=yes, 3=*no*
AHSROOM4 - Number of rooms in accommodation: 0=*3+*, 1=1-2
APHONED1 - HH has telephone: 1=no phone, 0=*phone*
AHSTYP - Type of accommodation: 1=missing, 2=det.hse/bungalow, 3=semi/bung/end of terrace, 4=terraced, 5=*other*
AHSOWND3 - House owned or rented: 0=*owned*, 1=rented
ANCARS - Car/van available for private use: 0=none, 1=one, 2=2+
ANCOUPL2 - A couple present in HH: 0=*no*, 1=yes
ACD8USE - Home computer in accommodation: 1=no, 0=*yes*
AXPHSN - Net monthly housing costs: 1=none, 2=1-125 pound, 3=126-250 pounds, 4=*251+*
ANWAGE2 - Number in HH of working age: 0=*0-3 people*, 1=4+ people
ANPENS2 - Number over pensionable age in HH: 0=*none*, 1=1+
AAGE2 - Respondent's age at date of interview: 1=15-25; 2=26-59;3=60-93
ASEX - Respondent's gender: 1=male, 0=*female*
AJBGLD2 - Goldthorpe social class: 1=NA, 2=*non-manual*, 3=small proprietors, 4>manual
AHGEST - Employment status: 1=*working*, 2=unemployed, 3=other
ARACE3 - Ethnic group membership: 0=*white*, 1=non-white
AFIYRDI2 - Income last year from dividends/interest: 1=nothing, 2=*under 100 pounds*, 3=100-1000 pounds, 4=1000+ pounds
PERDISC - % distribution, hotel, and catering industries in area
PERCFLAT - % flats in area
PERC0CAR - % no car households in area
PERCNWT - % non-white residents in area
PERCODC - % households of couples with dependent children

Households and Individuals. Bivariate analyses suggested that nonrespondents to Wave 2 of a panel study tend to be those who are less economically well-off or have other characteristics associated with low incomes. For example, it is harder to make contact with residents in flats, renters, those with only 1 or 2 rooms in their accommodation, those with no phone, the unemployed, those in receipt of income support, and those with no interest from dividends as well as those who live in areas where there was a higher proportion of flats, accommodation with less than 5 rooms, no cars, mainly local authority and other renters, a higher proportion of unemployed or a higher proportion of manual workers. Similarly, it is also harder to persuade those who only have 1 or 2 rooms in their accommodation, those with low monthly housing costs, those with low educational qualifications, those employed as plant and machine operatives or in one of the occupations in the 'other' category (based on the UK Standard Occupational Classification; SOC) and those who did not have more expenditure on consumer items such as separate freezers, washing machines, dish washers, home computers, and CD players, as well as those who live in areas with a higher proportion of local authority and other renters or with a higher proportion of distribution, hotel and catering industries. This economic influence was also clearly seen in the multiple logistic regression models.

The bivariate analyses also suggested that urbanicity plays a key role in nonresponse. In addition to the individual and areas characteristics pertaining to lower car ownership and a higher proportion of flats which have already been mentioned, other indicators were a higher population density, a higher rate of migration and being non-white or living in an area with a higher proportion of non-white residents. There was also some suggestion that those who were less socially integrated were harder to contact including those who

did not have phones, belonged to no organisations, and had no religion or were not Church of England. There was less support for the urbanicity and social connectiveness themes in the multiple logistic regression models.

Both the bivariate and multiple logistic regression analyses consistently picked out some of the other key characteristics of persons and households that are related to nonresponse. For example, these suggested that those who live in households with four or more adults of working age, males, and the old and the young, the employed, and proprietors of small businesses are more difficult to include in a survey. The bivariate analyses also suggested difficulty in areas with a higher proportion of 15-34 years olds, a higher proportion of persons with long-term illness, or a higher proportion of one person households, a higher proportion of households with no families or lone parents with dependent children.

Interviewers, Areas, and Measures of Co-operation and Contactability. Interviewer characteristics (such as age, gender, experience, and grade level) can predict nonresponse in bivariate analyses. Area characteristics (such as population density, proportion of flats in the area, percentage of non-white residents, etc. taken from 1991 Census small area statistics) are clear important bivariate predictors of nonresponse. Both of these sets of effects, however, virtually disappear when respondent and household characteristics are controlled for. If one were only to conduct such analyses, one would come to the conclusion that interviewer and area effects can be ignored. A different conclusion, however, is reached in Chapter 6, when the random components of the multilevel models are examined.

In contrast, particularly useful indicators of attrition nonresponse as identified by both bivariate analysis and multiple logistic regression are the various indicators of co-operation and contactability from the previous wave. These suggest that respondents who were difficult to obtain an interview from in the first wave of a survey, but who nonetheless participate, are much more likely not to participate at all in subsequent waves. The importance of these variables in contrast to the household and individual characteristics of the respondent were shown in Table 5b.

This also held true with respect to Wave 1 proxy cases. Those individual who were absent at Wave 1 and a questionnaire was filled by proxy were significantly more likely to be individual level non-contacts at Wave 2. As described under the 'Missing Data' heading in Section 5.3.3, these cases could not be analysed directly in the model without imputation. Even though a separate missing category was developed, the exact same cases were missing from the majority of Wave 1 individual level variables. On the other hand, missing was specified as a category for variables which had 3 percent or more missing cases. These occurred exclusively with respect to the interviewer observation variables, representing interviewers who failed to fill in their extra administrative tasks, and complete household non-contacts cases where these variables did not need to be completed by interviewers. Thus the finding that the missing category of AIV42 (the interviewer's rating of the respondent's co-operation) and AIV6D2 (the interviewer's rating that the respondent has problems with English) are related to household non-contacts and individual level nonresponse is not surprising.

5.5.2 Attrition Versus Initial Nonresponse

As nonresponse in this thesis is studied at Wave 2, this is technically attrition nonresponse

which is being examined. Ideally, of course, one would like to be able to generalise the findings of this Chapter to the characteristics of those individuals who failed to participate in the survey at Wave 1. Interestingly, several studies have suggested the similarity between the types of people who are lost in the first wave of a survey (initial nonresponse) and those who are lost at later waves (attrition nonresponse) (see, for example, Farrant and O'Muircheartaigh, 1991; Gray, *et al*; 1996). More importantly, Taylor (1994) who studied BHPS Wave 1 nonresponse in comparison to the 1991 Census sample of anonymised records found a very similar pattern to the results found in this thesis which used Wave 2 data. (It should be remembered, however, that this is just a rough comparison as Taylor's comparison to the Census confounded any unrepresentativeness of the selected sample with nonresponse bias and the interpenetrated sub-sample differs from the full sample by slightly over-representing urban dwellers (see Appendix A)). Despite these differences, both Taylor (1994) and this thesis found that the people who tended to be lost were the young and the old, those with no access to a car, those who were never married, those who were non-white, and semi-skilled and unskilled manual workers. Both studies also found that the married and cohabiting, and employers and managers of large establishments tended to be over-represented. Although Taylor (1994) found a difference with respect to household size, none was found in this study. There were also some instances where Taylor found differences in small categories which were not found in this study. In addition, Taylor found that private renters were over-represented whereas this study found them more likely to be nonrespondents. Overall, however, the results were more similar than different. Support for being able to generalise from the findings in this thesis at Wave 2 to initial nonresponse is also strengthened by the great similarity between the findings in this Chapter and other research based on initial nonresponse.

5.5.3 Interpreting Household and Individual Correlates from a Theoretical Perspective

As described in Section 3.4, the causes of nonresponse can be viewed from the sociological level in terms of changes in social structure and from the level of the individual in terms of attitudes and beliefs. The correlates of nonresponse discussed above, however, describe a socio-demographic and economic profile of households and individuals. This profile can be labelled the level of social group membership (i.e., being young or old, working or unemployed, etc.) and can be seen as fitting in between the level of social structure and the level of the individual person.

Given the similarity of the nonresponse profile developed in this thesis and the various nonresponse profiles developed over the years, there could be an argument that the classes of households and individuals described by the profile are *fixed*. Attitudes, beliefs and other personality factors are also typically considered to be *fixed* attributes of the individual. How then do social group membership and individual level factors interact at the level of individual respondent decisions?

There is obviously variability within a fixed social group category (e.g., not all elderly people refuse). Such variability could be due to sampling variability. But as argued in Section 3.3, one also needs to consider a *variable* response disposition within a person. Evidence for this latter model is clearly seen in the BHPS data. For example, the Wave 1 indicators of co-operation and contactability provide a profile of difficult to find and unco-operative respondents who nonetheless participated at Wave 1, but who were much less likely than the average respondent to participate at Wave 2. In terms of individual

level patterns, we also know that whether or not a respondent will agree to participate is based in part on 'situational' factors as well as fixed personality factors (see Morton-Williams, 1993 and Groves and Cialdini, 1991; see also Chapter 8 for a further discussion of situational factors).

One way to reconcile these various findings is to broaden the category of situational factors to include lifestyle factors as well. The social group membership factors can then be seen to directly influence the situational and lifestyle factors. In the case of non-contacts, these situational and lifestyle factors can directly translate into the likelihood of finding someone at home (e.g., young people are hard to contact, because the lifestyle young people lead makes them hard to contact). The case of refusals would be slightly different. One could view the situational and lifestyle factors as having the most impact on individuals who are located toward the centre on the reluctance dimension. Or one could see certain group membership attributes, such as those which profile refusers, as shifting these individuals toward the less co-operative end of the reluctance dimension. For example, the employed can be more reluctant because the pressures of work make them more reluctant. Dissecting what actually occurs at the nexus between the macro and micro influences on nonresponse is an interesting subject for further research, not only for social and psychological theory, but also from a statistical modelling perspective. As seen in Section 5.5.4, it is also interesting from a survey fieldwork standpoint.

5.5.4 Practical Implications of Nonresponse Correlates

Considering the fixed and variable aspects of survey nonresponse can be very useful to inform survey research practice. The profile of hard-to-include types of individuals and households could be considered as identifying *fixed* social groups which need to be

targeted for special response strategies. The good news is that response can be seen as *variable* within each particular profile category. Thus special targeted strategies aimed at people within these fixed classifications should be successful.

For a panel study, these analyses clearly suggest difficult groups which can be identified well in advance of Wave 2 or subsequent waves who could benefit from special targeted fieldwork strategies at Wave 2 or subsequent waves. It should be remembered that some of the best indicators are interviewers' subjective rating of the person's level of co-operation at the previous wave and variables constructed from the previous call-records and outcomes. Thus, these variables should routinely be collected and keyed as part of data collection.

The variable nature of nonresponse also suggests the usefulness of always revisiting past non-contacts and refusals at subsequent waves of a panel study as these households and individuals may now participate.

For a one-off survey, the analyses suggest the types of socio-demographic and economic variables that ideally should be sought for nonresponse weighting. This is true whether the nonresponse weighting proceeds through comparison to a more complete data source such as the Census, or through information on respondents and nonrespondents present on the sampling frame, or through the use of a special data collection form for the nonresponse cases which is typically completed via observation by the interviewer, but can consist of a set of key questions to ask nonrespondents (see, for example, Lynn, 1996).

Section 3.2 had mentioned the view that the bias due to refusals and the bias due to non-contacts could in many instances be compensating as suggested by the work of Wilcox (1977) and Lievesley (1986). With respect to economic conditions, the analyses in this Chapter suggest that those at the lower end of the economic scale are both more difficult to persuade to participate and more difficult to find. This suggests the idea of compounding rather than compensating biases. Although there are several differences between the characteristics of those who become refusals and those who become non-contacts, the sole example of compensating bias is the fact that teenagers and young adults tend to be co-operative but often difficult to track down, whereas the elderly are easy to find, but tend to be less co-operative. Thus given the lack of strong evidence for compensating biases, one could argue that any reduction in nonresponse is a reduction in nonresponse bias. But more research is badly needed in this area. Until such time, it is advisable for field offices ideally still to have the reduction of bias in their minds rather than simply the raising of response rates.

Finally, the lack of findings with respect to the visible characteristics of interviewers (such as their age and gender) suggests that these should be of no concern for interviewer recruitment in general-purpose household surveys.

CHAPTER 6 ISOLATING INTERVIEWER EFFECTS ON NONRESPONSE

6.1 Background ²⁰

The effects introduced by interviewers into surveys are typically viewed as correlated variance. From the perspective of the correlation model (see Section 1.2.2), the increase in the variance of a mean is due to the interviewer's creation of positive correlations between the response deviations contained in (almost all) survey data (see, for example, Hansen, Hurwitz, and Bershad, 1961; Fellegi, 1964, 1974; among others). Such studies of correlated interviewer variance have typically focused on the impact of interviewers on substantive results rather than nonresponse (see, for example, Kish, 1962; Wiggins, 1979, 1985; O'Muircheartaigh and Wiggins, 1981; Collins and Butcher, 1982; Groves and Magilavy, 1986; O'Muircheartaigh and Campanelli, 1998).

Interviewer variance with respect to nonresponse error has rarely been considered. It is best conceptualised from the perspective of the ANOVA model (see Section 1.2.1). Under the ANOVA model, the correlated interviewer variance is seen to arise from the systematic differences in interviewers' response rates. The focus is then on this variability among interviewers rather than on the overall level of nonresponse.

The aim of this Chapter is to tease out this variability among interviewers with respect to nonresponse from variability due to sample areas. This is far from straightforward, since interviewers in the UK are normally assigned clusters of addresses within particular areas and the demographic, socio-economic and other attributes of areas and their inhabitants are themselves known to influence the average probability of securing a successful interview. As pioneered by Mahalanobis (1946) and expanded by Kish (1962), Hartley and Rao (1978),

20 A summary of the results from this Chapter are published in Campanelli and O'Muircheartaigh (1999).

Hansen, Hurwitz and Bershada (1961), Fellegi (1964; 1974) and others, what is needed to measure interviewer variance is an interpenetrated design in which households/respondents are assigned at random to interviewers across areas (see Section 1.2). Due to field cost considerations, such designs are rare. As described in Section 1.3, this thesis makes use of data from the second wave of the BHPS which contains a modified interpenetrated design in a subset of areas.

The goal of separating the effects of interviewers from the effects of areas, requires acknowledgement of the cross-classified and hierarchical nature of the data created by the interpenetrated sample design in Wave 2 of the BHPS. More specifically, households are seen to be nested within the cross-classifications of interviewer by PSU, which are in turn nested within larger geographic pools. Following on from the typical ANOVA model to measure interviewer variance (see Section 1.2.1), a basic analysis can be conducted using hierarchical analyses of variance through the SPSS MANOVA procedure (see Section 1.4.2). A better approach is the use of cross-classified multilevel logistic regression models (see Section 1.4.5) which enable one to separate out the random effects of the interviewer on nonresponse from the random effects of the area. The cross-classified multilevel logistic regression modelling approach also allows for the proper framework for a dichotomous dependent variable. More importantly, it facilitates the inclusion of other variables into the model (such as the characteristics of interviewers and the areas as well as other covariates) to help explain these effects. (Note that ML3 was the version of the multilevel software available at the time the nonresponse analyses reported in Section 6.3 were conducted.)

6.2 Variation Between and Homogeneity Within

In order to separate interviewer from area effects on nonresponse effectively, it is first useful to ascertain that there is indeed variation in response rate. Wave 2 BHPS data

suggested that there was considerable variation in the response rate by PSU, 57 to 100 percent. This is captured mainly by variation between geographic pools where response rates range from 67 to 96 percent. The lowest response rates occurred in one inner London geographic pool, two of the geographic pools that were formed by pairing inner and outer London PSUs, and the geographic pool in the West Midlands conurbation. There was also great variation in response rate by interviewer: 58 to 100 percent.²¹

As described in Section 1.2, the intraclass correlation coefficient (ρ) provides a measure of the homogeneity within an interviewer's assignment. Such homogeneity of randomly assigned respondents is presumed to be due to the influence of the interviewer. The primary interest in ρ from a sampling standpoint, however, is as a measure of the homogeneity within a sample PSU. In the case of geography, homogeneity is seen to occur through the similarity of residents in local areas.

As generated from hierarchical analyses of variance (see Section 1.4.2), Table 7 shows the ρ values associated with PSUs, geographic pools, and interviewer assignments. All of these values are generally in keeping with the ρ values found in substantive studies. As outlined in Section 1.2.5 and Table 2, ρ values for interviewer effects on substantive results are typically less than 0.02. A similar summary could be made for the effects of geographic clustering on substantive results (see, for example, Kish, 1965 and Lynn and Lievesley, 1991). The findings are also in line with the work of O'Muircheartaigh and Campanelli (1998), summarised in Chapter 9, who found that for substantive variables,

21 These figures for interviewers and the ones in Tables 7-9 are based on clusters determined from interviewer assignments to areas. In practice, households within an interviewer assignment area could still be handled by an alternative interviewer if needed. This did occur, but infrequently (in less than 3 percent of cases). The estimates provided are still robust.

the ρ values due to interviewers tended to be of the same magnitude as the ρ values due to sample clustering.

An examination of Table 7 suggests that there is significant homogeneity for the individual level nonresponse categories within geographic pools, within PSUs, and within interviewers. Significance is seen less often for the household level nonresponse categories as the sample size is smaller. In general, the magnitude of the impact of the interviewers in comparison to areas (PSU and geographic pool) is largest for household and individual level refusals, followed by whole household non-contacts. The ρ values for individual level nonresponse and individual level non-contacts are comparable to ρ values for areas (either PSU or geographic pool) and smaller than the effects of geographic pool in the case of whole household nonresponse.

6.3 Results from the Cross-Classified Multilevel Models ²²

Under the multilevel modelling framework, first considered were simple variance components models including random effects terms for household/individuals (depending on the analysis), interviewers, PSUs, and the larger geographic pools. The results of these analyses are shown in Table 8.

Several things need to be said initially about these random parameters. First, it should be noted that the household/individual level variation is purposely constrained to 1.0 through the assumption of binomial variation. (To test for the presence of extra binomial variation, this constraint can be removed.) Second, we need to remember that these parameter estimates are on the logistic scale and thus can neither be directly interpreted as

22 These analyses are revisited in Chapter 8, with a multinomial approach and in conjunction with Wave 3 data. A more in-depth examination of the technical adequacies of the multilevel models

Table 7: Estimates of ρ from Hierarchical Analyses of Variance [‡]

	Household Level			Individual Level		
	Non-respondents	Refusals	Non-contacts	Non-respondents	Refusals	Non-contacts
GEOGRAPHIC POOL	0.023 [‡]	0.000	0.008	0.023 [‡]	0.012 [‡]	0.007 [†]
PSU	0.000	-0.006	0.017	0.012 [†]	0.013 [†]	0.011 [†]
INTERVIEWER	0.015	0.018	0.025	0.021 [‡]	0.030 [‡]	0.011 [†]

[‡] These estimates are based on the sample of 25 two-by-two geographic pools as the SPSS MANOVA procedure would not simultaneously handle two-by-two and three-by-three pools. In addition, Wave 1 proxy cases are present. Thus the base for these analyses differs somewhat from those for Tables 8 and 9. Also hierarchical analysis of variance should ideally be used with a continuous dependent variable or one where the proportion is not close to 0 or to 1.

[†] $p < .05$, [‡] $p < .01$, ^{‡‡} $p < .001$. As a ‘rough’ guide to the significance of the ρ values produced from the hierarchical analyses of variance, the significance test for the ‘Between Squares’ component of the hierarchical analyses of variance has been reported (i.e., larger variance between groups translates into a larger degree of homogeneity within groups and thus a larger ρ). Note that these F tests are dependent on the assumption of homogeneity of variance across groups, which is likely to be violated in the current analyses because of the binary nature of the dependent variables. In the presence of heteroscedasticity, the estimate of the mean square error is inflated, making it less likely to declare a given F test significant. Thus when this assumption is violated, the F tests err on the side of being conservative.

Table 8: Cross-Classified Multilevel Logistic Regression Models: Variance Components Model [‡]

	Household Level			Individual Level		
	Non-respondents	Refusals	Non-contacts	Non-respondents	Refusals	Non-contacts
GEOGRAPHIC POOL	0.09(.13)	0	0.12(.24)	0.03(.09)	0.07(.10)	0.05(.16)
PSU	0	0	0	0.02(.06)	0	0.25(.18)
INTERVIEWER	0.23(.16)	0.27(.18)	0.27(.26)	0.28(.12)*	0.39(.15)*	0.16(.16)
H'HOLD/INDIVIDUAL	1	1	1	1	1	1

[‡] Random parameters are based on the logistic scale so that they cannot be directly compared to Table 7.

* Significance at $p < .05$ assessed by Wald test ($p < .01$ and $p < .001$ not shown).

Table 9: Cross-Classified Multilevel Logistic Regression Models: Random Effects from Covariates Model (from Table 6) [‡]

	Household Level			Individual Level		
	Non-respondents	Refusals	Non-contacts	Non-respondents	Refusals	Non-contacts
GEOGRAPHIC POOL	0.08(.15)	0	0.24(.28)	0	0	0
PSU	0	0	0	0.03(.07)	0	0.20(.13)
INTERVIEWER	0.29(.20)	0.23(.19)	0.08(.31)	0.23(.09)*	0.40(.15)*	0
H'HOLD/INDIVIDUAL	1	1	1	1	1	1

[‡] Random parameters are based on the logistic scale so that they cannot be directly compared to Table 7.

* Significance at $p < .05$ assessed by Wald test ($p < .01$ and $p < .001$ not shown).

are also presented in Chapter 8.

proportions of variance nor directly compared with the ρ values in Table 7. The third consideration is the influence of sample size. With the individual level analysis having roughly twice the number of cases as the household level analysis, the household level parameters which are of about the same magnitude as the individual level ones are often not significant although the individual ones are. In addition, as the multilevel models allow the simultaneous modelling of both the 2x2 and 3x3 geographic pools, the estimates are based on a slightly larger sample size than the hierarchical analysis of variance estimates in Table 7.

Bearing all of this in mind, the multilevel variance components models shown in Table 8 generally show the interviewer making a larger effect than the PSU or geographic pool. The one exception is for individual level non-contacts, where the random effect of PSU is larger than that for interviewers, but the interviewer still makes a contribution. These are similar to the conclusions from Table 7, although the stronger impact of the interviewer rather than the area (PSU or larger geographic pool) is seen more clearly.

The simple variance components models were then rerun using all of the covariates described in Chapter 5. When all of the covariates are controlled for as shown in Table 6 in Chapter 5, a different picture emerges for the random effects. The random effects portion of Table 6 has been copied to create Table 9, for ease of comparison. Comparing Tables 8 and 9 shows that the importance of the interviewer in the refusal component of nonresponse is clearly maintained, but the role of the interviewer in the non-contact component of nonresponse virtually disappears and it is the area, rather than the interviewer, which is the important effect. Given the findings from Chapter 5, about the minor contributions of interviewer's age, gender, years of experience, and grade level

with respect to nonresponse, we suspect that these random terms are picking up the more intangible aspects of interviewer behaviour. This could include, for example, the quality of their doorstep behaviour (see, for example, Campanelli, Sturgis, and Purdon, 1997) or perhaps other unmeasured elements such as interviewer expectations and attitudes.

6.4 Is a Household Level Needed?

As described in the Section 1.4.6, there is a concern that the individual level multilevel models should ideally include the household level in the model. This was not feasible to test with ML3, when these models were first created, but this issue has now been re-visited with MLwiN. First, the variance components model for individual level nonresponse from Table 8 was rerun shifting the random effects of interviewers and areas to higher levels and including households as level 2. Second, the corresponding covariates model from Table 6 was rerun. The results are shown in Tables 10a and 10b respectively. As can be seen in comparing the random effects from column one with those in column three in both tables, a large effect for household level variation is found. The other random terms in the model and their standard errors (as well as the fixed effects terms and their standard errors) remain approximately the same with or without the presence of the household level term. The key exception is the random effects term for interviewers. In Table 10a it shrinks to about half of its original size and in Table 10b, it shrinks to about a third of its original size and in both cases is clearly non-significant. The addition of a household level term was also considered in two other models (not shown here). The first was the individual level refusals model developed in Chapter 7 (see Model 4 in Table 12). The second was the individual non-contacts model from Table 8. Each showed significant household level variation and each showed a sizeable reduction in the interviewer variation. Whereas the addition of the household level term had no effect on the other random terms in the individual level refusals model, there was a reduction in the PSU

Table 10a: Comparing an Individual Level Nonresponse Model With and Without Household Level Variation: Variance Components Model

	From Table 6	Relaxing the Binomial Assumption	With Household Level Variation	With Household Level Variation and Relaxing the Binomial Assumption
GEOGRAPHIC POOL	0.03 (0.09)	0.03 (0.09)	0.04 (0.09)	0.05 (0.09)
PSU	0.02 (0.06)	0.03 (0.06)	0.0	0.0
INTERVIEWER	0.28 (0.12)*	0.29 (0.12)*	0.16 (0.12)	0.16 (0.11)
HOUSEHOLD	NA	NA	1.94 (0.26)*	5.13 (0.27)*
INDIVIDUAL	1	0.94 (0.03)	1	0.32 (0.01)
CONSTANT	-1.83	-1.83	-1.81	-1.77

* Significance at $p < .05$ assessed by MLwiN Wald test ($p < .01$ and $p < .001$ not shown).

Table 10b: Comparing an Individual Level Nonresponse Model With and Without Household Level Variation: Covariates Model \oplus , ϕ

	From Table 6	Relaxing the Binomial Assumption	With Household Level Variation	With Household Level Variation and Relaxing the Binomial Assumption
FIXED EFFECTS				
SAMEINT	-0.14 (0.15)	-0.14 (0.14)	-0.11 (0.17)	-0.14 (0.17)
TOTCNTW2	0.21 (0.04)*	0.21 (0.04)*	0.22 (0.05)*	0.22 (0.05)*
AIVFHO3	1.51 (0.22)*	1.50 (0.21)*	1.46 (0.27)*	1.41 (0.29)*
AIV42(1)	1.65 (0.69)*	1.64 (0.66)*	1.54 (0.77)*	1.25 (0.57)*
AIV42(3)	0.74 (0.15)*	0.75 (0.15)*	0.72 (0.18)*	0.60 (0.15)*
AIV42(4)	2.38 (0.61)*	2.37 (0.57)*	2.37 (0.71)*	2.26 (0.61)*
AIV6D2(1)	-2.25 (0.80)*	-2.22 (0.75)*	-2.07 (0.86)*	-1.37 (0.59)*
AIV6D2(2)	0.69 (0.39)	0.67 (0.37)	0.86 (0.44)	0.92 (0.39)*
AHSTYP(1)	0.36 (0.39)	0.36 (0.37)	0.29 (0.46)	0.18 (0.48)
AHSTYP(2)	-0.05 (0.26)	-0.05 (0.25)	0.03 (0.30)	0.15 (0.30)
AHSTYP(3)	-0.32 (0.19)	-0.33 (0.18)	-0.32 (0.22)	-0.34 (0.22)
AHSTYP(4)	-0.57 (0.21)*	-0.58 (0.20)*	-0.61 (0.24)*	-0.61 (0.24)*
ANCARS(0)	0.57 (0.23)*	0.58 (0.22)*	0.58 (0.27)*	0.62 (0.27)*
ANCARS(1)	0.47 (0.19)*	0.48 (0.18)*	0.45 (0.23)*	0.46 (0.22)*
ACD8USE	0.49 (0.18)*	0.49 (0.17)*	0.47 (0.21)*	0.42 (0.20)*
AXPHSN(1)	0.11 (0.21)	0.11 (0.20)	0.15 (0.25)	0.17 (0.25)
AXPHSN(2)	0.61 (0.20)*	0.61 (0.19)*	0.65 (0.24)*	0.66 (0.25)*
AXPHSN(3)	0.24 (0.19)	0.24 (0.18)	0.32 (0.23)	0.39 (0.22)
ANWAGE2	0.55 (0.21)*	0.56 (0.20)*	0.55 (0.27)*	0.61 (0.30)*
ANPENS2	0.74 (0.26)*	0.73 (0.25)*	0.68 (0.31)*	0.60 (0.28)*
AAGE2(1)	0.86 (0.17)*	0.86 (0.16)*	0.82 (0.19)*	0.72 (0.15)*
AAGE2(3)	-0.04 (0.29)	-0.04 (0.28)	-0.08 (0.33)	-0.16 (0.25)
ASEX	0.37 (0.13)*	0.36 (0.13)*	0.39 (0.14)*	0.40 (0.10)*
AHGEST(2)	0.09 (0.23)	0.09 (0.22)	0.08 (0.27)	0.08 (0.22)
AHGEST(3)	-0.54 (0.19)*	-0.54 (0.18)*	-0.48 (0.21)*	-0.30 (0.15)
PERDISC	0.03 (0.02)	0.03 (0.02)	0.02 (0.02)	0.01 (0.00)
PERCNWT	0.00 (0.01)	0.00 (0.01)	0.01 (0.01)	0.01 (0.01)
RANDOM EFFECTS				
GEOGRAPHIC POOL	0.0	0.0	0.0	0.0
PSU	0.04 (0.07)	0.06 (0.07)	0.0	0.0
INTERVIEWER	0.31 (0.11)*	0.34 (0.11)*	0.13 (0.09)	0.14 (0.09)
HOUSEHOLD	NA	NA	1.94 (0.29)*	5.64 (0.34)*
INDIVIDUAL	1	0.89 (0.03)	1	0.39 (0.02)
CONSTANT	-4.09	-4.11	-4.03	-3.88

* Significance at $p < .05$ assessed by MLwiN Wald test ($p < .01$ and $p < .001$ not shown).

ϕ Note that when individual level variables are analysed at the household level, they refer to the characteristics of the 'household reference person', the person in whose name the property is owned or rented.

variation in the individual level non-contacts model. These examples raise potential concerns about the importance of the interviewer effects, in particular, in the individual level models.

But how adequate are the models with the household level terms? The fourth column in both Table 10a and 10b suggest that models with the household level terms appear to diverge significantly from the assumption of binomial variation whereas this is not the case for the simpler models without the household term which appear in the second column. (Although not shown, even further divergences were noted when RIGLS and PQL estimation were used.²³ With RIGLS and PQL estimation, the level 1 term shrinks to 0.12 with a standard error of (0.01) for a variance components model and 0.12 with a standard error of (0.00) for a covariates model. In turn, the level 2 term (random effects for households) increase to 11.9 with a standard error of (0.81) for a variance components model and 13.9 with a standard error of (1.00) a the covariates model.) Thus it appears that individual level models with a household level term are not well behaved and the ideal solution is not completely clear. Nevertheless, as all subsequent individual level models in this thesis do not include a household level term, the magnitude of the random effects terms, particularly for interviewers, should be considered with some caution.

6.5 Chapter Summary and Discussion

In this Chapter we have investigated the impact of interviewers on response rates. Interviewers can be seen to vary systematically in terms of their response rates and the focus has been on this variability among interviewers rather than on the overall level of nonresponse. In particular, the focus has been on separating this variability among interviewers from the corresponding variability among areas.

23 ^{2nd} order estimation did not converge. For more information about RIGLIS, PQL and the 2nd order approximation, see Section 1.4.8.

The results suggest that some of the variation in nonresponse is due both to area effects and interviewer effects. The bivariate analyses conducted in Chapter 5 suggested that several of the easily measurable characteristics of interviewers (such as age, gender, experience, and grade level) and area (such as population density, proportion of flats in the area, percentage of non-white residents, etc. taken from 1991 Census small area data) are important predictors of nonresponse. Interestingly, these relationships almost completely disappeared when the characteristics of specific households and respondents were controlled for. However, the importance of the interviewer and the area reappears in this Chapter through the cross-classified multilevel analysis which treats the effects of interviewers and areas as random. This suggests that what is making a difference in terms of nonresponse is more subtle and elusive than the easily measured characteristics of interviewers and areas. Studies which simply look at standard interviewer characteristics would miss this.

Looking at the variance components models, it was seen that the effects of the interviewer often tend to predominate over those of the area (PSU or larger geographic pool). After controlling for the characteristics of households and individuals, however, the interviewer effects remained strong for individual and whole household refusals, but virtually disappear in the case of individual and whole household non-contacts where the area is the more important factor.

In general, however, the size of these effects are modest. Only the random effects due to the interviewer reach significance and this is only for the individual level models where the sample size is twice that of the household level models.

It should be remembered, however, that the significant interviewer effects at the individual level were greatly reduced when a random effects terms for households was added to the model. Although there was some concern over the adequacy of the model with the household level term, caution should nonetheless be used in interpreting the random effects, for interviewers in particular, in the individual level models. Thus the findings in this thesis should be considered as indicative, rather than definite.

These basic analyses are revisited in Chapter 7 from a multinomial perspective and in conjunction with Wave 3 data. Various technical aspects of the multilevel models are also considered in Chapter 7.

CHAPTER 7 EXPLORING INTERVIEWER AND AREA EFFECTS WITH A MULTINOMIAL APPROACH; EXPLORING THE PERSISTENCE OF INTERVIEWER AND AREA EFFECTS AT WAVE 3 ²⁴

7.1 Introduction

This Chapter looks again at Chapter 6 which had focused on separating the variability in response rates due to interviewers from that due to areas. It considers this issue from the perspective of a cross-classified multilevel multinomial model which will allow us to see the extent to which the two main components of nonresponse, refusals and non-contacts, are related within interviewers and within areas. Also considered is the extent to which the influences of interviewers and areas are maintained over time by making use of Wave 3 data in addition to Wave 2 data.

This Chapter also serves to illustrate various points about multilevel analysis using the software, MLwiN (see Goldstein *et al*, 1998). More specifically, it focuses on a comparison of cross-classified multilevel logistic regression models with cross-classified multilevel multinomial regression models and the comparison of cross-classified multilevel multinomial regression models with single level (non-hierarchical) multinomial regression models.

7.1.1 *The Relationships Between Refusals and Non-contacts*

As described in Section 3.1, the two main components of nonresponse are refusals and non-contacts. It can be argued that these offer very different challenges for the interviewer. Minimising whole household refusals requires interviewers to draw on their

24 The material from this Chapter was first presented in October of 1998 at a one day conference on the Application of Random Effects/Multilevel Models to Categorical Data in the Social Sciences and Medicine, London sponsored by the Analysis of Large and Complex Datasets programme of the ESRC, the Medical and Social Statistics sections of the RSS and Series A of the Journal of the Royal Statistical Association. It has been subsequently submitted for publication as

persuasive skills (see, for example, Groves, Cialdini and Couper, 1992; Morton Williams, 1993; Campanelli, Sturgis, and Purdon, 1997), while minimising whole household non-contacts requires their persistence in making follow-up calls as well as their ingenuity in how they structure their calls. Although survey organisations offer guidelines to interviewers about how many calls to make and the timing of these calls (see, for example, Social and Community Planning Research, 1995), good interviewers often go well beyond the required number of calls. In addition, studies have shown that despite this general guidance, interviewers do indeed differ in how they structure their calls and in their final non-contact rates (see, for example, Lievesley, 1986; and Campanelli *et al*, 1997). Given that whole household refusals can offer very different challenges to whole household non-contacts, we would hypothesise a negative relationship between interviewers' whole household refusal rate and their whole household non-contact rate, with some interviewers being better at reducing the refusal side of nonresponse and others being better at reducing the non-contact side of nonresponse. Once a household has been contacted, there is then the issue of *finding* and *convincing* all members of the household to participate, including the elusive ones. At this individual level, we would hypothesise that the two skills (i.e., that of *finding* and that of *convincing*) are much more similar and that a positive correlation should be found. To our knowledge, neither of these hypotheses has ever been experimentally tested.

Certain characteristics of areas have been shown to increase both the refusal and non-contact portions of nonresponse. Take, for example, urban areas. Increased urbanicity usually translates into increased refusal rates because of reduced norms of helping behaviour and through greater fear of crime (see, for example, House and Wolf, 1978).

Increased urbanicity also increases non-contact rates as the life style of urban residents may mean they are harder to find at home or that they are simply harder to contact because they live in multi-storey blocks of flats (see, for example, Lynn, 1996). Chapter 5 explored several other aspects of areas in relation to nonresponse using 1991 Census small area statistics data. Findings suggested that non-contacts were more likely to occur in the poorer economic areas, i.e., those with a higher proportion of housing with a small number of rooms, or no car ownership, or local authority renters, or manual workers, or the unemployed, or individuals with lower qualifications. Non-contacts were also more likely in areas with a higher proportion of one person households, households with no families, and households with a lone parent with dependent children or with a higher percentage of the long term ill, or a higher percentage of people aged 15-34. Also found was the fact that refusals were more likely in areas with a higher percentage of renters, or lone parents with dependent children, or non-white residents, or the elderly, or migrants, or areas with a higher percentage of industries focusing on distribution, hotel and catering and other services. Along the urban/rural dimension of areas, we would hypothesise that the refusal and non-contact components would be positively correlated. With respect to other characteristics of areas, however, it is unclear how the two components would interact. With respect to individual effects, a negative correlation would be expected as non-contacts are often hidden refusals and thus a high individual non-contact rate could mean a low individual refusal rate.

7.1.2 Interviewer and Area Effects at Wave 3

Over time, one would hypothesise that the influence of interviewers and areas would be less, as one would expect the truly uncooperative and hard-to-find would have already been lost to survey attrition. Rarely, however, has this hypothesis been examined.

7.2 The Data

This Chapter makes use of data from both Waves 2 and 3 of the BHPS from the interpenetrated sample design sub-sample. As described in Section 1.3.2, the interpenetrated design at Wave 2 creates a hierarchical cross-classification in which households or individuals are seen to be nested within the cross-classification of interviewer by PSU, which are in turn nested within larger geographic pools, thus allowing for the separation of interviewer and PSU effects within geographic pools. As described in Section 1.3.3, the random assignment of interviewers at Wave 3 was exactly reversed. At Wave 3 the data can be viewed as a hierarchical cross-classification in which households or individuals are seen to be nested within the cross-classification of interviewer/PSU by interviewer continuity, which are in turn nested within geographic pools. This allows for the continued assessment of interviewer continuity (see Section 8.2), but means that the interviewer and PSU effects are again confounded.

At Wave 1, all individuals in 69 percent of households in the full sample were either interviewed directly or by proxy. Excluding ineligible cases and a very modest number of 'unable to participate households/individuals' (i.e., 1.1 percent or less in all cases), the response rates for the interpenetrated sub-sample for Waves 2 and 3 are as shown in Table 11. Note that all of the household analyses in this Chapter are conditioned on Wave 1 fully and partially co-operating households and all of the individual analyses are conditioned on Wave 1 full interview cases.

Table 11: Response Figures for Waves 2 and 3 of the Interpenetrated Sub-Sample †

	Interviews	Refusals	Non-contacts	n
Wave 2 households §	87%	7%	5%	1468
Wave 2 individuals ‡	87%	7%	6%	2429
Wave 3 households §	87%	6%	7%	1432
Wave 3 individuals ‡	85%	9%	7%	2258

† Excluding a small handful of ineligible cases ineligible at Waves 2 and 3 (1.1% or less in all instances)

§ Based on Wave 1 co-operating and partially co-operating households

‡ Based on Wave 1 fully co-operating individuals

7.3 Methods

7.3.1 Cross-classified Multilevel Models

Isolating interviewer effects from area effects on refusals and non-contacts can be accomplished through separate cross-classified multilevel logistic regression models.

Exploring the covariance (correlation) between these two components for both interviewers and PSUs, however, requires a multinomial approach with the polytomous dependent variable: refusals, non-contacts, and interviews. Note that this correlation cannot be obtained from a standard (non-hierarchical) multinomial model. For comparison purposes we have conducted both the separate cross-classified multilevel logistic regression analyses in addition to the multinomial one. To ensure comparability between these two approaches, the dependent variables for the two logistic models are set up to model the odds of being a refusal as opposed to being a respondent and the odds of being a non-contact as opposed to being a respondent.

The regression equation for the cross-classified multilevel logistic model is similar to that of a standard logistic regression model, except that random effects terms are added to capture the random departures due to geographic pool, and the cross-classification of PSU by interviewer, in addition to the individual level error term (see Section 1.4.5). In the multinomial case examined in this Chapter, two equations are needed, one for each of the dependent odds being considered. When this is implemented in MLwiN, the first two levels are used to form the multivariate structure (see Section 1.4.5). To my knowledge, this Chapter represents one of the first uses of a cross-classified multilevel multinomial model using the MLwiN software.

7.3.2 *Analysis Plan*

Variance Components Models and the Comparison with the Multinomial Model. First considered are binary variance components models at both the household and individual levels, separately for refusals and non-contacts. These are then compared to a multinomial model at both the household and individual level, in which the covariation between the random term for refusals and the random term for non-contacts can be estimated for interviewers, for PSUs, and for geographic pools. As discussed in Section 7.1.1, these covariances are of particular interest from a substantive standpoint and will be expressed as correlations for ease of interpretation. These covariances (correlations) are also of interest from an analytic standpoint. Due to allowing these covariances, the results from the cross-classified multilevel multinomial regression will in principle differ from the separate cross-classified multilevel logistic regressions. This is not necessarily the case for single-level separate logistic regressions versus multinomial regression. As suggested by Begg and Gray (1984, p. 12), “*the two models are parametrically equivalent*” and “*if*

maximum likelihood estimation is employed, the estimates, \tilde{B}_d , will be asymptotically unbiased.” Alvarez and Nagler (1998 p. 56) argue that separate logistic regressions and multinomial regression “*produce almost identical estimates . . . and the only real difference between the two techniques is that the multinomial logit produces more efficient estimates.*” Thus, it appears that the separate logistic regression method can be used as a way to generate estimates of parameters and their standard errors for multinomial regression, but there is some controversy around this. Menard (1995, p. 90) argues along with Hosmer and Lemeshow (1989) that separate logistic regressions do “*not appear to produce results sufficiently consistent with the multinomial logit/polytomous logistic regression model to warrant its use except for exploratory or diagnostic purposes*”. In addition, there are concerns with the individual logistic regression method in terms of providing proper joint tests of parameters from different regressions and in developing confidence intervals for predicted probabilities.

It is also interesting to note that both the individual logistic regressions method and the multinomial method implicitly assume that the ratio of the probabilities for a given alternative in the dependent variable are not affected by the addition of another alternative such as a third candidate in a two candidate election. This is the property of the “Independence of Irrelevant Alternatives” (see Alvarez and Nagler, 1998, p. 57; and Greene, 1997). In reality, however, there may be times when this may not hold, particularly when 2 or more of the alternatives are very similar. Take, for example, the entrance of a third candidate just before the election who represents the same ideological position as one of the existing candidates. As suggested by Alvarez and Nagler (1998), a multinomial probit model offers a better alternative than a multinomial logit model when there are correlations between the alternative choices posed by the dependent variable, as

in the election example. However, software to estimate multinomial probit models is rare given the estimation resources needed. (Of particular interest is the fact that MLwiN offers a way to build a complex variance-covariance structure to estimate correlations between the random terms for each of the categories of the dependent variable.)

Note that variance components models reported in this Chapter differ from those reported in Chapter 6. First, the dependent variables are constructed differently for comparison with the multinomial models (e.g., refusals versus respondents and non-contacts versus respondents in this Chapter as opposed to refusals versus respondents and other nonrespondents, and non-contacts versus respondents and other nonrespondents in Chapter 6). Second, due to this comparison, the binary models have been restricted to the same covariates (see below) and sample sizes as were used in the multinomial case. Third, the current binary models have benefited from a full investigation of the various options available in MLwiN for more accurate estimation (see Section 1.4.8). Fourth, a slightly different list of covariates has been used.

Covariates Models and the Comparison with a Single Level Multinomial Model. Given that multilevel modelling easily facilitates the use of covariates, the next step was to introduce known correlates of refusals and non-contacts to see if any of the random variation could be explained. The covariates of most interest first appeared to be the characteristics of interviewers and the characteristics of areas. As reported in Chapter 5, however, the easily measurable characteristics of interviewers (such as their age, gender, years of experience and grade level) offer virtually no explanation for the variability of interviewer response rates in the BHPS data. Some relationships between interviewer characteristics and refusals and non-contacts were found in bivariate analyses, but these

interviewer characteristics did not prove to be significant predictors once the characteristics of individuals, households and measures of respondent contactability and co-operation were controlled for. In addition, the findings with respect to grade level were discovered to be due simply to the differential allocation of the few higher grade interviewers to the more difficult areas.

With respect to areas, Chapter 5 suggested that although area characteristics were significant predictors of refusals and non-contacts in bivariate analyses of BHPS data, they were simply proxies for the characteristics of the actual sample members. Once characteristics of individuals, households and measures of respondent contactability and co-operation were controlled for, area effects virtually disappeared. We therefore decided to include as covariates just those variables that had proved to be useful in earlier work (see Chapter 5). Given the complex estimation requirements of the cross-classified multilevel multinomial models, the covariate list was further restricted to those found to make a significant contribution at the $p < .01$ level. (Note that the variable 'total number of Wave 2 calls' is also excluded because the level 1 variance becomes highly over-dispersed (34 times so) with its inclusion. The variable 'total number of Wave 2 calls until first contact' is conceptually clearer and does not cause this over-dispersion and has therefore been used in its place.)

The covariate models are then compared to the variance components models under both the binary and multinomial conditions. Note that any missing data incurred from the covariates was excluded from the variance components models as well to facilitate the comparison. In addition, to further the comparability, the same covariates have been used in all of the household models and all of the individual models. For example, some

covariates are only predictors of refusals while some are only predictors of non-contacts, but as both sets must be included in the multinomial model, they have also been included in the binary response models. This, however, is not problematic as the focus of the analysis is to establish a suitable set of controls rather than to obtain the best fitting fixed effects.

The cross-classified multilevel multinomial model is also compared to a non-hierarchical one using the software, STATA (Hamilton, 1993). The purpose of this comparison is to see if by adding the proper random effects structure, the conclusions about the fixed effects are changed. Another point of interest is that it is possible to only enter one of the pair of covariates (e.g., age only with respect to refusals rather than with respect to both refusals and non-contacts, when using the MLwiN software, but it is clearly not possible to do this in the standard single-level multinomial models). If, of course, one is willing to use the separate logistic regressions approach to estimate your multinomial parameters as suggested by Begg and Gray (1984), then it is clearly possible to enter only one of the pair of covariates.

Exploring the Combined Effects of Interviewers and PSUs at Wave 3. At Wave 3 the random effects of interviewers and areas are again confounded (see discussion in Section 7.2). Simple three level models are considered, with households or individuals nested within interviewers/PSUs within geographic pools. (The cross-classification with interviewer continuity is considered in Section 8.2.) These are conducted as binary rather than multinomial models as the correlation across refusals and non-contacts for interviewers and PSUs is now confounded and no longer of interest.

7.4 Results

7.4.1 *Technical Aspects*

Twelve multilevel models are presented in Tables 12 and 13. As described in Section 1.4.8, the literature suggests that the best estimation for the BHPS nonresponse data with its rather small probabilities of refusals and/or non-contacts would be Restricted Iterative Generalised Least Square (RIGLS) with Penalised/Predictive Quasi-Likelihood (PQL) and the Second-Order approximation. We explored intermediate models as well. RIGLS made a very slight increase in the size of the random and fixed coefficients in relation to their standard errors. PQL made a less modest increase. In contrast, the second-order estimation made a large increase in the size of both the random and fixed estimates in all models.

The assumption of binomial/multinomial variation at level 1 was tested by relaxing the constraint that the level 1 variance be 1.0. As shown Tables D1 and D2 in Appendix D, the individual level models yielded level 1 variance terms close to 1.0. The household level models, however, appear to be under-dispersed. This is interesting given that in practice, the most common tendency is for over-dispersion (McCullagh and Nelder, 1989) and one would expect over dispersion in the individual-level data, because of the household clusters have not been included (see Goldstein, 1995). This under-dispersion was partly alleviated by the second-order estimation. Curiously, however, this problem was eliminated from more of the models with simple MQL/first-order estimation (see Tables E1 and E2 in Appendix E). However, a comparison of both the fixed and random coefficients from Tables E1 and E2 to their counterparts in Tables D1 and D2, show that

the fixed and random coefficients estimated with MQL and first order estimation are under-estimated (grossly in some cases). Thus one could debate which should take precedence, meeting the distributional assumption or ensuring the appropriate magnitude of the parameter estimates.

Given that we have no theoretical explanation for the apparent under-dispersion, we have opted for the more parsimonious model which constrains the level 1 variance to 1.0. This is true for all of the models presented in this Chapter. As the random and fixed effects coefficients change in size in relation to the size of the level 1 variance, this constraint also facilitates the comparison of coefficients across models.

Comparability across models is complicated by the fact that the second-order approximation did not converge for the household level multinomial variance components model. The first-order versions of the binary variance components models were therefore used. It should also be noted that the PQL/second-order models produce larger estimates of the fixed parameters than MQL/first-order estimation and that this needs to be considered in the comparison of the single and multilevel multinomial regressions.

An examination of the level 2 and level 3 residuals for Tables 12 and 13 suggested that normality was a valid assumption, despite the small probabilities associated with being a refusal or a non-contact. The level 1 errors are assumed to follow a binomial distribution which approximates a normal distribution in large samples. The level 1 residuals for these models, however, were clearly non-normal as they were for similar single-level models (see Section 5.3.3). Illustrations for two of the models are shown in Figure 16.

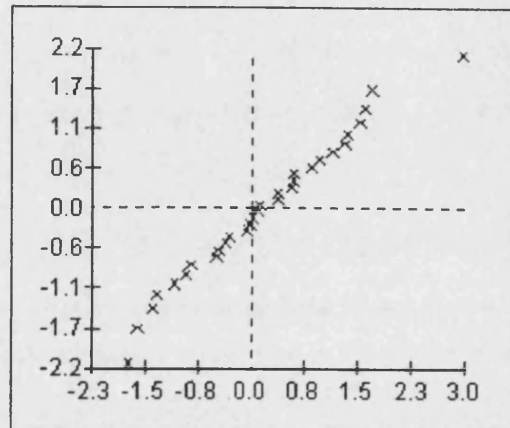
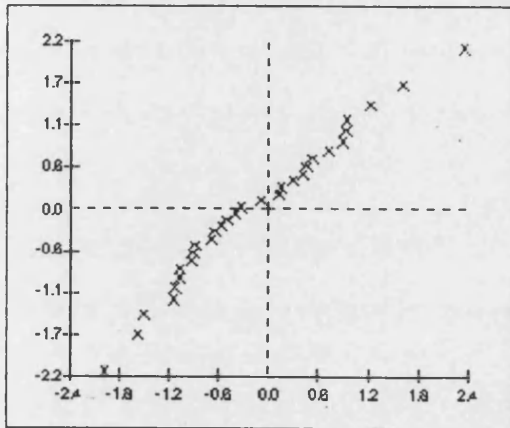
Figure 15: Exploring the Normality of Residuals: Normal Scores Versus Standardised Residuals - Illustrations for Two Models

Covariates Model of Household Level Non-Contacts (Model 8, Table 12)

Covariates Model of Individual Level Refusals (Model 10, Table 13)

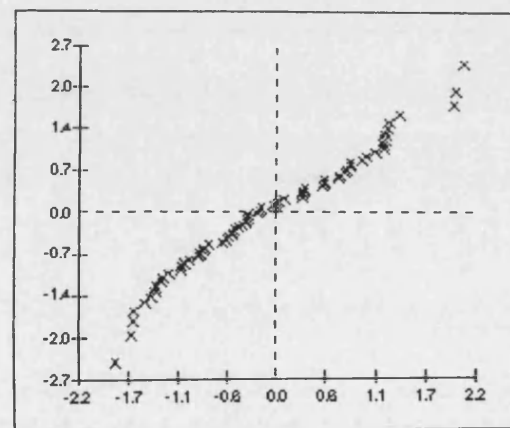
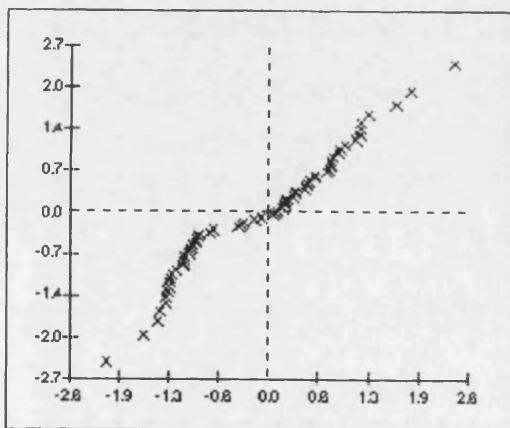
Level 3 - Combined Area Variation (Geographic Pool and PSU)

Level 3 - Combined Area Variation (Geographic Pool and PSU)



Level 2 - Interviewer Variation

Level 2 - Interviewer Variation



Level 1 - Household Variation

Level 1 - Individual Variation

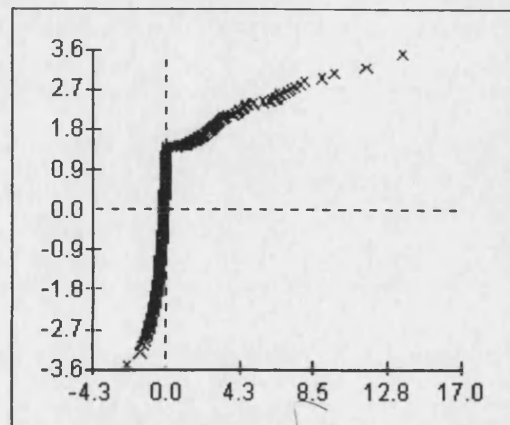
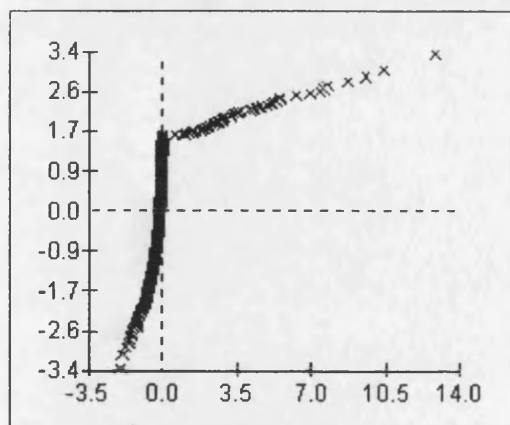


Table 12: Wave 2 Binary and Multinomial Cross-Classified Multilevel Regression: Variance Components Models using RIGLS, PQL and 2nd Order Estimation and Assuming Binomial/Multinomial Variation at Level 1

Household Level Nonresponse (n = 1,365) †				
	Model 1 Binary Refusals	Model 2 Binary Non-contacts	Model 3 Multinomial Estimate	Model 3 Multinomial Correlation between random terms
Fixed Effects				
Refusals	-2.66 (0.13)		-2.66 (0.13)	
Non-contacts		-2.94 (0.15)	-2.94 (0.15)	
Random Effects				
Geographic pools				
Refusals	0.0		0.0	NA
Non-contacts		0.02 (0.26)	0.0	
PSUs				
Refusals	0.0		0.0	NA
Non-contacts		0.10 (0.27)	0.08 (0.23)	
Interviewers				
Refusals	0.32 (0.20)		0.31 (0.19)	0.96
Non-contacts		0.27 (0.31)	0.31 (0.26)	
Households	1	1	1	
Individual Level Nonresponse (n = 2,421)				
	Model 4 Binary Refusals	Model 5 Binary Non-contacts	Model 6 Multinomial Estimate	Model 6 Multinomial Correlation between random terms
Fixed Effects				
Refusals	-2.71 (0.13)		-2.70 (0.13)	
Non-contacts		-2.94 (0.14)	-2.94 (0.14)	
Random Effects				
Geographic pools				
Refusals	0.0		0.0	NA
Non-contacts		0.03 (0.21)	0.0	
PSUs				
Refusals	0.10 (0.13)		0.08 (0.12)	-0.49
Non-contacts		0.31 (0.22)	0.28 (0.18)	
Interviewers				
Refusals	0.49 (0.19) *		0.48 (0.18) *	0.63
Non-contacts		0.24 (0.20)	0.27 (0.18)	
Individuals	1	1	1	

† 2nd order estimation would not converge for this household level multinomial model. RIGLS and PQL were used for the binary household level models to facilitate comparison with the multinomial model.

* Coefficient significant at p < .05 using MLwiN's separate (rather than simultaneous) Wald test (p < .01 and p < .001 not shown).

Table 13: Wave 2 Binary and Multinomial Cross-Classified Multilevel Regression: Covariates Models using RIGLS, PQL and 2nd Order Estimation and Assuming Binomial/Multinomial Variation at Level 1

Household Level Nonresponse (n = 1,365) ϕ					
	Model 7 Binary	Model 8 Binary	Model 9 Multinomial	Model 9a Multinomial	
	Refusals	Non-contacts	Estimate	Correlation between random terms	Single Level
Fixed Effects					
Refusals (constant)	-3.80 (0.33)		-3.90 (0.33)		-3.67 (0.31)
Non-contacts (constant)		-4.91 (0.42)	-4.95 (0.41)		-4.50 (0.35)
W1 num of calls - Ref	-0.02 (0.06)		-0.01 (0.06)		-0.01 (0.06)
W1 num of calls - Ncon		0.16 (0.06) *	0.15 (0.06) *		0.15 (0.05) *
Tot calls till 1 st contact W2 - R	0.24 (0.08) *		0.26 (0.07) *		0.22 (0.07) *
Tot calls till 1 st contact W2 - N		0.29 (0.07) *	0.31 (0.07) *		0.26 (0.06) *
W1 not comp co-op HH-R	1.95 (0.38) *		2.00 (0.37) *		1.90 (0.35) *
W1 not comp co-op HH-N		0.30 (0.74)	0.47 (0.65)		0.28 (0.64)
W1 co-operation rating (base category = very good)					
missing - Ref	0.53 (1.42)		0.58 (1.39)		0.53 (1.39)
good/fair - Ref	1.14 (0.27) *		1.19 (0.27) *		1.11 (0.26) *
poor/very poor - Ref	1.94 (0.75) *		2.39 (0.71) *		1.97 (0.70) *
missing - Ncon		3.57 (1.35) *	3.50 (1.36) *		3.63 (1.23) *
good/fair - Ncon		1.08 (0.35) *	1.11 (0.34) *		0.97 (0.31) *
poor/very poor - Ncon		1.61 (1.30)	2.10 (0.99) *		0.37 (1.26)
Problems with English (base category = no)					
missing - Ref	-1.32 (1.39)		-1.34 (1.38)		-1.30 (1.36)
yes - Ref	0.71 (0.80)		1.01 (0.67)		1.01 (0.73)
missing - Ncon		-3.38 (1.64) *	-3.26 (1.64) *		-3.56 (1.48) *
yes - Ncon		1.96 (0.68) *	2.15 (0.62) *		2.25 (0.61) *
Any pensioners in HH? - Ref	1.30 (0.61) *		1.28 (0.61) *		1.38 (0.58) *
Any pensioners in HH? - Ncon		0.82 (1.01)	0.74 (1.01)		0.93 (0.91)
Age of Head of Household (base category = 26-59)					
15-25 - Ref	-0.22 (0.54)		-0.20 (0.51)		-0.16 (0.49)
60-93 - Ref	-0.22 (0.60)		-0.20 (0.60)		-0.33 (0.58)
15-25 - Ncon		1.81 (0.33) *	1.78 (0.33) *		1.70 (0.30) *
60-93 - Ncon		-1.91 (1.13)	-1.79 (1.12)		-1.93 (1.00)
Random Effects					
Geographic pools					
Refusals	0.0		0.0	NA	NA
Non-contacts		0.0	0.0		
PSUs					
Refusals	0.0		0.0	NA	NA
Non-contacts		0.21 (0.37)	0.14 (0.33)		
Interviewers					
Refusals	0.38 (0.24)		0.46 (0.25)	0.72	NA
Non-contacts		0.54 (0.42)	0.63 (0.41)		
Households	1	1	1		1

ϕ Note that when individual level variables are analysed at the household level, they refer to the characteristics of the 'household reference person', the person in whose name the property is owned or rented.

* Coefficient significant at $p < .05$ using MLwiN's separate (rather than simultaneous) Wald test ($p < .01$ and $p < .001$ not shown).

Table 13: (continued)

Individual Level Nonresponse (n = 2,421)				
	Model 10 Binary	Model 11 Binary	Model 12 Multinomial	Model 12a Multinomial
	Refusals	Non-contacts	Estimate	Correlation between random terms Single Level
Fixed Effects				
Refusals (constant)	-4.71 (0.45)		-4.68 (0.44)	-4.42 (0.36)
Non-contacts (constant)		-4.95 (0.43)	-4.95 (0.40)	-4.50 (0.35)
Interviewer yrs with company (base category = 0-2 years)				
3-5 years - Ref	-0.37 (0.37)		-0.39 (0.36)	-0.29 (0.21)
6-7 years - Ref	0.25 (0.42)		0.26 (0.41)	0.16 (0.25)
8-10 years - Ref	-0.51 (0.58)		-0.54 (0.57)	-0.48 (0.35)
11-13 years - Ref	-0.82 (0.76)		-0.74 (0.73)	-0.75 (0.50)
14-20 years - Ref	-0.34 (0.57)		-0.31 (0.55)	-0.23 (0.35)
3-5 years - Ncon		0.36 (0.34)	0.43 (0.31)	0.20 (0.26)
6-7 years - Ncon		0.53 (0.38)	0.71 (0.35) *	0.31 (0.30)
8-10 years - Ncon		0.71 (0.48)	0.90 (0.44) *	0.29 (0.35)
11-13 years - Ncon		1.10 (0.48) *	1.10 (0.44) *	1.04 (0.35) *
14-20 years - Ncon		0.15 (0.52)	0.14 (0.47)	0.32 (0.40)
Tot calls till 1 st contact W2 - R	0.23 (0.06) *		0.22 (0.06) *	0.21 (0.05) *
Tot calls till 1 st contact W2 - N		0.25 (0.05) *	0.25 (0.05) *	0.26 (0.05) *
W1 not comp co-op HH - R	2.15 (0.25) *		2.14 (0.25) *	1.94 (0.23) *
W1 not comp co-op HH - N		0.63 (0.39)	0.64 (0.36)	0.61 (0.36)
W1 co-operation rating (base category = very good)				
missing - Ref	-0.56 (0.81)		-0.54 (0.78)	-0.69 (0.74)
good/fair - Ref	0.85 (0.21) *		0.84 (0.21) *	0.80 (0.19) *
poor/very poor - Ref	2.46 (0.70) *		2.78 (0.59) *	2.61 (0.64) *
missing - Ncon		0.78 (0.54)	0.77 (0.52)	0.68 (0.49)
good/fair - Ncon		0.92 (0.24) *	0.91 (0.23) *	0.80 (0.22) *
poor/very poor - Ncon		2.99 (0.84) *	3.09 (0.68) *	2.72 (0.80) *
Home computer in accommodation - R	0.77 (0.28) *		0.76 (0.27) *	0.78 (0.25) *
Home computer in accommodation - N		0.53 (0.25) *	0.56 (0.24) *	0.50 (0.23) *
Net monthly housing costs (base category = 251-1650)				
none - Ref	0.57 (0.34)		0.57 (0.33)	0.61 (0.30) *
1-125 pounds - Ref	1.40 (0.33) *		1.39 (0.32) *	1.30 (0.29) *
126-250 pounds - Ref	0.72 (0.33) *		0.70 (0.32) *	0.69 (0.29) *
none - Ncon		-0.11 (0.31)	-0.12 (0.30)	-0.20 (0.28)
1-125 pounds - Ncon		-0.09 (0.32)	-0.09 (0.30)	-0.19 (0.28)
126-250 pounds - Ncon		0.12 (0.27)	0.07 (0.25)	0.01 (0.23)
A couple in the household - Ref	-0.62 (0.23) *		-0.61 (0.23) *	-0.61 (0.21) *
A couple in the household - Ncon		0.18 (0.22)	0.21 (0.21)	0.26 (0.20)
Any pensioners in HH? - Ref	0.92 (0.35) *		0.93 (0.34) *	0.89 (0.31) *
Any pensioners in HH? - Ncon		0.14 (0.49)	0.18 (0.47)	0.07 (0.46)
Respondent gender, men - Ref	0.13 (0.18)		0.14 (0.18)	0.18 (0.17)
Respondent gender, men - Ncon		0.67 (0.20) *	0.63 (0.19) *	0.62 (0.19) *
Age of Head of Household (base category = 26-59)				
15-25 - Ref	0.11 (0.27)		0.14 (0.25)	0.14 (0.24)
60-93 - Ref	-0.24 (0.35)		-0.25 (0.35)	-0.27 (0.32)
15-25 - Ncon		1.37 (0.21) *	1.35 (0.20) *	1.29 (0.19) *
60-93 - Ncon		-0.84 (0.55)	-0.88 (0.53)	-0.72 (0.51)

Table 13: (continued)

	Model 10 Binary	Model 11 Binary	Model 12 Multinomial	Model 12a Multinomial
	Refusals	Non-contacts	Estimate	Correlation between random terms Single Level
Random Effects				
Geographic pools				
Refusals	0.0		0.0	NA NA
Non-contacts		0.12 (0.22)	0.19 (0.20)	
PSUs				
Refusals	0.06 (0.14)		0.05 (0.12)	-0.22 NA
Non-contacts		0.32 (0.23)	0.28 (0.21)	
Interviewers				
Refusals	0.70 (0.24) *		0.69 (0.23) *	1.0*‡ NA
Non-contacts		0.10 (0.18)	0.02 (0.13)	
Individuals	1	1	1	1

* Coefficient significant at $p < .05$ using MLwiN's separate (rather than simultaneous) Wald test ($p < .01$ and $p < .001$ not shown).

‡ The interviewers contribution towards non-contacts is essentially zero and as such has distorted this correlation so that it is greater than 1.0. The actual covariance was 0.38 with a standard error of 0.04.

7.4.2 Substantive Aspects at Wave 2

Random effects in the binary models. Models 1, 2, 4, and 5 shown in Table 12 show effects for the interviewer across all categories of nonresponse, although these only reach significance in the case of individual level refusals (see Section 1.4.8 for a discussion of significance testing). None of the effects for areas reach significance, but are most pronounced in the case of individual non-contacts. After controlling for important covariates of nonresponse in Models 7, 8, 10 and 11 in Table 13, similar conclusions are reached to those in Models 1, 2, 4, and 5, with the exception that interviewers' contribution to individual level non-contacts is reduced, whereas the geographic pool contribution to individual level non-contacts is increased.

The similarity of these results to those found in Chapter 6 are reassuring given the different construction of the dependent variables, the different sample size, the slightly different list of covariates, and the improved estimation procedure. The key difference is found with respect to whole household non-contacts in the covariates model (Table 13,

Model 8). In Table 6 (Chapter 6), the influence of the interviewer on whole household non-contacts was greatly reduced after the various covariates were controlled for. But this does not hold true in Table 13. An investigation into this issue suggests that the main cause of the difference in conclusions is that the household level non-contacts model in Table 6 includes the variable BIVLNC and the one in Table 13 does not. BIVLNC is the number of Wave 2 calls at the last known address. With the exception of a small number of recent movers, it is the total number of calls at Wave 2. Interestingly it was the variable, BIVLNC, which caused an abnormal amount of over-dispersion in the household non-contacts model which was initially being considered for Table 13 (see discussion in Section 7.3). It was also BIVLNC which was discussed in the sub-section entitled, “Assumptions Behind the Model,” in Section 5.3.3 as a variable which would have been better entered as a series of dummy variables rather than as a continuous one. Thus, the version of the household non-contacts model in Table 13 without BIVLNC appears to be the better option.

It is interesting to note that the alternative models shown in Appendix D in which the constraint on the level 1 variance is relaxed, show significant effects of the interviewers for household level refusals and non-contacts in addition to individual level refusals for the variance components models (Table D1) and for household level non-contacts in addition to individual level refusals for the covariate models (Table D2). This suggests the importance of the interviewer effects in general and strengthening the view that interviewers do play a role in whole household non-contacts.

Random effects in the multinomial models. As expected, allowing covariance between the random terms for refusals and contacts, within interviewers and within areas, changes

the size of the random coefficients in the multinomial models. Despite this the cross-classified multilevel multinomial models yield the same conclusions as the binary ones. There are, however, some differences. For example, in the household level covariates models (Models 7-9), one can see that the coefficients for interviewers are larger in the multinomial model than in the separate logistic regression models and that the coefficient for PSUs is smaller. The reverse happens in the individual level covariates models (Models 10-12) where the coefficients for interviewers are smaller in the multinomial model than in the separate logistic regression models and that the coefficient for geographic pools is larger.

It is in the correlations between random terms in the multinomial model, however, that the main interest lies. Large positive correlations were found between refusals and non-contacts for interviewers in each of the multinomial models. In the case of the individual level variance components model (Model 6) and household level covariates model (Model 9), the correlations were not significant. In the case of the household level variance components model (Model 3), the correlation approached significance ($p = 0.057$) and in the case of the individual level covariates model (Model 12), it reached significance.

Note, however, that in this latter case, the extremely small contribution of the interviewer towards the non-contact component distorted the correlation to a value greater than 1.0. (The covariance was 0.38 and its standard error was 0.14). Overall, these findings suggest that good interviewers are actually good at reducing both aspects of nonresponse and poor interviewers are generally poor at both tasks. This is contrary to our hypothesis at the household level, suggesting that although persuading a household to co-operate and finding them at home in the first place require different types of skills from the interviewer, interviewers are generally good or poor at mastering both. Our hypothesis at the individual level within the household was supported.

In the case of PSU (area) variance, a negative correlation between refusals and non-contacts is observed in both the individual level variance components model (Model 6) and the individual level covariates model (Model 12). This is consistent with our hypothesis that individual non-contacts can be hidden refusals. No correlations are observed for the PSU (area) variance for either of the household level models (Model 3 and Model 9) because areas show not impact on refusals. The models in Appendix D, do however show a positive correlation between refusals and non-contacts for the household level covariates model (Table D2, Model 9). This suggests that in areas with higher refusals, more non-contacts are also seen and is consistent with what was hypothesised along the urban/rural dimension of areas.

Fixed Effects in the Binary and Multinomial Multilevel Models. Given the difference in the random parameters between the cross-classified multilevel logistic regression models and the cross-classified multilevel multinomial models due to the covariances and the fact that the fixed parameters are estimated based on the random effects, we would expect differences between these models. When we compare the individual and household level binary covariate models (7, 8, 10, and 11) with the multinomial covariate models (9 and 12) in Table 13, the two sets of analyses yield similar conclusions about the significance of the fixed effects. However, some differences do exist. For example, different conclusions would be reached between Model 8 and Model 9 about whether respondents who were rated with ‘poor/very poor’ co-operation as opposed to ‘very good’ co-operation in Wave 1 are more likely to be non-contacts in Wave 2. Between Model 11 and Model 12, different conclusions would also be reached about whether the effects of interviewers with 6-7 and 8-10 years of experience as opposed to 0-2 years of experience are more likely to have higher non-contact rates.

Non-hierarchical Models. Although smaller in size and based on a variance-covariance matrix that ignores the random effects of interviewers and areas, the coefficients for the non-hierarchical (single level) multinomial models (Models 9a and 12a) produce broadly similar results to their multilevel counterparts (Models 9 and 12, respectively). Three of the four differences between the multilevel and single level multinomial models have to do with the following categories: the ‘poor/very poor’ category of co-operation with respect to non-contacts and the 6-7 and 8-10 years of experience categories with respect to non-contacts. (Interestingly, these are the same three categories discussed above with respect to differences between the separate and multinomial cross-classified multilevel regressions.) In contrast, there is one instance where the single level multinomial model (Model 12a) produces a significant effect and the cross-classified multilevel multinomial one (Model 12) does not. This is for the category ‘no monthly housing costs’ (as opposed to £251-£1650 in housing costs) for refusals. Given the proper variance-covariance matrix used in the cross-classified multilevel multinomial model, these small differences could suggest that the systematic influence of individual interviewers is having some effect on the relationship between substantive variables in the population.

7.4.3 Wave 3 Interviewer/PSU Random Effects

The combined effects of interviewers/PSUs within geographic pools are shown in Table 14 for household and individual refusals and non-contacts (Models 13, 14, 15, and 16). A direct comparison of the estimated coefficients from Table 14 with those given in Table 12 (Models 1, 2, 4, and 5, respectively) is complicated by the fact that 2nd order estimation was not possible for all models. In general, however, we can see that the combined Wave 3 coefficients are smaller. Significance is only reached in the case of individual level refusals. This lack of significance is a combination of the smaller Wave 3 coefficients and the smaller sample size inflating the standard errors. (Similar results are

found in Table 20, Section 8.2.4, which investigates the possibility of random slopes for interviewer continuity while controlling from combined PSU/Interviewer effects at Wave 3.)

Table 14: Combined Random Effects of Interviewers and Areas at Wave 3

	Households (n = 1039)		Individuals (n = 1731)	
	Model 13 Refusals	Model 14 † Non-contacts	Model 15 Refusals	Model 16 Non-contacts
Fixed Effects				
Constant	-2.97 (0.18)	-3.57 (0.20)	-2.43 (0.12)	-3.44 (0.17)
Random Effects				
Geographic pools	0.20 (0.26)	0.07 (0.32)	0.0	0.05 (0.26)
PSUs/Interviewers	0.09 (0.31)	0.02 (0.60)	0.30 (0.14) *	0.22 (0.35)
Households/Individuals	1	1	1	1

† 2nd order estimation would not converge for this household level model. RIGLIS and PQL with 1st order estimation were used.

* Coefficient significant at $p < .05$ using MLwiN's separate (rather than simultaneous) Wald test ($p < .01$ and $p < .001$ not shown).

7.5 Chapter Summary and Discussion

This Chapter revisited the material in Chapter 6 with the use of cross-classified multilevel multinomial regression models and the use of Wave 3 data. Fundamental to these investigations was the use of cross-classified multilevel statistical models. Although multivariate analyses of variance can be used to estimate interviewer variability, the use of cross-classified multilevel logistic regression models (using the MLwiN software) was essential for the accurate estimation of the random effects of interviewer and areas. In addition to facilitating variance components models, MLwiN allows the examination of covariates which can be used to explain the random effects. It was only through the use of cross-classified multilevel multinomial regression models, that we were able to examine the correlation between refusals and non-contacts within interviewers and areas. Given models where the fixed effects rather than the random effects are the main focus,

multilevel modelling allows for the incorporation of the complex variance-covariance structure present in almost all survey data directly into the substantive analysis. For example, a majority of large social surveys in the UK employ face-to-face multi-stage cluster sample designs which can incur positive correlations among units due to the design itself (area clustering) and due to the execution of the design through interviewers.

Results suggested that at Wave 2 the variance introduced by the systematic influence of individual interviewers clearly predominates over the effects due to geographic areas (PSUs and geographic pools) in the case of whole household refusals, whole household non-contacts and individual refusals. Although not significant, the effects of areas are most clearly seen with respect to individual and whole household non-contacts. This pattern holds after controlling for the characteristics of Wave 1 households and individuals and various measures of respondent co-operation and contactability, although the role of interviewers in individual non-contacts is further reduced. The stronger effects of interviewers over areas suggests the need for survey organisations to focus on interviewer training as a successful avenue for minimising nonresponse even in what classically have been considered the difficult areas (such as London and other urban areas). Despite the different construction of dependent variables, smaller sample size, slightly different covariate list, and improved estimation, these results are very similar to those reported in Chapter 6, the main difference being the continued impact of the interviewer in the household level non-contact models. As discussed in Section 7.4, the results of this Chapter are preferred over those in Chapter 6.

Evidence from separate cross-classified multilevel logistic regression models produced very similar conclusions to a cross-classified multilevel multinomial regression models.

The cross-classified multilevel multinomial models, however, allowed us to ascertain the correlation between the refusals and non-contact cases for interviewers and areas, which has not been estimated before. Large positive correlations were found at both the household and individual level for interviewers' role in refusal and non-contact cases. Although not conclusive, this finding suggests that interviewers who are generally good at converting refusals are also generally good at chasing up non-contacts. Thus our hypothesis that interviewers would be differentially good at these two very different types of tasks posed at the household level may be unfounded, although our hypothesis at the individual level was supported. With respect to areas (PSUs), a negative correlation was found at the individual level. This is in line with the view that non-contacts are often hidden refusals. Overall, these correlations provide useful information to survey managers.

These estimated correlations were one contribution of the cross-classified multilevel multinomial models. Such models also allowed us to incorporate covariates to help explain the observed random effects. Comparison with a single-level multinomial model allowed us to assess the extent to which the elaborate variance-covariance structure affects the relationships between these substantive covariates and the dependent variable. The conclusions we would have reached for the analyses we chose were broadly similar, but differed with respect to four specific conclusions. We believe these small differences and the potential for differences in other analyses, justifies the more complex analysis.

Our results with respect to Wave 3 suggest that the impact of interviewers and areas on survey nonresponse are smaller. With the exception of the significant effect of interviewers/PSUs with respect to individual level refusals, this suggests that interviewers

and areas may have less and less of an effect over time as the truly non-persuadable and hard-to-find respondents have already dropped out of the panel survey.

CHAPTER 8 IN LONGITUDINAL STUDIES, DOES HAVING THE SAME INTERVIEWER MAKE A DIFFERENCE TO RESPONSE RATES?

The conventional wisdom in survey research suggests that it is advisable to have the same interviewers return to the same respondents in order to maintain good response rates in longitudinal surveys. There has been, however, very little documented experimental research to support this. The interpenetrated sample design of the BHPS allows for an exploration of this issue. Section 8.1 represents the initial exploration of interviewer continuity using quantitative data from Wave 2 and the qualitative impressions of the interviewers themselves. Section 8.2 revisits interviewer continuity by employing data from Waves 2 and 3 of the interpenetrated sample in comparison to data from Waves 2 through 4 of the full sample.²⁵

8.1 Assessing the Effects of Interviewer Continuity

8.1.1 Background

The Longitudinal Context: 'New' Interviewers to 'Old' Respondents. There are several types of survey designs which are implemented over time (Kalton, 1993; Duncan and Kalton, 1987): repeated surveys (no overlap of units), repeated surveys (partial overlap of units), longitudinal surveys (no rotation group, i.e. classic panel study design), and longitudinal surveys (with rotation). Longitudinal studies which use a rotation group design have been common among government statistical agencies for some time. For

25 The material from Section 8.1 was first presented in September of 1994 at the 5th International Workshop on Household Survey Nonresponse, Ottawa, Canada. It was subsequently published as Campanelli and O'Muircheartaigh (1999).

The material from Section 8.2 was first presented in October of 1998 at a one day conference on the Application of Random Effects/Multilevel Models to Categorical Data in the Social Sciences and Medicine, London sponsored by the Analysis of Large and Complex Datasets programme of the ESRC, the Medical and Social Statistics sections of the RSS and Series A of the Journal of the Royal Statistical Association.

example, the US Current Population Survey (CPS) has employed a longitudinal rotation group design since 1953 to improve the precision of cross-sectional estimation.

Longitudinal designs such as the classic panel study are now becoming more common in non-governmental settings. An example of this trend can be seen in the establishment within the last 15 years of national household panel studies in most European countries. There has also been a growth of interest in panel study methodology as evidenced by recent publications (see, for example, Kasprzyk *et al*, 1989; Magnusson and Bergman, 1990; van de Pol, 1989).

In longitudinal designs (and particularly classic panel studies) it is generally considered advisable to send the same interviewers back to the same respondents at later data collection waves. Such 'interviewer continuity' is thought to improve response rates mainly by reducing the refusal component of nonresponse²⁶ (discussed in this Chapter); and to improve response quality (discussed in Chapter 10). There has been, however, very little documented research to investigate these assumptions. Using the panel component of the British Social Attitudes survey, Waterton and Lievesley (1987, p. 270) reported results "*consistent with the view that the strategy is beneficial - response rates were 3 percent higher when the same interviewer returned.*" As part of their work on doorstep interactions, Couper, Groves and Raghunathan (1996) also mentioned that interviewer continuity had a significant impact on panel survey nonresponse. The clearest illustration of the impact of not using an interviewer continuity strategy is given by Rendtel (1990). Based on German Socio-Economic Panel data, Rendtel found higher

26 At the time of the first wave of data collection, refusals usually make up the largest component of overall nonresponse. Over time panel studies suffer from attrition, but the degree of additional sample loss usually stabilises at a low level. For some, but not all studies, the refusal component still remains the major cause of nonresponse at later waves.

attrition rates among respondents who had experienced a change of interviewer and this difference varied by age group. Among persons 35-54 years of age, those who received a new interviewer were 4 times more likely to be nonresponse cases than those who received the same interviewer. Among persons over the age of 64, those experiencing interviewer 'discontinuity' were 8 times more likely to be nonresponse cases.

Rope (1993) examined the effects of a different interviewer on nonresponse among CPS respondents. The CPS employs a design whereby households are in the sample for 4 consecutive months, out-of-sample for 8 months and then back in the sample for another 4 months and a new panel starts each month (US Bureau of the Census, 1978). Rope found that there was a relationship between having a different interviewer and nonresponse for each month in sample although the probabilities were less, relatively, for the interview waves after the 8 month break. This suggests that households may be comfortable with a different interviewer after being out of the rotation for eight months and that a different interviewer towards the end of the panel survey has less of an effect on response than one at the initial waves of a panel survey. Rope also observed an interesting pattern among households that only had one refusal over time. These refusals tended to occur the month after the interviewer changed.

These four studies provide evidence to support the conventional survey wisdom that interviewer continuity over time facilitates panel study response. This in turn has implications for fieldwork practice. For example, it suggests examining how interviewers are assigned to respondents in longitudinal studies and the need for ways to increase interviewer longevity with the study. On the other hand, it should be pointed out that each of the cited studies is based on non-experimental data. For example, in the majority

of longitudinal studies respondents receive a new interviewer simply because the previous interviewer who covered the assignment has left employment. Such interviewer attrition is rarely random and it is possible that the factors which are associated with interviewer change are also associated with subsequent nonresponse (e.g., it is difficult to keep good interviewers in the difficult survey areas). Thus, confidence in the conclusions of the studies just quoted should be tempered with caution. It can be seen that there is a need for experimental data to shed light on the interviewer continuity issue and to provide insight into the mechanisms of longitudinal survey response/nonresponse.

Why Should Interviewer Continuity Improve Response Rates?

Refusals. As described in the previous section, sending the same interviewers back to the same respondents at subsequent data collection waves is thought to improve response rates by reducing the refusal component of nonresponse, but how specifically does this work?

First it can be hypothesised that an interviewer continuity strategy will help to maintain interviewer/respondent rapport which was developed during the previous interview(s). Research suggests that up to a certain level such rapport should positively affect survey quality (Harkess and Warren, 1993). The impact of rapport is illustrated by the following comment of a face-to-face panel interviewer:

“[interview] relationships matter. People are then more patient with you. They remember more and try to be accurate. They are also able to tell you personal details, which otherwise they might avoid.”²⁷

27 From the British Household Panel Study Interviewer Debriefing Study. This study is described in Section 8.1.2.

As the work of Morton-Williams (1993) suggests, interviewer/respondent rapport may have a direct effect on the refusal component of nonresponse for respondents with certain types of motivation pattern. She found that respondents have one of two basic motivation patterns for participating in a survey. Using labels developed by Kahn and Cannell (1957), these are described as 'intrinsic' and 'extrinsic'. Extrinsically motivated respondents participate because they find the topic or purpose of the survey of inherent interest/value. For this type of respondent, who the interviewer is makes relatively little difference. For intrinsically motivated respondents, however, who the interviewer is can make the key difference with respect to participation. Intrinsically motivated respondents respond to the idea of what it will be like to spend time with a particular interviewer. If the previous interaction with the interviewer had been a pleasurable experience, the respondent would probably anticipate that this year's interview will be likewise. If the previous interaction with the interviewer had not been perceived as a pleasurable experience, the respondent may see the survey as something to avoid. In addition, Morton-Williams (1993) found that the majority of respondents were intrinsically motivated, suggesting the need to take an interviewer continuity effect seriously.²⁸

The work of Groves and Cialdini (1991) is in line with that of Morton-Williams. They suggest that individuals typically use varying degrees of two information processing strategies, referred to as *systematic* and *heuristic*, in making decisions. They go on to suggest that a *systematic* strategy is based on a rational assessment and depends on the individual's interest, time, energy, and cognitive capacity. However, when someone is distracted, tired, or indifferent, a *heuristic* strategy becomes increasingly probable, leading

28 These intrinsic reasons for participation could be less relevant in the case of longitudinal telephone surveys, as the extent of rapport building is not as great as it is in face-to-face surveys. This would suggest that an interviewer continuity effect should be less pronounced with longitudinal telephone surveys although potentially still present. See, for example, the work of Rope (1993) with respect to the U.S. Current Population Survey which starts as a face-to-face

individuals to make their decisions on past behaviour in similar situations. It is likely that the latter strategy applies predominantly to survey respondents and incorporates situational cues such as reactions to the interviewer at the time of previous interviews.

Non-Contacts. Sending the same interviewers back to the same respondents at subsequent data collection waves could also improve response rates by reducing the non-contact component of nonresponse. For example, it seems reasonable that face-to-face interviewers will call on their experience from the previous wave in locating hard-to-find addresses and hard-to-contact respondents. This is less relevant in the case of telephone interviewing.

8.1.2 *The Data*

Quantitative Data. As the interpenetrated sample design experiment was implemented at Wave 2 of the BHPS, the design implicitly allows for the assessment of the differential effects of sending or not sending back the same interviewer at a subsequent wave of a panel study (see Section 1.3.2). One caveat, however, is necessary. Because there was a fieldwork concern that the interpenetrating experiment (whose main purpose was to measure interviewer effects on substantive data – see Chapter 9) would jeopardise response rates, a strategy was employed whereby the ‘new’ interviewers were encouraged to talk to the ‘old’ interviewers to find out about hard to find addresses and particularly difficult respondents. This could artificially reduce the apparent effects of interviewer ‘dis-continuity’, but the qualitative data (described below, results in Section 8.1.4) suggest that any amelioration is more due to the interviewers’ ingenuity in handling new situations, which could be present in any study. In addition, the analyses in Section 8.2 suggest that little or no artificial reduction was present.

Qualitative Data from British Household Panel Study Interviewers. The issue of the impact of interviewer continuity in a panel study was also investigated qualitatively through the impressions of the interviewers themselves. In line with this, a series of interviewer debriefing questions were developed to look at BHPS interviewers' perceptions of respondents' reactions to the change in interviewer. It should be remembered that as a result of the interpenetrated design, each interviewer's assignment would involve them in re-visiting half of their 'old' respondents as well as visiting new respondents. They would thus be in an ideal situation to see how respondents react when faced with a familiar interviewer and when faced with a new interviewer.

BHPS staff gave us permission to include these interviewer continuity questions on a standard BHPS interviewer debriefing questionnaire which was sent to all 262 interviewers and supervisors who had worked on Wave 2 of the survey. The debriefing questionnaire achieved a 66 percent response rate.²⁹ The relevant questions from the debriefing questionnaire are shown in Figure 15.

Figure 16: Interviewer Continuity Items from British Household Panel Study Interviewer Debriefing Questionnaire

- Q19. Did you work on 'Living in Britain' (BHPS) last year?
- Q20. Were all your interviews in the same area as last year?
- Q21. Why did you work in a different area?
- Q22. Do you think the fact that you weren't the same interviewer as last year affected respondent reaction at all? If yes please explain how.
- Q23. What advice would you give to any interviewer working in a new area next year?
-

29 This low response rate is unfortunate. The BHPS staff were loath to pressurise NOP interviewers into a higher response rate and as I was no longer a member of the BHPS staff, I was unable to have a strong influence.

8.1.3 Methods

The Interviewer Continuity Indicator. We would have expected 50 percent of respondents to receive a different interviewer at Wave 2 as part of the experimental design. In actuality, 63 percent of households received a different interviewer. This is due to several factors: interviewer attrition between waves³⁰ and some interviewers therefore needing to cover more than one assignment; some interviewers moving to different areas, and supervisors coming in to finish an assignment or do refusal conversions. Thus we have 167 unique orderings of Wave 1 and Wave 2 interviewers rather than the 130 which would have been predicted by the design.

Analyses. The main focus of this chapter is on assessing the impact of interviewer continuity. This is first examined from a bivariate perspective with two-way tables (see Section 1.4.1). It is then studied after controlling for various covariates and interactions between covariates using multiple logistic regression (see Section 1.4.3). This is useful, for although in expectation the randomisation of the interpenetrated design will give each interviewer an equivalent sample of respondents, some differences may result in the actual samples drawn. The covariates used in this analysis are those developed and discussed in Chapter 5. The analysis of interviewer continuity was also facilitated through the use of cross-classified multilevel models (see Section 1.4.5), as multiple logistic regression throws away valuable information about the hierarchical nature of the design. Hierarchical analyses of variance (see Section 1.4.2) were used to explore a wide range of possible interactions.

30 Section 8.2 compares and contrasts the portion of the interpenetrated sample without this concern to the portion with this concern.

The dependent variables for this analysis are the same six variables considered in Chapters 5 and 6, i.e. total nonresponse, refusals, and non-contacts at both the household and individual levels.

8.1.4 Results

Quantitative Findings.

Interviewer Continuity Effects. In bivariate analyses (see Appendix B), multiple logistic regression analyses (see Table 5a in Chapter 5), and cross-classified multilevel analyses (see Table 6 in Chapter 5), the difference in response rate between households receiving the same interviewer at Waves 1 and 2 and those receiving a different interviewer was essentially 0, suggesting no interviewer continuity effect.

Variation in Interviewer Continuity Effects. Variation in the interviewer continuity effect was discovered when each geographic pool was considered separately. The range of the difference statistic (proportion nonresponse for different interviewer - proportion nonresponse for same interviewer) by geographic pool is shown in Table 15. For example, the difference in nonresponse rates for different versus same interviewer groups across geographic pools ranged from a minimum of -17 percent to a maximum of 19 percent (with a mean of -1) in the case of household level nonresponse, with the negative numbers indicating that the same interviewer has higher nonresponse than the different interviewer and positive numbers indicating that the same interviewer has lower nonresponse than the different interviewer.

This variation was first explored with hierarchical analyses of variance (see Section 1.4.2). In these analyses, although non-zero sums of squares for the interaction of interviewer

continuity and geographic pool were found in all cases, significant effects were only noted in the case of individual level nonresponse and individual level refusals.

Table 15: Variability of the Mean Difference in Nonresponse Rates Between the Same and Different Interviewer Groups by Geographic Pool at Wave 2

	Household Level			Individual Level		
	Non-respondents	Refusals	Non-contacts	Non-respondents	Refusals	Non-contacts
Mean difference	- 1 %	1 %	- 2 %	0 %	0 %	0 %
Minimum difference	-17 %	-15 %	- 13 %	-20 %	-14 %	- 21 %
Maximum difference	+19 %	+19 %	+14 %	+22 %	+19 %	+10 %

The significance of this variation was also considered in the context of cross-classified multilevel analyses (see Section 1.4.5). The possibility of a complex variance structure was explored by introducing a term to represent the random variation in the fixed coefficient of interviewer continuity across geographic pools. A covariance term between the variability of the coefficient of interviewer continuity and the random effects of geographic pools was also considered. As shown in Table 16 variability in the interviewer continuity coefficient was only observed in the case of individual level nonresponse, individual level refusals, and individual level non-contacts. This is similar to the findings from the hierarchical analyses of variance. However, the multilevel models suggest that the random slope terms and the covariance terms do not reach significance.

Thus although variability in the interviewer continuity effect is seen across geographic pools in the bivariate case, this variation does not prove to be important when the random effects of interviewers, PSUs, and geographic pools are controlled for.

Table 16: Random Slopes Models for Interviewer Continuity at Wave 2: Modifications to the Variance Components Models Shown in Table 8, Chapter 6

	Household Level			Individual Level		
	Non-respondents	Refusals	Non-contacts	Non-respondents	Refusals	Non-contacts
Fixed Effects						
Constant	-2.01 (0.13)	-2.63 (0.15)	-3.04 (0.18)	-1.82 (0.12)	-2.53 (0.16)	-2.82 (0.14)
Interviewer Continuity	0.02 (0.18)	-0.13 (0.24)	0.18 (0.27)	-0.06 (0.16)	-0.10 (0.25)	0.09 (0.23)
Random Effects						
Geographic Pools	0.09 (0.13)	0.0	0.0	0.13 (0.13)	0.18 (0.22)	0.05 (0.16)
Interviewer Continuity	0.0	0.0	0.0	0.25 (0.20)	0.87 (0.47)	0.51 (0.40)
Covariance of Geo Pools and Int Cont	0.0	0.0	0.0	-0.17 (0.13)	-0.52 (0.27)	-0.01 (0.18)
PSUs	0.0	0.0	0.11 (0.24)	0.01 (0.06)	0.05 (0.11)	0.24 (0.18)
Interviewers	0.23 (0.16)	0.29 (0.19)	0.26 (0.25)	0.28 (0.12)*	0.50 (0.22)*	0.00
H'holds/Individuals	1	1	1	1	1	1

* Coefficient significant at $p < .05$ using MLwiN's separate (rather than simultaneous) Wald test ($p < .01$ and $p < .001$ not shown).

Interactions with Interviewer Continuity. Using hierarchical analyses of variance, various interactions with interviewer continuity were explored.

As described in Section 8.1.1, the work of Rendtel (1990) suggests that the lack of continuity on response is most strongly felt among the oldest age group. An investigation of the BHPS data, however, showed no trace of an age by interviewer continuity interaction for any of the six nonresponse variables used as dependent variables

An interaction between interviewer continuity and interviewer grade level was found. This interaction, however, was found to be due to the disparities in workload allocation (described in Section 5.4.1), rather than to interviewer continuity effects

Also explored were the possibility of interactions with 20 household level variables, 13 individual level variables and 46 area level characteristics. Only two of the area level characteristics formed a significant interaction with the interviewer continuity variable.

As these two interactions (i.e., between interviewer continuity and the percentage of 2 person households in the area and the percentage of households with 1 car in the area) did not make theoretical sense, they could easily be chance events given the number of comparisons conducted.

This exploration of possible interactions with interviewer continuity shows little evidence to suggest that interviewer continuity is more effective in particular areas or with particular kinds of people. It should be remembered, however, that the more intangible aspects of the interviewing process (such as the doorstep conversation), have not been considered.

Qualitative Findings. The results in this section are based on a content analysis of responses to Q22 and Q23 among the 50 interviewers who had participated in the interpenetrated sample design experiment and answered the Interviewer Debriefing Questionnaire. The pattern of responses for Q22 are shown in Table 17a along with some illustrative quotes. These responses suggest that lack of interviewer continuity was not necessarily a large problem. The vast majority of interviewers felt that the change of interviewer did not have a negative effect on response, and among those who did, the impact was described as limited to only one or two cases. This is generally reassuring in terms of nonresponse error, but one does not know what effect the lack of interviewer continuity can have on response error. (This latter point is investigated in Chapter 10.)

As shown in Table 17b, interviewers provided a variety of recommendations for interviewers who find themselves in the position of being a 'new' interviewer to an 'old' respondent. The most helpful answers from Q23 were found among the 60 percent who

Table 17a: Interviewer Responses to Debriefing Q22 – Outcome of Lack of Continuity (n=50)

Negative Impact on Response

General negative comments 4%

E.g. 'Difficult to say. Possibly on the refusal in that not having interviewed them last time I was not aware of what they may have been objecting to that could have been the same with last year's interview. They may not have wanted to voice objections.'

Negative reaction in 1 or 2 cases 14%

E.g. 'One of my first year respondents did say she wouldn't be interviewed by another interviewer. She said she didn't want to give personal information to just anyone. This was an isolated case.'

Total 18%

No Negative Impact on Response, Other Reactions Noted

Interviewer discomfort, i.e. loss of rapport 8%

E.g. 'Only on a couple, I noticed on the ones I had done in previous year, the respondent seemed to have formed some sort of rapport. This was missing on the 'new' respondents.'

Noted a reaction, but had no effect on response 8%

E.g. 'Some did say they had expected the previous interviewer back but it didn't seem to have any effect.'

Respondents didn't remember interviewer 4%

E.g. 'Some couldn't remember the interviewer. Some asked if we have met.'

Positive outcomes 4%

E.g. 'I think it was a good idea especially one or two respondents with hardships. They might have felt embarrassed.'

Not a problem 58%

E.g. 'Respondents only too willing to co-operate'

Total 82%

**Table 17b: Responses to Debriefing Q23 –
Strategies to Cope with Lack of Continuity (n=50)**

General interview strategies	25%
Necessity of gathering knowledge from the previous interviewer	
<i>E.g. Specific details on difficult households</i>	15%
Necessity of acknowledging the change	45%
<i>E.g. Mentioning the name of the other interviewer Passing along regards from the previous interviewer Having a good explanation as to why the previous interviewer was not there.</i>	
Other answers	15%
Total	100%

provided specific suggestions. These answers suggest some useful field strategies to be used whenever the situation of ‘new’ interviewers to ‘old’ respondents can not be avoided.

As a minimum interviewers, for example, can be:

- given the name of the previous interviewer
- instructed to mention the name
- instructed to give regards from the previous interviewer
- explain why the previous interviewer is not there.

Mentioning the name of the previous interviewer lends legitimacy to the role of the new interviewer. Passing along regards from the previous interviewer helps to make the

respondent feel special. Explaining why the previous interviewer was unable to come follows normal social conventions.

In addition to these steps, it is highly advisable to have the new interviewer chat with the old interviewer about the assignment so that the experience gained by the first interviewer in terms of finding addresses and handling particular respondents is not wasted.

These various examples can also be seen as examples of good interviewer tailoring. As Groves, Cialdini, and Couper (1992) suggest, expert interviewers have access to a large repertoire of cues, phrases, or descriptors corresponding to the survey request. Which statement they use to begin the conversation is the result of observations about the housing unit, the neighbourhood, and immediate reactions upon first contact with the person who answers the door. Knowing in advance that they would be contacting different respondents, the interviewers would have been able to plan advance strategies to ameliorate this problem.

8.1.5 End of Section Summary

As the interpenetrated sample design was employed in Wave 2 of the BHPS, it offered a rare opportunity to test experimentally the extent to which sending the same or different interviewer to households in later waves of a panel study makes a difference on nonresponse.

Across all of the different types of analyses (bivariate, multiple logistic regression, and cross-classified multilevel) there was no observable effect of interviewer continuity on nonresponse at Wave 2.

Although in bivariate analyses there was considerable variation in the effectiveness of an interviewer continuity strategy across geographic pools, hierarchical analyses of variance and cross-classified multilevel analyses in which the random effects of interviewers, PSUs and geographic pools were controlled for suggested that a small amount of variation was only present in the case of individual level nonresponse, refusals and non-contacts. Furthermore, this variability could not be explained by the measurable characteristics of individuals, households or areas. The same is true with respect to the background characteristics and experience of the interviewer.

The qualitative findings suggest that BHPS interviewers have a variety of resourceful and creative strategies at their disposal to help combat any possible interviewer continuity problems. Their suggestions offer good strategies which can be used by other survey organisations to ameliorate a possible interviewer dis-continuity effect.

8.2 Refining and Extending the Analysis of Interviewer Continuity

8.2.1 Background

As described in Section 8.1, the interpenetrated sub-sample experiment from Wave 2 of the BHPS, showed no effect of interviewer continuity. Independently, Laurie, Smith and Scott (1997), however, did find an interviewer continuity effect for the full BHPS sample, when looking across Waves 2 through 4. Of interest is whether the effects found by Laurie and her colleagues are due to their multiple wave perspective or due to the potential non-random nature of interviewer attrition present in the full sample. This Section explores this issue by comparing data from Waves 2 and 3 of the interpenetrated sample design experiment with data from the full sample.

Although we found no overall interviewer continuity effect at Wave 2, a small amount of variation in the effectiveness of interviewer continuity was found across geographic pools for individual level nonresponse, refusals, and non-contacts. The possibility of random variation in the interviewer continuity coefficient across geographic pools will be revisited with Wave 3 data in the context of cross-classified multilevel models.

8.2.2 *The Data*

The interviewer continuity analyses in Section 8.1 are based on the full interpenetrated sample design. The analyses in this Section make a distinction between those geographic pools in which no interviewer attrition occurred and those with potentially non-random interviewer attrition.

For Wave 2, as described in Section 1.3.2, the interpenetrated design creates a hierarchical cross-classification in which households or individuals are seen to be nested within the cross-classification of interviewer by PSU, which are in turn nested within larger geographic pools. Within 14 geographic pools where no accidental interviewer attrition occurred between Wave 1 and Wave 2, the main diagonal formed by the cross-classification of interviewer by PSU represents households/individuals receiving the same interviewer at Wave 2 as at Wave 1 and the off-diagonals represent households/individuals receiving a different interviewer at Wave 2. Within the 16 geographic pools with interviewer attrition, this design does not hold.

This Section also makes use of data from Wave 3 of the interpenetrated sample. For Wave 3 (as described in Section 1.3.3) the random assignment of interviewers was exactly reversed. Through the reversing of the random assignment Wave 2 to Wave 3, the measurement of interviewer continuity is possible in the 24 geographic pools with no

accidental interviewer attrition between Waves 2 and 3. For these geographic pools, the data can be viewed as a hierarchical cross-classification in which households or individuals are seen to be nested within the cross-classification of interviewer/PSU by interviewer continuity, which are in turn nested within geographic pool. This design does not hold for the 6 geographic pools in which interviewer attrition occurred.

Those geographic pools with interviewer attrition can be used to suggest what effects might be expected from non-random interviewer attrition in the full sample. This thesis compares the interviewer attrition portions of the interpenetrated sub-sample for both Wave 2 and Wave 3 with data from the full BHPS sample.

8.2.3 Methods

Replication of Laurie et al, 1997. As described in Section 8.2.1, the focus of this half of the Chapter is on a replication of Laurie *et al* (1997). Thus data management decisions have been made with this in mind. For example, in line with Laurie and her colleagues we (1) defined interviewer continuity as having the same interviewer at the immediately preceding wave, (2) restricted the analyses to original Wave 1 respondents, excluding ‘new entrants’ (i.e., new comers to original households, other people in households where the original sample member is now the new entrant, and so on) and (3) restricted the analyses to non-movers. It can be argued that new entrants and movers would have a different experience of interviewer continuity.

There are inherent problems in interviewer continuity research. The determination of interviewer continuity is made through a comparison of interviewer ID numbers across time. Interviewers unfortunately can be rather poor about writing their ID numbers on coversheets, especially for nonresponse cases! These missing interviewer ID numbers

have to be reconstructed based on interviewer assignments. This will generally be accurate, except for any re-issues to other interviewers which might occur. Personal communication with Laurie and her colleagues suggested that we were using similar strategies in trying to construct the missing IDs. We also searched for and corrected several keying errors in interviewers' ID numbers.

More problematic are the telephone conversions which took place on the BHPS, particularly from Wave 3 onwards. Even though after the persuasive call the case was returned to the same interviewer, these telephone conversion cases have been excluded as they muddy the issue of interviewer continuity. But this means that the nonresponse figures reported in Laurie *et al* (1997) and in this thesis, underestimate the true amount of nonresponse. Similarly, standard refusal conversions have been eliminated together with a small handful of cases from Wave 3 onwards which were simply not reissued because they were strong nonresponse cases. Because these latter cases have no interviewer involvement at the current wave they have been excluded from analysis, but this again means that the nonresponse level reported here underestimates the true amount. On a more minor note, respondents that have moved out of scope (to another country or to an institution) or have died are treated as ineligible by BHPS staff, but some surveys would classify these non-interview cases as nonresponse (see Groves, 1989).

Finally it should be noted that the interviewer continuity analyses are sensitive, rather than robust, to the decisions made above. It should be remembered, however, that the key point is whether, keeping all the data decisions constant across analyses, the same conclusions about interviewer continuity will be reached in the experimental sub-sample as in the full sample.

The Analyses. As described in Section 1.4.1, chi-square tests of independence were used in two-way tables and loglinear modelling was used in three-way tables. The specific focus of the three-way tables was to search for evidence for the presence of non-random interviewer attrition. Variability in the effectiveness of interviewer continuity across geographic pools at Wave 3 was assessed through cross-classified multilevel logistic regression models in which the coefficient for interviewer continuity was allowed to vary randomly.

8.2.4 Results

Fixed Effects

Differences in Interviewer Continuity Effects. Table 18 shows a comparison of several different ways of viewing the interviewer continuity effect. Looking at the panel for Wave 2, for example, it can be seen that there is no effect of interviewer continuity at Wave 2. This is true for the portion of the interpenetrated sub-sample (IPS) without attrition in which the hypothesis can be experimentally tested (Row A) as well as in the interpenetrated sample with attrition (Row B), the full sample (Row C), and in the analyses of Laurie and her colleagues (1997) (Row D). The key question, however, is whether or not Wave 2 is the start of growing interviewer continuity effect over subsequent waves. Although not significant, Rows A and B are in the opposite direction from an expected interviewer continuity effect, with lower rather than higher response rates for households and individuals who received the same interviewer at both Waves 1 and 2. The reverse pattern is seen in both our full sample analysis (Row C) and that of Laurie and her colleagues (Row D).

Table 18: The Impact of Interviewer Continuity Across Waves of the British Household Panel Study for Original Wave 1 Respondents Who Have Not Moved Out of Their Original Areas †

		Households			Individuals				
		Co-operating	Refusal	Non-contact	Full interviews	Proxy interviews	Refusals	Non-contacts	Other NR
WAVE 2									
A. IPS no attrition	same	88.1	6.6	5.3	87.7	0.7	7.0	3.9	0.7
	diff	88.4	6.8	4.8	88.5	1.0	6.6	3.8	0.0
		n=654			n=1142				
B. IPS non-random attrition	same	88.4	5.1	6.6	85.1	1.7	6.4	6.1	0.6
	diff	89.6	6.2	4.2	86.4	2.6	7.3	3.5	0.3
		n=746			n=1267				
C. Full sample	same				89.2	1.4	7.6	1.3	0.5
	diff				88.6	1.7	7.6	1.7	0.4
					n=9146				
D. Laurie <i>et al.</i> , 1997	same				89.3	NA	NA	NA	NA
	diff				88.2	NA	NA	NA	NA
					n=NA				
WAVE 3									
A. IPS no attrition	same	90.7	6.0	3.4	87.0	1.1	9.2	2.0	0.7
	diff	93.3	4.9	1.9	88.2	2.0	7.9	1.3	0.6
		n=1039			n=1742				
B. IPS non-random attrition	same	95.7	2.1	2.1	*	83.3	1.1	5.6	1.1
	diff	90.7	7.0	2.3	*	82.3	3.4	11.9	1.4
		n=219			n=384				
C. Full sample	same				*	88.6	1.3	8.0	1.0
	diff				*	87.5	2.2	8.2	1.3
					n=8189				
D. Laurie <i>et al.</i> , 1997	same				*	86.1	NA	NA	NA
	diff				*	83.3	NA	NA	NA
					n=NA				
WAVE 4									
C. Full sample	same				*	90.4	1.5	6.3	1.3
	diff				*	87.6	1.3	8.0	2.6
					n=7560				
D. Laurie <i>et al.</i> , 1997	same				*	86.1	NA	NA	NA
	diff				*	83.3	NA	NA	NA
					n=NA				

* p < .05

† Note that interviewer continuity is defined as having the same interviewer as the previous wave. Other non-interview households are excluded at Wave 2 and Wave 3 from household tables. Telephone interviews are excluded from all waves for the individual and household tables. Rows A and B exclude refusal conversion cases by supervisors. Also excluded are firm nonresponse cases which were not re-issued to interviewers at a particular wave as well as initially co-operating households that become ineligible and cases where missing interviewer ID numbers could not be imputed.

A more interesting picture unfolds at Wave 3. Wave 3 clearly shows that different conclusions would be reached depending on whether one examines Row A or Row B, with Row A suggesting lower response rates among same interviewer combinations and Row B suggesting the reverse. In a loglinear model, the three-way interaction (nonresponse by Row A versus B by same versus different interviewer) approaches significance (i.e., the attained level of significance is 0.06 for the change of 8.98 in the likelihood ratio χ^2 between a saturated model and one with all two-way interactions, the change in the degrees of freedom was 4). This was not true in the case of household nonresponse. In Wave 3, the pattern detected in Row B is mirrored in Row C and D as well, all three of which can suffer from non-random interviewer attrition and all three of which now reach significance at the individual level.

Interviewer continuity can no longer be examined experimentally in Wave 4. If one could assume, however, that the pattern of Row A and B in Wave 3 would continue into Wave 4, this would suggest that the effect found by Laurie and her colleagues (1997) is more due to the potential non-random nature of interviewer attrition than purely to a lack of interviewer continuity.

From Wave 3 onwards, many patterns of interviewer continuity are possible. Although not shown here, we also explored the trichotomy: all the same interviewer, all different interviewers, and at least two interviews with the same interviewer. The group of respondents receiving 'all different interviewers' had significantly lower response rates than the other two groups for the full sample at Waves 3 and 4. For Wave 3, there were similar significant differences noted for the portion of the interpenetrated sub-sample which experienced interviewer attrition. Importantly, however, there were no significant

differences within the interpenetrated sub-sample without attrition. Again, this suggests that non-random interviewer attrition may be the cause of the results.

Loglinear Investigation into the Randomness of Interviewer Attrition. The possibility of non-random interviewer attrition was explored more directly through the use of loglinear analyses to explore the relationship between interviewer continuity, individual level nonresponse and region. Of concern here is whether the fact that respondents received the same or a different interviewer is related to region and whether this in turn is related to nonresponse. Thus, ideally we would want to show evidence for a three-way interaction between region, nonresponse, and interviewer continuity in the portion of the interpenetrated sample with non-random interviewer attrition and in the portion of the full sample excluding the interpenetrated sub-sample cases. In contrast, we would only expect to find a relationship between region and nonresponse for the no interviewer attrition portion of the interpenetrated sub-sample. Results with respect to these three sub-samples are shown respectively in Rows B, C', and A in Table 19. Although not a perfect match to expectations, there is definite evidence to suggest that the presence of non-random interviewer attrition could be affecting results. For example, at Wave 2 a significant 3-way interaction was found for the portion of the full sample excluding the interpenetrated sub-sample cases (Row C') (change χ^2 if three-way interaction is deleted = 30.01, df = 17, p < .05). For Row C' at Wave 3, the best fitting model included the interaction of nonresponse by region (change χ^2 = 33.44, df = 17, p < .01) and interviewer continuity by region (change χ^2 = 626.56, df = 17, p < .001). For Row C' at Wave 4, a significant 3-way interaction was again found (change χ^2 = 29.14, df = 17, p < .05). In contrast, the Row A results only show the effects of nonresponse and region, as expected.

Table 19: Exploring the Possibility of Non-Random Interviewer Attrition with Loglinear Modelling: Three-Way Tables - Interviewer Continuity by Individual Level Nonresponse by Region †

Sample	Sample Size	Number of Regions	Generating Class of Best Fitting, Parsimonious Loglinear Model (Lower-order terms are present in model, but not shown in table)	χ^2 of Final Model	Change χ^2 if Term Deleted	% of Empty Cells	% Cells with Expected Counts < 5
WAVE 2							
A. IPS no attrition	n=1142 (14 geo pools)	10 Regions	NONRESPONSE X REGION	28.38, df = 20, p = .101	27.69, df = 9, p < .01	3%	10%
B. IPS non-random attrition	n=1267 (16 geo pools)	10 Regions	NONRESPONSE X REGION	11.60, df = 10, p = .312	27.26, df = 9, p < .01	3%	15% ‡
			REGION X INT CONTINUITY		20.86, df = 9, p < .05		
C'. Full sample (excluding A and B portions)	n=6722	18 Regions	NONRESPONSE X REGION X INT CONT	.0000, df = 0, p = 1.0 (saturated)	30.01, df = 17, p < .05	0%	6%
WAVE 3							
A. IPS no attrition	n=1742 (24 geo pools)	15 Regions	NONRESPONSE	51.54, df = 44, p = .203	1111.90, df = 1, p < .001	2%	7%
			REGION		335.56, df=14, p < .001		
B. IPS non-random attrition	n=384 (6 geo pools)	4 Regions	NONRESPONSE X REGION	2.56, df = 4, p = .635	27.80, df = 3, p < .001	13%	19% ‡
			REGION X INT CONTINUITY		28.56, df = 3, p < .001		
C'. Full sample (excluding A and B portions)	n=6055	18 Regions	NONRESPONSE X REGION	15.42, df = 18, p = .633	33.44, df = 17, p < .01	10%	24% ‡
			REGION X INT CONTINUITY		626.56, df = 17, p < .001		
WAVE 4							
C'. Full sample (excluding A and B portions)	n=5610	18 Regions	NONRESPONSE X REGION X INT CONT	.0000, df = 0, p = 1.0 (saturated)	29.14, df = 17, p < .05	6%	15% ‡

† Throughout the analysis the 2 categories, same and different interviewer, were used for interviewer continuity, the 2 categories, full interview and nonresponse, were used for nonresponse, and the original 18 BHPS categories were used for region. When fewer regions appear this is because only a portion of the geographic pools were included.

‡ Given the large proportions of cells with expected counts less than 5, these models were re-run with a collapsed version of region. The same models was chosen in 3 of the 4 cases, the exception being that a model with just all two-way interactions was selected for Wave 4.

Note that the original 19 category region variable from the BHPS data is used throughout Table 19 to promote comparability. The number of regions in any particular analysis, however, varies because different geographic pools are included in the different analyses, depending on the amount of interviewer attrition between waves (see Section 8.2.2). Given that the models for the non-random attrition portion of the interpenetrated sample (Row B) for Waves 2 and 3 and for the full sample, excluding the interpenetrated sample (Row C') for Waves 3 and 4 have a high proportion of cells with expected counts less than 5, additional analyses were conducted with a collapsed version of region with 8 categories. This was created by combining adjacent rural areas and combining adjacent urban areas. The analyses based on the collapsed version of region yielded the same conclusions about the generating class for the best fitting model in three of the four cases. The exception was the model for Wave 4. A model with just all two-way interactions was selected over one with the addition of the three-way interaction.

Revisiting the Attenuation Issue. The findings from Table 18 also help to reduce the concern described in Section 8.1.2 that the BHPS strategy of having interviewers talk to each other before swapping assignments may have attenuated the interviewer continuity affect. For the portion of the interpenetrated sub-sample with non-random attrition, interviewers would not have been able to have such a chat because their counterpart interviewers would have already left employment. Yet at Wave 2, we can see that they clearly achieved the same pattern of response results as the interviewers for the no attrition sub-sample. In addition, there is little difference between these results and those for the full sample where no random swapping took place.

Random Slopes for Interviewer Continuity. The variable effectiveness of interviewer continuity from Wave 1 to Wave 2 was examined in Section 8.1. This section looks for evidence for random variation in the interviewer continuity coefficient across geographic pools at Wave 3. Here random variation of the interviewer continuity coefficient can be estimated as part of the cross-classification (see Section 8.2.2). The results from this work are shown in Table 20. Although not shown here, very similar results were found with a simple nested model (with individuals/households within interviewers/PSUs within geographic pools) in which the coefficient of interviewer continuity was allowed to vary across geographic pools and a covariance term between geographic pools and interviewer continuity was allowed. As can be seen in Table 20, a non-zero value for the variability in the interviewer continuity coefficient is only found in the case of household level non-contacts and individual level refusals, but neither of these reach significance.

Table 20: Random Slopes Models for Interviewer Continuity at Wave 3

	Households (n = 1039)		Individuals (n = 1731)	
	Model 29 Refusals	Model 30 † Non-contacts	Model 31 Refusals	Model 32 Non-contacts
Fixed Effects				
Constant	-3.11 (0.24)	-3.92 (0.33)	-2.54 (0.16)	-3.44 (0.22)
Same Interviewer W2 to W3	0.24 (0.29)	0.62 (0.41)	0.14 (0.20)	-0.06 (0.29)
Random Effects				
Geographic pools	0.16 (0.27)	0.09 (0.41)	0.0	0.01 (0.26)
PSUs/Interviewers	0.10 (0.32)	0.0	0.29 (0.16)	0.25 (0.36)
Interviewer Continuity	0.0	0.07 (0.54)	0.06 (0.12)	0.0
Households/Individuals	1	1	1	1

† 2nd order estimation would not converge for this household level model. RIGLIS and PQL with 1st order estimation were used.

* Coefficient significant at $p < .05$ using MLwiN's separate (rather than simultaneous) Wald test ($p < .01$ and $p < .001$ not shown).

8.2.5 Chapter Summary and Discussion

In Section 8.1 we had found no observable interviewer continuity effect on response rates between Waves 1 and 2 of the BHPS using data from the interpenetrated sub-sample

experiment. In contrast Laurie, Smith and Scott (1997) had found a significant interviewer continuity effect when using the full BHPS data across Waves 2 through 4. Using Wave 2 and Wave 3 data, Section 8.2 investigated whether the difference in findings might be explained by the presence of non-random interviewer attrition in the full sample or by the more longitudinal perspective examined by Laurie and her colleagues. Distinctions were also made between the portions of the interpenetrated sample with and without interviewer attrition. Although the findings cannot be conclusive, patterns in the data suggested that the non-experimental nature of the full sample is the major source of the findings of Laurie and her colleagues.

Having rejected the hypothesis of an interviewer continuity benefit in the aggregate, we used cross-classified multilevel modelling to explore variation in the effectiveness of same versus different interviewers combinations across geographic pools. As it was, although random variation in the interviewer continuity coefficient was found at both Wave 2 (see Section 8.1.4) and Wave 3 (see Section 8.2.4), it did not reach significance.

The fact that no net interviewer continuity effect was found in this thesis is somewhat reassuring. Attrition of the interviewing workforce is a perennial problem in most survey organisations. There was a concern in Section 8.1.2, that the lack of an interviewer continuity effect may have been partly due to the fact that the 'new' interviewers were encouraged to talk to the 'old' interviewers about possible hard-to-find and problem respondents prior to fieldwork. The qualitative data from the interviewers themselves (discussed in Section 8.1.4) suggests that if BHPS interviewers actually practised the resourceful and creative strategies they described, this may account for the lack of quantitative findings. In addition, however, the data from the loglinear modelling

(discussed in Section 8.2.4) clearly show that this conversation strategy had little or no effect on the resulting interviewer continuity effect.

Despite these conclusions about the fixed and random aspects of interviewer continuity, we feel it would be unwise for field organisations simply to dismiss concerns about interviewer continuity. There were many obstacles to the exact estimation of interviewer continuity effects in Section 8.2 and the work of Laurie and her colleagues (1997). In addition, both studies have focused on a small slice of the issue. For example, only respondents with full Wave 1 interviews who had not moved out of their local area were considered. Also excluded were cases which needed telephone conversion, supervisor conversion, were found to be so unco-operative that they had been dropped before re-issuing to the next wave of the panel study or who had become ineligible at either the current or previous waves. In the future the impact of the excluded nonresponse cases can be examined.

Although the interviewer continuity results of this study should generalise to other types of longitudinal surveys as well as other classic panel study designs, such generalisation needs to be considered in light of a given interviewing staff and their level of resourcefulness. In addition, survey organisations should keep in mind that even if interviewer continuity does not affect response rates directly, it could have an impact on interviewers in terms of their own feelings of vulnerability and morale. It may also have an impact on response quality (see Chapter 10). These issues need to be explored as well as the extent to which respondents expect the return of the same interviewer.

PART 3 EMPIRICAL INVESTIGATIONS INTO INTERVIEWERS AND RESPONSE ERROR

CHAPTER 9 WHAT TYPES OF RESPONSE VARIANCE DO INTERVIEWERS MANIFEST? ³¹

9.1 Background

As described in Chapter 2, the interviewer is seen as one of the major sources of error in data collected from structured face-to-face interviews. The other major component of imprecision in survey estimates is sampling variance. It is known that for most complex sample survey designs the precision of estimators is low compared to simple random sample designs of the same size. Area clusters typically form the sampling units for complex sample designs and the loss of precision is due to positive correlations among people belonging to the same area clusters.

Though there are some studies in which the complex sampling variance and the complex interviewer variance are both computed (see, Bailey, Moore and Bailar, 1978, for the US National Crime Survey; Collins and Butcher, 1982, for a consumer attitude survey; and O'Muircheartaigh, 1984a, 1984b, for the World Fertility Survey in Lesotho and Peru), such studies are rare. This is due to a combination of design and analytic challenges. The norm for face-to-face interview surveys in both the US and UK is to have the workload from a given Primary Sampling Unit (PSU) assigned to a single interviewer and, moreover, to have each interviewer work in only one PSU. This confounds the sampling and interviewer variances and thus standard estimates of sampling variance often include both (see Collins and Butcher, 1982). Such confounding is removed by an interpenetrated design in which respondents are assigned at random to interviewers. Due to cost considerations, these designs are rarely

31 The material from this Chapter was first presented in May of 1995 at the annual meeting of the American Association for Public Opinion Research, Fort Lauderdale, Florida. It has been recently published as O'Muircheartaigh and Campanelli (1998).

employed in face-to-face surveys. Even for telephone surveys, where the practical problems are less severe, though non-trivial (see Groves and Magilavy, 1986), such studies are uncommon.

The *design effect* is the most commonly used measure of the impact of within PSU homogeneity on survey results; this is

$$\text{deff} = 1 + \rho_s(b-1) \quad (47)$$

where s = the sample clustering,

ρ_s = the intra-cluster correlation, and

b = the average number of elements selected from a cluster.

As described in Section 1.2.3, the analogous expression for interviewers is the *interviewer effect*, this is

$$\text{inteff} = 1 + \rho_i(k-1) \quad (48)$$

where i = the interviewer,

ρ_i = the intra-interviewer correlation, and

k = the average interviewer workload.

Both ρ_s and ρ_i measure the within-unit (PSU or interviewer) homogeneity of the observations. Within-PSU homogeneity is a characteristic of the true values of the elements in the population. Within interviewer workloads the homogeneity results from the interaction between the interviewer and his/her respondents. The number of elements selected from a cluster and the interviewer workload size arise as a result of decisions by the designer of the survey; thus ρ_s and ρ_i are quantities intrinsic to the population structure and to the quality of interviewers. As such the latter are more portable than the variance components themselves;

the variance components themselves can of course be calculated once the ρ values are known.³²

During the past thirty years or so evidence has accumulated about the order of magnitude of both the intra-cluster correlation coefficient and the intra-interviewer correlation coefficient in sample surveys in the US and elsewhere. The findings with respect to ρ_i are described in Section 1.2.5.

The range of values reported in the literature for ρ_s is similar to that for ρ_i , though one would expect ρ_i to have more values near zero. Again, the evidence suggests that values greater than 0.1 are uncommon and that positive values are almost universal. The large values tend to be for certain types of demographic variables, notably tenure and ethnic origin. This is to be expected since adjacent groups of houses in a small area will tend to be of similar type and tenure (Lynn and Lievesley, 1991). Other demographic variables such as sex and marital status tend to show very low values. It is typically found that behavioural and attitudinal variables have ρ_s values somewhere between these extremes, with attitudinal variables showing slightly lower values than behavioural ones. In the World Fertility Survey (see Verma, Scott, and O'Muircheartaigh, 1980), the median ρ_s across various countries was 0.02 for various nuptiality and fertility variables. The median was much higher (around 0.08) for variables concerning contraceptive knowledge.

In comparing these two sources of variability, Hansen, Hurwitz and Bershad (1961) found that interviewer variance was often larger than the sampling variance. Bailey, Moore, and Bailar (1978), on the other hand, found response variance components that were 50 percent of their

32 Note that technically, ρ_s and ρ_i also reflect any increase or decrease in homogeneity which might occur due to non-random nonresponse.

sampling variance for only a quarter of their statistics. Collins and Butcher (1982) found that values of the intra-interviewer correlation coefficient were of a similar scale to those of the intra-cluster correlation coefficient. However, the interviewer effects tended to be larger for the attitudinal items and the sample design effects larger for the factual items.

9.2 The Data and Methods

This Chapter compares the relative impact of interviewer effects and sample design effects on survey precision by making use of the interpenetrated PSU/interviewer experiment in the second wave of the BHPS (see Section 1.3 for a full description of the data and the experiment).

The main focus is on the estimation of the intra-class correlation coefficients for PSUs and interviewers, ρ_s and ρ_i , respectively. There are, of course, other features which add to the complexity of multistage survey designs such as stratification and the use of sampling with probabilities proportional to size. However, in a national sample the available stratifiers (primarily region and measures of urbanisation) produce quite modest gains in precision, typically less than 5 percent. The complexity of adding these to the analyses did not justify their inclusion. Ignoring them may inflate the estimates of ρ_s but only if the effect is greater on the estimate of between cluster variance than on the estimate of the total variance. The survey literature refers to an estimate of ρ that may include the impact of stratification as a 'synthetic' ρ (see, for example, Kish, 1965).

Selection with probability proportional to size followed by selection with probabilities inversely proportional to size within clusters will not affect the results though it does affect slightly the interpretation of ρ : ρ is now the estimated intra-cluster correlation coefficient for the

subclusters created by the subsampling procedure. As the subsampling procedure for BHPS was essentially random within cluster, this distinction is unimportant here.

First, estimates of the intra-cluster correlation coefficient (ρ_s) and the intra-interviewer correlation coefficient (ρ_i) were calculated. Patterns in these values were examined. Then multilevel analyses were conducted the results of which were compared to a standard single level analyses. In both cases, appropriate individual and household level variables were entered as covariates. In addition, the characteristics of interviewers were entered to see what extent these offered an explanation for the interviewer effects.

9.2.1 Estimation of ρ

Using hierarchical analyses of variances through the SPSS MANOVA procedure, intraclass correlation coefficients were calculated for the effects of the interviewers (ρ_i) and for the effects of the PSUs (ρ_s). (See Section 1.4.2 for a full discussion of the advantages and disadvantages of the hierarchical analysis of variance approach). These coefficients were estimated for all variables in the dataset. This Chapter, however, only reports on variables with at least 700 or more cases to ensure the stability of the estimates. This decision was based on a rough rule of thumb that n should be greater than or equal to the degrees of freedom times 10. Categorical and most ordinal variables were transformed into binary variables prior to the analyses; ordinal attitude scales (Likert scales) were, however, treated as continuous assuming latent continuous distributions. This resulted in 820 variables and categories to be analysed.

The sums of squares were partitioned using a ‘regression approach’ in which each term is corrected for every other term in the model. This makes sense substantively and also facilitates the planned comparison with the multilevel models. It also means that the values for ρ_i and ρ_s

which are reported are conditional on each other. Data from the hierarchical analysis of variance runs were then assembled to create a meta dataset of ρ estimates. Other information was added to this dataset. For example, a distinction was made by question type. Overall there were 98 attitude questions, 574 factual questions, 88 interviewer check items (i.e., items completed by the interviewers without a formal question), and 60 'quasi-facts' (mostly from a self-completion form). Also added was information about the topic area of the questionnaire, and whether the variable was continuous or categorical. There were not enough open-ended questions in the BHPS questionnaire to allow for an analytic comparison of interviewer effects for open and closed questions as described in Section 2.3.2.

9.2.2 *Multilevel Models*

Whereas it has been possible to carry out a simultaneous analysis of interviewer and cluster effects for sample means and other simple statistics³³, it is only recently that software has become available to estimate interviewer and cluster effects simultaneously while incorporating these effects directly into a substantive model of interest. This is possible through the use of a cross-classified multilevel modelling (see Sections 1.4.5-1.4.8). The analyses in this Chapter were conducted using the software package ML3 (see discussion in Section 1.4.9) as this was the version of the software available at this time.

9.3 Results

9.3.1 *Findings from Hierarchical Analyses of Variance*

Figure 17 shows the cumulative frequency distributions for ρ_s and ρ_i . The orders of magnitude for the two coefficients were strikingly similar. (As these values are themselves estimates, they are subject to imprecision and can take on negative values. Some authors set these negative

33 Technically, means and proportions estimated from survey data are ratio estimates as there is uncontrolled variation in the sample size. For the BHPS with the selection of PSUs with probability proportional to size and equal probabilities overall, this variation is fairly tightly

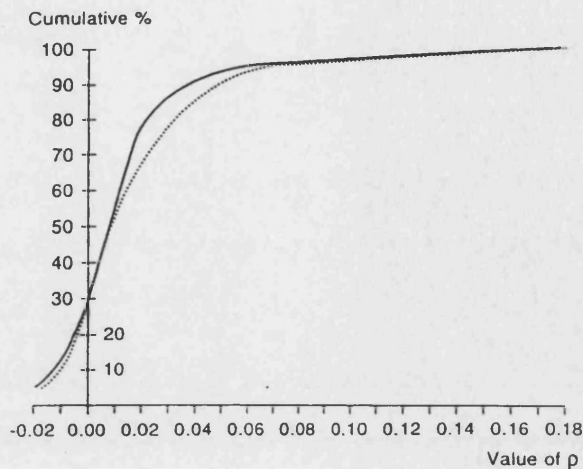
values to zero (see Table 2), but we preferred to leave the actual values.) To gain an approximate idea of the significance of the ρ values, the standard F test from the hierarchical analysis of variance was employed.³⁴ Using a 5 percent level of significance, 4 in 10 of the values of ρ_s and 3 in 10 of the values of ρ_t were significantly greater than zero. In the case of ρ_s this is not surprising as positive values are expected for most survey variables. What is somewhat surprising is that, within the study, ρ_t is of the same order of magnitude. For these data, because of the way the investigation was designed, the average interviewer workload and the average number of elements selected from a cluster were the same; thus the estimates of ρ_s and ρ_t imply that the *effect* of the sample design and the interviewers were also about the same.

All types of questions show the presence of significant values of ρ_t . For the attitude questions, 28 percent of the values of ρ_t were significantly greater than zero; for the factual questions it was 26 percent; for the interviewer check items, a staggering 58 percent; and for the quasi-factual questions, 25 percent (with the exclusion of the self-completion items). What is interesting is the similarity of the findings for the attitudinal and factual items, which is in contrast to the findings of some studies. As discussed in Section 2.3.1, some research has suggested that attitudinal items were more susceptible to interviewer effects than factual items. There was, however, some variation between types of attitudinal item in this study. Among the 'classic' attitude items measured on Likert scales, 33 percent showed significant values of ρ_t ; this compares with 25 percent of the other attitude items.

controlled.

34 The results of these tests should be treated with some wariness. As described in Section 9.2.1, several of the dependent variables are binary, although mostly in the 20-80 percent range (see also Section 1.4.2). In addition, given the varying size cells of the cross-classification it is possible that homogeneity of variance does not hold. In the presence of heteroscedasticity the error variance may be inflated, thereby reducing the size of the F statistics. This latter point suggests that the results are conservative.

Figure 17: Intra-interviewer and Intra-cluster Correlations: Cumulative Distribution of ρ (—) and ρ (. . . .) □



□ The negative values of ρ and ρ occur because ρ and ρ are estimates.

We also looked for differences by source of the question. For example, 32 percent of the items in the individual schedule had ρ values which were significantly greater than zero. The same was true for 17 percent of the self-completion items, 27 percent of the coversheet items, 28 percent of the derived variables from the individual's questionnaire, 32 percent of the household questionnaire items, and 34 percent of the derived variables from the household questionnaire. The notable difference here in susceptibility to interviewer effects is between the self-completion items and those that are interviewer-administered. The fact that there is an interviewer effect at all on the self-completion form is interesting, given the much lower level of interviewer involvement. As discussed in Section 2.3.3, Kish (1962), found little evidence to suggest the presence of such an effect on the written questionnaires he examined. In contrast, O'Muircheartaigh and Wiggins (1981), however, did find an effect for a health supplement completed in the presence of the interviewer (as were the BHPS self-completion items).

There was also basically no difference in the proportion of significant ρ_1 values between the different sections of the questionnaire: demographics, health, marriage and fertility, employment, employment history, values, and income and household allocation (with the percentage significant ranging from 22 percent to 35 percent). In contrast the section at the end of the questionnaire for interviewers to record their observations was highly susceptible to interviewer effects. Seventy-six percent of the items in the interviewer observation section showed significant values of ρ_1 . There was also a difference between dummy and continuous variables, with a higher proportion of effects being noted for the continuous variables.

Furthermore, there was a clear positive correlation of 0.35 between ρ_1 and ρ_2 . A positive correlation between ρ_2 and ρ_1 in this meta dataset implies that variables that show large intra-cluster homogeneity (i.e., show relatively substantial clustering among true values) are also sensitive to differential effects from interviewers. Such a correlation has not, to our knowledge, been observed before. As the elements in the computation of this correlation are themselves variables, the absence of such evidence may be because it is necessary to have a large number of variables to estimate such a correlation coefficient with any precision. In this analysis, the correlation shows remarkable consistency across types of variables.

An explanation for this may be as follows: Homogeneous clusters contain individuals similar to one another; it is not unreasonable to suggest that individuals with *similar* values on the variable in question may respond in a *similar* way to whatever qualities the interviewer brings to bear in the interviewer-respondent interaction. This would mean that variables that manifested intra-cluster homogeneity would on balance be more likely than other variables to be display intra-interviewer homogeneity. An alternative explanation may be found from the perspective of interviewer expectations. As discussed in Section 2.2.2, interviewer

expectations are known to influence the responses obtained by interviewers. For a variable to have a relatively large value of ρ_3 , the individuals within a cluster will have relatively homogeneous values; it is possible that this consistency will affect the interviewers' expectations as the interviewer's workload progresses, leading to enhanced correlations within interviewer workloads.

This latter explanation is consistent with the technical interpretation of the correlation between the response deviation and the sampling deviation for a single variable postulated in the Census Bureau model and included in Hansen, Hurwitz, and Bershada (1961), Fellegi (1964), and Bailey, Moore, and Bailar (1978). It is not possible to estimate this correlation directly for a single variable without at least two waves of data collection, though it is included in the model estimate of ρ_1 (see Section 1.2.4). Hansen, Hurwitz, and Bershada (1961) give an example of how this latter correlation may arise for a single variable.³⁵

9.3.2 Findings from the Multilevel Models

For illustration, three cross-classified multilevel logistic regression models are included, one for each of the main types of variables: interviewer check items, facts, and attitudes. These are shown in Tables 21 through 23, respectively. The corresponding non-hierarchical (single level) model are also shown for the purpose of comparison. Of interest is whether any substantive conclusions will be affected when the appropriate data structure is incorporated in the analysis.

35 There was some concern that this correlation between ρ_3 and ρ_1 might be an artefact from the method of estimation used as the sums of squares in the hierarchical analyses of variance were calculated with a regression approach (see Section 1.4.2). However, our understanding of the regression approach is that the sums of squares for each component reflects its unique contribution, while holding the other components constant and therefore this method should not be problematic.

The variable modelled in Table 21 is a binary subcategory indicating whether children were present during the demographics section of the interview, as noted by the interviewer. From the hierarchical analyses of variance, the estimated ρ values for this *children present* subcategory were $\rho_1 = 0.171$ and $\rho_2 = 0.062$ ($n = 725$).

The hierarchical version of Model 1 is a basic variance components model showing the cross-classification of PSU and interviewer. Although the estimated standard errors of the random parameters are included in the table, the significance of the random parameters is based on a contrast test.³⁶ Significant variation was found between interviewers but not between PSUs.

The estimate for variation between geographic pools in the model was zero, even when employing the 2nd order estimation procedure. In the standard formulation of the model the individual variation is assumed to have a binomial distribution and is constrained to 1.0 (see Section 1.4.8)

In model 2 the individual level explanatory variable, number of children in the household, was included. This addition was useful for controlling any systematic differences among interviewers in the composition of their workloads. An interviewer whose interviews take place in households without children would be expected to differ on this item from those interviewers whose workloads contained a large number of households with children. This control variable has a significant coefficient in the hierarchical model. For fixed effects, significance may be judged by comparing the estimate with its standard error in the usual way.

36 As the distribution of the standard errors for the random parameters may depart considerably from normality, especially in small samples, a better option is to use a specific contrast on the random parameters.

Table 21: Cross-Classified Multilevel Logistic Regression Model: Response to the Interviewer Check Item ‘Whether Children Were Present’

EXPLANATORY VARIABLES	MODEL 1		MODEL 2		MODEL 3	
	FIXED EFFECT (Std. error)		FIXED EFFECT (Std. error)		FIXED EFFECT (Std. error)	
	Non-hier	Hier	Non-hier	Hier	Non-hier	Hier
Grand mean	-1.05 (0.08)	-1.05 (0.14)	-3.24 (0.37)	-3.30 (0.41)	-5.42 (0.94)	-5.49 (1.27)
No. of children in HH	--	--	1.20 (0.10)	1.23 (0.11)	1.23 (0.10)	1.25 (0.11)
Respondent gender (F)	--	--	0.62 (0.21)	0.59 (0.22)	0.62 (0.21)	0.60 (0.22)
Interviewer gender (F)	--	--	--	--	1.11 (0.43)	1.14 (0.62)
RANDOM EFFECTS SOURCE	VARIANCE COMPONENT (Std. error)		VARIANCE COMPONENT (Std. error)		VARIANCE COMPONENT (Std. error)	
Respondent	--	1	--	1	--	1
PSU	--	0.09 (0.12)	--	0.08 (0.17)	--	0.08 (0.17)
Interviewer	--	0.49 (0.20)#	--	0.89 (0.32)#	--	0.81 (0.31)#
Geographic Pool	--	0.0	--	0.0	--	0.0

Significant random parameters based on contrast test.

Also included is the individual level explanatory variable, respondent's gender. We expected that the presence of children during the interview would be a function of the respondent's gender, with women respondents being more likely to have children with them than male respondents. As can be seen by the values in Table 21, this expectation was confirmed.

It is interesting to note that the random coefficient for interviewers in the hierarchical version of the Model 2 increases in comparison to Model 1. This suggests that it is not haphazard variation in interviewer workloads that explains this interviewer variability, but rather that the variation among interviewers in recording the presence of children is greater when opportunity (i.e., children in household) is taken into account as well as respondent's gender. The basic conclusions about the fixed effects for Model 2 are the same for both the hierarchical and non-hierarchical versions of the model.

Several interviewer explanatory variables were then added in. These included interviewer age, gender, status (whether basic interviewer, supervisor, or area manager), and years of experience with the company. Also included was a measure of whether the same interviewer had visited the household for last year's interview. Of these various characteristics, only interviewer gender is considered in Model 3. It was clearly significant in the non-hierarchical model and only approached significance in the hierarchical one. It is interesting to note that in this case, different conclusions might have been reached depending on which model was considered. Also investigated was the possibility of an interaction between interviewer gender and respondent gender. This coefficient was not significant under either version of the model.

There are at least two possible explanations for the correlated interviewer effect in this case. First, there is quite likely a difference in the ability of interviewers to arrange the circumstances

of the interview so that the respondent is alone at the time - flexibility in making appointments, degree to which the interviewer emphasises the need for an undisturbed setting for the interview, etc. There is also the possibility that most of the between-interviewer variability is due to differences in the extent to which, or the circumstances in which, interviewers record the presence of children; one source of variation could be in the definition of others being present.

The key contrast here is between the message that we would obtain from ρ_1 and ρ_2 and the message from the multilevel analysis. With the former we would be concerned that the standard analysis would give spurious significance to the relationships estimated. In this case at least, however, the interviewer effect - though present for the dependent variable - does not affect the substantive analysis.

Table 22 deals with one of the respondent level factual items, newspaper readership. The variable modelled is a binary subcategory indicating whether or not the respondent typically reads the *Independent*. From the hierarchical analyses of variance, the estimated ρ values for this readership subcategory were $\rho_1 = 0.129$ and $\rho_2 = 0.106$ ($n = 1,268$).

Unlike the variance components model shown for the interviewer check item (see Model 1), the basic variance components model given in Model 4 shows significant variation between PSUs as well as between interviewers. However, there was no significant variation between geographic pools.

In model 5, the individual level explanatory variable, respondent's age, was included. Several other explanatory variables had also been explored in both the hierarchical and non-hierarchical

Table 22: Cross-Classified Multilevel Logistic Regression Model: Response to the 'Reads the Independent' Question

EXPLANATORY VARIABLES	Model 4		Model 5		Model 6	
	FIXED EFFECT (Std. error)		FIXED EFFECT (Std. error)		FIXED EFFECT (Std. error)	
	Non-hier	Hier	Non-heir	Hier	Non-hier	Hier
Grand mean	-3.04 (0.13)	-2.99 (0.30)	-1.70 (0.35)	-1.94 (0.45)	-2.99 (0.67)	-3.19 (0.90)
Respondent's age	--	--	-0.03 (0.01)	-0.03 (0.01)	-0.04 (0.01)	-0.03 (0.01)
Whether same interviewer as last year	--	--	--	--	0.21 (0.28)	0.63 (0.34)
Interviewer status Whether regular interviewer (compared to area manager)	--	--	--	--	1.35 (0.60)	1.06 (0.84)
Interviewer status Whether supervisor interviewer (compared to area manager)	--	--	--	--	2.25 (0.76)	2.23 (1.25)
RANDOM EFFECTS	VARIANCE COMPONENT (Std. error)		VARIANCE COMPONENT (Std. error)		VARIANCE COMPONENT (Std. error)	
SOURCE						
Respondent	--	1	--	1	--	1
PSU	--	1.55 (0.64)#	--	1.48 (0.63)#	--	1.59 (0.66)#
Interviewer	--	1.97 (0.71)#	--	1.78 (0.68)#	--	1.67 (0.67)#
Geographic Pool	--	0.0	--	0.0	--	0.0

Significant random parameters based on contrast test.

versions of the model (e.g., gender, social class, political party identification, and income) but only respondent's age was significant. With this addition, the interviewer random variation is reduced slightly and the PSU random variation remains essentially the same.

Of the various interviewer explanatory variables considered, two approached significance in the hierarchical version of Model 6. These were the binary variable for whether the same interviewer had visited the household for last year's interview (interviewer continuity) and one of the two dummy variables modelling the 3 category interviewer status variable (regular interviewer, supervisor, area manager). Here we can see that the interviewer variance component is again slightly reduced.

Interestingly we would have had a very different interpretation of which interviewer characteristics are having a significant impact if only the non-hierarchical model had been considered. With the non-hierarchical model, the interviewer continuity variable was clearly not significant and the two interviewer status variables were clearly significant. In addition, (although not shown in Table 22), interviewer's age approached significance. Middle-aged interviewers were more likely than elderly ones to record respondents as readers of the *Independent*.

Table 23 presents a behavioural intention item looking at whether or not the respondent expects to have any more children. As this is a subjective assessment, the question has been classified in the attitude category for the analysis. From the hierarchical analyses of variance, the estimated ρ values for this item were $\rho_1 = 0.075$ and $\rho_2 = 0.048$ ($n = 1,177$).

Table 23: Cross-Classified Multilevel Logistic Regression Model: Response to the 'Likely to Have More Children' Question

	Model 7		Model 8		Model 9	
EXPLANATORY VARIABLES	FIXED EFFECT (Std. error)		FIXED EFFECT (Std. error)		FIXED EFFECT (Std. error)	
	Non-hier	Hier	Non-hier	Hier	Non-hier	Hier
Grand mean	-0.39 (0.06)	-0.44 (0.11)	7.73 (0.46)	7.59 (0.46)	8.81 (0.60)	7.39 (0.48)
No. children in HH	--	--	-0.85 (0.09)	-0.83 (0.10)	-0.86 (0.10)	-0.84 (0.10)
R's gender (F)	--	--	-0.65 (0.19)	-0.63 (0.19)	-0.64 (0.19)	-0.62 (0.19)
R's age	--	--	-0.24 (0.01)	-0.23 (0.01)	-0.24 (0.01)	-0.24 (0.01)
Interviewer years w/ co.	--	--	--	--	0.042 (0.020)	0.043 (0.027)
RANDOM EFFECTS SOURCE	VARIANCE COMPONENT (Std. error)		VARIANCE COMPONENT (Std. error)		VARIANCE COMPONENT (Std. error)	
Respondent	--	1	--	1	--	1
PSU	--	0.15 (0.09)	--	0.00 (0.00)	--	0.00 (0.00)
Interviewer	--	0.22 (0.10)#	--	0.38 (0.16)#	--	0.34 (0.15)#
Geographic Pool	--	0.0	--	0.0	--	0.0

Significant random parameters based on contrast test.

As was the case for the variance components model under Model 1, Model 7 shows significant variation between interviewers, possible variation between PSUs, but not among geographic pools.

In model 8, individual level explanatory variables, number of children in the household, respondent's gender, and respondent's age were included. Each of these is highly significant in both the hierarchical and non-hierarchical versions of the model. With the addition of these explanatory variables in the hierarchical model, random variation due to PSUs goes to zero and random variation due to interviewers increases. The disappearance of the PSU effect may mean that the characteristics that led to the possible PSU effect have been adequately specified in the substantive model. Again, this suggests that it is not haphazard variation in interviewer workloads that is contributing to interviewer variability, but rather that there is variation among interviewers in their measurement of people's intentions to have more children.

In the non-hierarchical version of model 9, interviewer experience is a significant predictor with more experienced interviewers being more likely to record a 'yes' to the *more children* question than inexperienced interviewers. Although not shown, in the non-hierarchical model, the interviewer continuity variable approached statistical significance. When the same interviewer returned on the second wave of the survey he/she was less likely to record a 'yes' to the 'more children' than a different interviewer. These findings, however, do not hold for the hierarchical model.

Perhaps the most important point to note here is that, despite the strong interviewer effect, the substantive description represented by the substantive fixed part of the model is unaffected by the interviewers (at least not affected differentially). However, there are differences in the

conclusions about the impact of interviewer characteristics depending on whether an interviewer variance term is explicitly included.

In addition to these examples above, a further exploration was conducted of the impact of the background characteristics of the interviewers and years of experience (see Section 2.2) on model conclusions. For each of the different types of item (attitudes, facts, quasi-facts, and interviewer checks), a sample of variables was drawn from among those shown to have highly significant interviewer variability. Across the four categories, 26 items were drawn from 84. A cross-classified multilevel analysis (interviewer by PSU) was conducted on each one of these with the interviewer characteristics as the explanatory variables. These included interviewer age, gender, status, years with the company and an indicator of interviewer continuity over time. Of the 26 models considered, interviewer age was significant in 7 of the 26 cases (27 percent). The comparable percentages of significant effects that were found for the other interviewer characteristics were as follows: interviewer continuity (12 percent), gender (8 percent), interviewer status (8 percent), and years with the company (4 percent). Although such data should be treated with caution, it may indicate that interviewer age is a general predictor of some of the interviewer variability on the high variability items. These age findings are similar to those of Freeman and Butler (1976) and Collins and Butcher (1982) (see Section 2.2.1), but could in principle be due to the particular interviewers in this study. Significant effects of interviewer continuity were also found at a greater than chance level. The relationship between interviewer continuity and response quality will be explored further in Section 10.1.

Again we saw differences in these 26 models depending on whether a hierarchical or non-hierarchical model was used. The comparable figures for the non-hierarchical models were:

age significant in 27 percent of cases, interviewer continuity in 19 percent, gender in 12 percent, interviewer status in 35 percent, and years with the company in 15 percent. In 11 of the 26 models, different conclusions about the effects of interviewer characteristics on substantive results would have been reached, depending on whether an interviewer variance term was explicitly included in the model.

9.4 Chapter Summary and Discussion

The assumption underlying most statistical software - that the observations are independent and identically distributed (*iid*) - is certainly not appropriate for most sample survey data. Variances computed on this assumption do not take into account the effects of survey design (e.g., inflation due to clustering) and execution (e.g., inflation due to correlated interviewer effects).

There are two different reasons why we might be interested in interviewer effects and sample design effects. The first is to establish whether the sample design (typically clustering in the design) and/or the interviewer (because many respondents are interviewed by each interviewer) have an effect on the variance-covariance structure of the observations. This is the traditional sample survey approach and includes consideration of the *design effect* and the *interviewer effect* following the ANOVA and correlation models (see Sections 1.2.1 and 1.2.2). The emphasis is on the estimation of means or proportions and on the standard errors of these estimates; variance components models do not add anything to these analyses.

Using the specially designed study in Wave 2 of the BHPS permitted assess to both these inflation components. Across the 820 variables in the study, there was evidence of a significant impact of both the population clustering and the clustering of households/individuals

in interviewer workloads. The intraclass correlation coefficient, ρ , was used as the measure of homogeneity. We found that sample design effects and interviewer effects were comparable in impact, with overall inflation of the variance as great as five times the unadjusted estimate.

The median effect across the 820 variables was an 80 percent increase in the variance. The magnitude of the intra-interviewer correlation coefficients was comparable across question types, though the most susceptible items tended to be the interviewer check items. There was a tendency for variables that were subject to large design effects to be sensitive also to large interviewer effects and a possible interpretation of this correlation was offered in Section 9.3.1.

The large values of ρ on particular items and the fact that ρ is of the same order of magnitude as ρ_d suggests that survey organisations should attempt to incorporate measurement of ρ into their designs. If the necessary modifications of the survey design are too expensive to allow this, organisations should at least try to minimise its impact; this could be accomplished by reducing interviewers' workloads. Current practice tends to favour smaller dedicated interviewer forces with large assignments; in the presence of substantial interviewer effects this is a misguided policy.

The second reason is to ensure that effects on the univariate distributions do not contaminate estimates of relationships among variables in the population; in this case the objective is to control the effects or to eliminate them from the analysis. The standard approach of the survey sampling statistician is to estimate the parameters assuming *iid* and produce design-based variance estimates using re-sampling methods such as the jack-knife or bootstrap; this however is only an approximate solution. The explicit modelling of effects is both more precise and more informative. In this situation there are two aspects of interest: whether explicitly including the sample clustering and the interviewer workloads in the model changes the

estimates of the relationships (the contamination issue), and whether the clustering and interviewers have an effect on the distribution of values obtained for the dependent variable.

Software developed for multilevel analysis (hierarchical modelling) presents an alternative framework within which to consider the sample design and interviewer effects by incorporating them directly into substantive models of interest. For illustration three binary items were chosen - an interviewer check item on *whether children were present during the interview*, a behavioural item, *readership of the Independent*, and a subjective item, *whether respondents thought it was likely that they would have another child*. For each of these items, a significant interviewer effect was found, which persisted when the inequalities in the interviewers' workloads and various background characteristics of the interviewers were controlled for. For other items not presented here, some support for interviewer's age was found as a possible explanation for the interviewer effects. In addition, it was found that conclusions about the influence of the various background characteristics and years of experience of interviewers would have differed in many cases if only the standard non-hierarchical model rather than a hierarchical one had been used.

From a modelling standpoint incorporating explanatory variables is one of specifying appropriately the underlying factors in the substantive models of interest. From a sample survey standpoint the issue is that of incorporating in the analysis a recognition of the special features of the sample design and survey execution that make a particular data set deviate from *iid*. Multilevel models have a natural congruence with many important aspects of the survey situation; both the sample design and the fieldwork implementation can be described appropriately as introducing *hierarchical levels* into the data and thus multilevel analysis

provides a framework that makes it possible to include both substantive and design factors in the same analysis.

CHAPTER 10 DOES INTERVIEWER CONTINUITY AFFECT RESPONSE QUALITY?

10.1 Background

In Chapter 8 the effect of interviewer continuity on nonresponse was explored. This Section explores the effect of interviewer continuity on response quality; a topic which has rarely been empirically explored. The conventional survey wisdom is that interviewer continuity should be beneficial to response quality. As suggested by the discussion in Section 8.1.1, one would expect the intervening factor between interviewer continuity and response quality to be the level of rapport between the interviewer and the respondent. Multiple interviews over the life of a survey would obviously be beneficial in building such rapport. However, as described by Harkess and Warren (1993), too much rapport can be detrimental to response quality (see Section 2.1.1). This latter concern is also supported by the work of Weiss (1968) and Anderson, Silver, and Abramson (1988a, 1988b) as described in Section 2.2.1. They found that inaccurate reporting can occur when purposely matched interviewers and respondents identify too much with each other.

The results described in Chapter 9 suggest that interviewer continuity may indeed have an effect on response quality. In Section 9.3.2, interviewer characteristics were examined as predictors of interviewer variance in a sample of 26 of the substantive variables which had been susceptible to very high interviewer values of ρ , the intra-interviewer correlation coefficient. Interviewer continuity was a significant predictor in three instances (12 percent of the cases). This is a small percentage but nonetheless greater than what would have been expected based on chance alone.

10.2 Data and Methods

All variables from Wave 2 of the BHPS were used for this analysis. Variables from the household questionnaire cover various aspects of the household's financial situation. They also include various questions on consumer behaviour and on the type of accommodation. Several of the derived variables at the household level summarise the characteristics of household members (e.g., household size, number of children, number of pensioners, number of couples, number of lone parents, number of employed individuals, numbers or wage earners). The individual questionnaire covers neighbourhood and individual demographics; health and caring; a lifetime marital, fertility, and employment status history; current employment and employment over the previous year; values and opinions; and household finances and benefits. The self-completion questionnaire, which showed modest interviewer effects in Chapter 9, includes the General Health Questionnaire (GHQ) items, family attitude questions, and social network questions.

As was done in Section 8.2, cases were restricted to fully participating Wave 1 respondents who did not move out of their local area and to the portion of the interpenetrated sample with no random attrition. This creates the group where any true differences due to interviewer continuity will be discovered. It also reduces the sample size considerably to 1,006 individuals and 564 households. Analyses were also restricted to variables with a total n of cases of at least $C \times 25$, where C was the number of categories. Sparse categories were collapsed where needed. This resulted in 111 variables and derived variables at the household level to be considered and 584 variables and derived variables at the individual level to be considered.

Initial analysis of these 695 variables consisted of simple cross-tabulations of the substantive characteristic by interviewer continuity.³⁷ Constructed indicators of response quality were also examined. These included a total count of refusals across all variables, a total count of don't know responses across all variables, a total count of the missing/wild values across all across all variables, and a total count of the number of imputations for the income items. These four quality indicators were constructed separately for the household and individual levels. Also considered was whether the respondent allowed the interviewer to check his/her payslip.

10.3 Results

10.3.1 *Changes in Substantive Answers*

Although random assignment of respondents to interviewers took place throughout the interpenetrated sample, some respondents received a different interviewer because their first interviewer had left employment. As noted in Section 10.1.1, the analysis was restricted to the portion of interpenetrated sample with no interviewer attrition between Wave 1 and Wave 2. Random assignment generates equivalent samples in expectation. In practice, however, sub-samples created by random assignment may not be equivalent. This is a concern for the analyses here. We need to be sure that any differences noted between respondents receiving the same or a different interviewer are indeed due to some effect of the interviewer and not simply due to chance differences between the two samples. This is also a concern for the interviewer continuity analyses reported in Chapter 8.

37 In addition to examining direct differences in reporting based on whether the respondent received the same or a different interviewer, it would have been interesting to see if the magnitude of interviewer variability differed between the two interviewer groups. However, due to the basic design of the interpenetrated sample, it is not possible to study this latter option. Partitioning the sample into (same and different interviewer portions) means that interviewer and PSU variance are confounded within each portion.

An analysis of the 111 variables and derived variables from the household question yielded one significant result at the $p < .05$ level. Given the wide range of characteristics covered, this is re-assuring with respect to the equivalency of the two portions of the sample. The one significant difference was the month of interview. As shown in Table 24 below, interviewers had a distinct preference for visiting their former respondents first.

Table 24. Month of Interview by Interviewer Continuity

Month	Different Interviewer %	Same Interviewer %
September 1992	26	52
October	56	41
November	15	5
December	1	0
January 1993	1	0
February	1	0
March	1	1
April	0	0
<i>n</i>	289	273

$$\chi^2 = 51.96, df = 7, p < .001$$

With respect to the 584 variables and derived variables from the individual and self-completion questionnaires, only 32 cross-tabulations were significant at the $p < .05$ level.³⁸ This equates to 5.5 percent of the comparisons. Certainly not pervasive, but

38 The three significant results of interviewer continuity reported in Chapter 9 were month of

rather just slightly above what we would expect to find from chance alone. In fact given the multiple tests, if one were to consider a Bonferroni approach (setting $\alpha = .05$ for the whole set of tests and dividing it by the number of tests to obtain the proper limit for a Type 1 error), then none of the findings would be significant. Thus, one could argue that the matter should simply be dismissed here. However, it is interesting to note that the 32 significant comparisons are not randomly spread throughout the questionnaire. This will be considered below. But first, evidence for any true differences in the two samples must be weighed.

Five of the significant differences found are likely to reflect true differences between the same and different interviewer portions of the sample. For example, significant differences at the $p < .05$ level were found for the Register General's Social Class variable.

Significant differences were also found for two of the five occupational classifications as defined by Major Group of the Standard Occupation Classification (SOC, see OPCS, 1990). These included the respondent's current and second jobs. Differences for the respondent's job in the previous year approached significance ($p < .10$). The results with respect to social class suggest that the same interviewer sub-sample slightly over-represents persons in professional occupations and slightly under-represents those in managerial and technical occupations. It also over-represents partly skilled manual workers. A similar picture is painted by SOC Major Group with the same interviewer sample slightly over-representing those in professional occupations and under-representing managers and administrators and over-representing those in craft and related occupations. A similar, though less pronounced pattern, was found for the occupational

interview, adults present during interview for the values section, and month left full time education. The first of these shows similar conclusions here. The second just misses being significant in the current analyses due to the smaller sample size employed in this Chapter (see Section 10.1.1) and the third was not analysed because its sample size was too small in this Chapter.

classification of the respondent's job in the previous year. With respect to the respondent's second job, we find that the same interviewer sample over-represents clerical and secretarial occupations and under-represents craft and related occupations and other occupations.

There is also a variable which looked at change in the respondent's financial situation since the previous year. Respondents from the same interviewer group were more likely than respondents for the different interviewer group to say that their financial situation had changed for the worst since the previous year. One might suspect that this reflects a true change as manual occupations are often more vulnerable to changes in the economic climate. However, a cross-tabulation of social class and change in financial situation shows a slight trend in the opposite direction, but it is far from being significant. It should also be remembered that across the numerous financial variables there is no significant difference in income between the same interviewer and different interviewer groups.

Another variable which is likely to present real differences, although not necessarily important ones is the 'number of children in the household'. Although there is no difference in the mean, the same interviewer sample has more 'one child households' and the different interviewer group has more 'two children households'. These discrepancies perfectly balance each other.

A description of the remaining 27 findings are shown in Table 25. As can be seen, the findings are not randomly spread across the questionnaire. Several are clustered within particular topic areas. Given the clustering by topic, it is possible that some of the differences listed in Table 25 are simply reflective of the differences in social class and the

change in financial situation. Several of the variables in Table 25 were indeed related to social class and change in financial situation. However, when social class or financial situation were used as control variables the same patterns with respect to interviewer continuity still emerged.

Table 25: Differences in Substantive Answers Depending of Whether the Respondent Received the Same or a Different Interviewer

Topic	No. of Variables on Topic	No. of Findings $p < .05$	No. of Findings $.05 < p < .10$
Interviewer's observation of 'other adults' of whether other adults present during interview	6	3	2
Attitudes towards moving	9	4	0
Looking for work, type of job wanted	8	3	1
Adoption of children	4	2	1
Political party membership	4	1	3
Union membership	3	1	1
Membership in groups	26	1	2
Political attitudes	6	1	0
Attitudes towards world concerns	1	1	0
Reports of health problems: problems seeing	13	1	0
Attitudes towards wealth distribution	4	1	0
Determining employment status: having a job although off last week	3	1	0
Job satisfaction: overall satisfaction	8	1	0
Attitudes towards usefulness of training	7	1	0
Family commitments	6	1	0
Income sources: income from lodgers	33	1	0
Savings: amount saved	3	1	0
Money transfers: first recipient	21	1	0
Income Imputation Flag	25	1	0
<i>Total</i>	<i>190</i>	<i>27</i>	<i>13</i>

Considering just the topic areas covered by Table 25, then the proportion of significant findings would rise to 14 percent, perhaps providing some support for a true interviewer continuity effect. What is unclear, however, is whether there is any particular pattern to these effects. The effects are clearly spread across attitudinal variables as well as factual

ones. Another dimension to consider is the sensitivity of the question. What would be the socially desirable response in each case? For example, socially desirable responses could be to describe the job you are looking for as grand, not to say you had adopted children, say you are a union member, say you are a member of a political party, say you are a member of other types of groups, not report having any sight problems, agree that wealth distribution is unjust, say you had a job even though you didn't earn money last week, say you are very satisfied with your job, say that you believe training is useful, not mention that you need money from lodgers, and say that you save a lot. In comparing the direction of the actual responses to these 12 predictions, it was found that four were in the direction of a socially desirable response, seven were in the direction of a non-socially desirable response, and one showed mixed results.

Another aspect to consider is the longitudinal one. Persons tend to be consistent in their attitudes and behaviour over time. Aside from the interviewer check variables, the imputation flag, and the question about working last week, there are 16 categories of variables where meaningful comparisons can be conducted. Fourteen of these have at least one variable from Wave 2 with an identical counterpart in Wave 1. All of the Wave 1 and 2 pairs are indeed highly related to each other. Given this consistency, one would hypothesise that the presence of the same interviewer would boost this consistency. The findings are again mixed, showing both increased consistency and increased inconsistency when the same interviewer was present. Some examples of increased inconsistency are as follows. People who reported problems with their sight in the previous year were more likely than those who didn't, to report this again, but there was still variation based on who the interviewer was. Those respondents receiving the same interviewer were actually less likely to say they have difficulty seeing this year. A similar phenomenon was

observed with respondents' preference for moving house. Although those wishing to move, still prefer to move and those wishing to stay, still prefer to stay, there is nonetheless a small trend for those who said they wished to stay in Wave 1 to report wanting to move when they have the same interviewer, but not when they have a different interviewer. Two examples of increased consistency are as follows. Those satisfied with their job in Wave 1 are more likely to be satisfied with their job in Wave 2. When they receive the same interviewer they are even more likely to say they are completely satisfied. Persons who said there was no union at their workplace in Wave 1 were most likely to say this again in Wave 2, however, if they received the same interviewer they were even more likely to say this.

10.3.2 Indicators of Response Quality

Results with respect to the five 'other indicators' of response quality are shown in Table 26. Although 7 of the 9 indicators are in the direction suggesting that poorer quality is obtained with the same interviewer, only one of these reaches significance at the $p < .05$ level. It is the variable regarding the checking of the respondent's payslip. Same interviewers were significantly less likely to have checked the respondent's payslip than different interviewers. Two other indicators are significant at the $p < .10$ level with same interviewers having more refusals for the household questionnaire and being more likely to have 2 or more don't knows for the individual questionnaire and self-completion form. Perhaps increased rapport with the same interviewer allows respondents to feel comfortable enough to refuse and say 'don't know'. Or similarly, not wishing to break the rapport, same interviewers may be more tolerant of these behaviours and less willing to test the rapport by asking for the respondent's payslip.

Table 26: The Relationship between Interviewer Continuity and Other Measures of Response Quality

	<u>Household Level</u>	<u>Individual Level</u>
Total count of refusals	More with same int (p < .10)	More with same int (NS)
Total count of don't know	More with same int (Not signif)	More with same int (p < .10)
Total count of the missing/wild values	Less with same int (Not signif)	Less with same int (Not signif)
Total count of the of imputations	More with same int (Not signif)	More with same int (Not signif)
Interviewer checked payslip	—	Less with same int (p < .01)

10.4 Chapter Summary and Discussion

This Section has explored the possible effects of interviewer continuity on response quality. First examined was the extent to which having the same or a different interviewer at Wave 2 actually affected the substantive answers given. Across all of the variables and derived variables from the individual and self-completion questionnaires, only 5.5 percent showed a significant result. However, these results were not randomly spread across the questionnaire. When just those topics which showed effects were considered, the percentage of significant effects was boosted to 14%. However, no consistent pattern could be found for these effects. Effects were found across both attitudinal and factual questions. The direction of the effects were sometimes in the direction of a socially desirable answer and sometimes in the direction of a non-socially desirable answer. Also effects were not simply due to increased longitudinal consistency due to the presence of the same interviewer.

Other indicators of data quality were also examined, such as counts of refusals and don't knows across the questionnaire, counts of missing or wild values, and counts of

imputation flags on the income variables. None of these proved significant at the $p < .05$ level. Some interviewers, however, were significantly less likely to check the respondent's payslip.

Thus given these two areas of exploration, there is no evidence to suggest that the impact of interviewer continuity on response quality is pervasive. Similarly, there is little evidence from the few effects which were found to say whether the influence of interviewer continuity is positive or negative.

PART 4 DISCUSSION AND INTEGRATION

CHAPTER 11 SUMMARY AND CONCLUSIONS

11.1 Integrating What the Various Strands of Empirical Work Have to Say

11.1.1 The Background Literature

This thesis explored the impact of interviewers on the survey process with a focus on their contribution to response variance and nonresponse variance and bias.

It began with a discussion of total survey error. In Section 1.1, we saw that different breakdowns have been considered by different authors (for example, response error versus statistical sampling error, see Frankel and Dutka, 1983; the bias of nonresponse versus the standard error of the response, see Deming, 1953; and all biases versus all variable errors, see Kish, 1965). Despite these differences there is general agreement on the following sources of error in survey data: coverage error, sampling error, nonresponse error, response error and processing error (see Groves, 1989; Andersen *et al*, 1979). In turn, response error can be seen to include effects due to the interviewer, the respondent, the questionnaire, and the mode of data collection. In addition each of these can be conceptualised to contribute to both variable error and to bias error.

Total survey error provides a useful context from which to view the specific contributions of the interviewer to response variance and nonresponse variance and bias which have been considered in this thesis. This context reminds us that the interviewer is just one component in the larger picture of total survey error. It also reminds us that as a component, interviewers are intimately linked to the other components. For example, if interviewers pressurise respondents on the doorstep in order to reduce nonresponse, they may well be increasing response error due to an uncooperative respondent. Similarly, interviewers need to be studied

in the context of the full survey interaction process, as interviewers and respondents interact with each other, with the questionnaire, and with a given mode of data collection. This interactive nature is confirmed by the models of the survey interaction process which were discussed in Chapter 4.

Interviewers contributions to response error are typically measured in terms of correlated variance rather than bias. Chapter 1 described the two main models for estimating the correlated interviewer effect (the ANOVA model and the Correlational model). More general measurement error models from both a sampling statisticians' and psychometrician's point of view were considered in Chapter 4.

Chapter 2 provided insight into how interviewers can influence data. The actual roles and tasks of the interviewer were examined. How interviewers affect response quality and nonresponse was considered from the perspective of their background characteristics and psychological factors, in addition to their actual behaviour.

Chapter 3 provided background on nonresponse, noting recent trends in the growth of nonresponse and defining response rates, nonresponse bias, and nonresponse variance. It revisited the role of the interviewer as part of an integrated theoretical model of survey nonresponse (for a further discussion of this, see Section 11.1.3).

11.1.2 The Opportunities and Challenges Presented by the Data

The contribution of the interviewer to response variance and nonresponse variance and bias was greatly facilitated by the use of the interpenetrated sample design experiment implemented

in Wave 2 of the BHPS. The interpenetrated sample design at Wave 2 allowed the opportunity

- to separate interviewer from area effects for both nonresponse and response investigations (these are typically confounded by design) (see Chapters 6, 7, and 9),
- to study of wide range of the correlates of nonresponse (because Wave 1 data were available for both Wave 2 respondents and nonrespondents and through access to auxiliary sources of data) (see Chapter 5),
- to test experimentally the effects of interviewer continuity on response rates and on response quality (see Chapters 8 and 10, respectively), and
- for continued investigation into the random effects of interviewers/PSUs and the ‘interviewer continuity’ issue at Wave 3 (see Chapter 7 and 8, respectively).

The data represented a challenge to analysis because of their hierarchical and cross-classified nature. For example the data from the interpenetrated sample design at Wave 2 are structured as individuals/households nested within an interviewer by PSU cross-classification within geographic pools. At Wave 3, where the randomisation of household to interviewer was exactly reversed, the data are structured as individuals/households nested within an interviewer/PSU by interviewer continuity cross-classification within geographic pools. These analytic challenges were explored through various techniques with a main focus on the use of hierarchical analyses of variance (see Section 1.4.2) and cross-classified multilevel models (see Sections 1.4.5-1.4.9)

11.1.3 Interviewers and Nonresponse

Several issues were explored with respect to interviewers and survey nonresponse.

The Theoretical Model. An integrated theoretical model of survey participation (see Figure 6 in Chapter 3) was first introduced by Groves, Cialdini, and Couper (see Groves and Cialdini, 1991; Groves, Couper and Cialdini, 1992; Groves and Couper, 1994; 1995). In Chapter 5 of this thesis, this theoretical model was extended to cover the unique sources of data available (see Figure 14a). For example, as described in Section 5.2, the larger *Social Context* in which the survey is embedded can be seen to influence the respondent directly as well as through the household in which he/she lives. The *Survey Design* itself is seen to influence both the *Respondent* and the *Interviewer*. The *Respondent* and the *Interviewer*, in turn, bring with them various background characteristics which affect how they behave in the *Respondent-Interviewer Interaction* which takes place on the 'doorstep'. The characteristics of *Respondents* and *Interviewers* also influence the likelihood of the interviewer actually making contact with the respondent. *Past interviewer/respondent interactions* are also seen to influence the current ones.

Such a theoretical model offers an ideal way to study the disparate influences on survey nonresponse in a unified way. (In an ideal research design, all elements and all cells in Figure 14a would be estimated with data. The possibilities of this are explored further in Section 11.3). Although the BHPS data and auxiliary sources of data added elements and cells to the original model, these sources did not include data to cover all aspects of the theoretical model. Thus, in this thesis the reduced model presented as Figure 14b has been used. It includes the aspect of geographic areas (as available through 1991 Census small area data) under the heading of the *Survey Context*, the issue of interviewer continuity under the heading of *Survey Design*, the socio-demographic and economic characteristics of *Respondents*, the socio-demographic characteristics and experience of *Interviewers*, and call record and co-operation data from past and current *Respondent-Interviewer Interactions*.

Results with respect to each of these points will be discussed in turn and summarised in the sub-section below entitled, “Summarising the Results with Respect to the Reduced Theoretical Model.”

Interviewers versus Areas. Chapter 5 provided general background about the correlates of nonresponse, examining the effects of the background characteristics of interviewers and their years of experience as well as exploring the effects due to the characteristics of respondents, households, and areas. Bivariate analyses suggested that several of the easily measurable characteristics of interviewers (such as age, gender, experience, and grade level) and areas (such as population density, proportion of flats in the area, percentage of non-white residents, etc. taken from UK 1991 Census small area data) were important predictors of nonresponse. Interestingly, these relationships almost completely disappeared when the characteristics of specific households and respondents were controlled for in multiple regression analyses. However, the importance of the interviewer and the area reappeared in Chapter 6 in the cross-classified multilevel analysis in which terms for the interviewers and areas were treated as random effects. This suggests that what is making a difference in terms of nonresponse is more subtle and elusive than the easily measured characteristics of interviewers and areas.

Looking at the overall contribution to the variance (using the random effects model in Chapter 6), it was seen that the effects of the interviewer often tend to predominate over those of the area (PSU or larger geographic pool). After controlling for the characteristics of households and individuals, however, the interviewer effects remained strong for individual and whole household refusals, but virtually disappeared in the case of

individual and whole household non-contacts, where the area was the more important factor. Using a slightly different sub-sample and different analysis strategies, the results of Chapter 7 parallel these findings, except that the influence of the interviewer is maintained for whole household non-contacts. An investigation into the discrepancy between these two sets of results suggested that the main difference was due to the fact that the household non-contact models in Chapter 6 included the variable BIVLNC (Total calls at Wave 2) and the household non-contact models in Chapter 7 did not. As described in Section 7.4.2, BIVLNC is a problematic variable in several respects and the preferred models are those without it. Thus the importance of the interviewer in household non-contacts should be considered.

Chapter 7 also suggests a positive relationship between the random effects for interviewers for refusals and the random effect for interviewers for non-contacts at both the individual and household levels suggesting that interviewers from this population are generally either good or poor at these different tasks. With respect to areas, a positive relationship between the random effect for PSUs for refusals and the random effect for PSUs for non-contacts was found at the household level in the models in Appendix D, suggesting that areas which are difficult in one aspect are difficult in the other as well. A negative relationship between these two components for PSUs, however, was found within households.

The combined effects of interviewers and PSUs were seen to persist at Wave 3, though they tended to be of a smaller magnitude.

Interviewer Continuity. Despite the conventional survey wisdom that allocating the same interviewers to the same respondents across waves of a panel study improves response rates, no observable effects of interviewer continuity on response rates were found. This held true across all of the different types of analyses considered in Section 8.1: bivariate, multiple logistic regression, and cross-classified multilevel. As noted in cross-classified multilevel models in which the coefficient for interviewer continuity was allowed to vary, there was some variability in the effectiveness of an interviewer continuity strategy across geographic pools with respect to individual level nonresponse, refusals, and non-contacts although this did not reach significance. In addition, no interactions were found between interviewer continuity and the measurable characteristics of individuals, households or areas, with respect to nonresponse. The same was true for the measurable characteristics of the interviewer.

There was concern that the lack of an interviewer continuity effect in Section 8.1 may have been partly due to the fact that the ‘new’ interviewers were encouraged to talk to the ‘old’ interviewers about possible hard-to-find and problem respondents prior to fieldwork. This could have artificially reduced the interviewer continuity effect. On the other hand, if BHPS interviewers actually practised the resourceful strategies (which were their own creation) and which they described in the qualitative debriefing study, this may account for the lack of quantitative findings as well as interviewers’ feelings that interviewer discontinuity is not problematic. Re-visiting the interviewer continuity issue in Section 8.2, however, gives further support that this field strategy was not the cause of the lack of interviewer continuity findings. First it should be noted that only interviewers in the portion of the interpenetrated sample without any interviewer attrition from Wave 1 to Wave 2 could practice this strategy. In the portion of the interpenetrated sample where

there was interviewer attrition, the Wave 2 interviewer would have no one to talk to as her/his counterpart had left employment. Interestingly, the same pattern of no interviewer continuity effects was found in the portion of the interpenetrated sample with interviewer attrition as in the portion of the interpenetrated sample without interviewer attrition. In addition, there were no interviewer continuity effects observed in the full Wave 2 sample, three-quarters of which was not interpenetrated.

This suggests a general lack of interviewer continuity impact on response rates for Wave 2, but what about for later waves? Laurie and her colleagues (1997) had found significant interviewer continuity effects when they studied trends from Waves 2 to 4 using the full BHPS sample. The results in Section 8.2, however, suggested that the source of their findings was the non-random nature of interviewer attrition, with respondents being more likely to receive different interviewers in the more difficult survey areas.

Section 8.2 also considered the possibility of random slopes for the interviewer continuity coefficient at Wave 3 using cross-classified multilevel models. Again variability was present, although it was not of a large enough magnitude to be significant.

Summarising the Results with Respect to the Reduced Theoretical Model. The results described above cover all of the cells of the reduced theoretical model. The findings, however, are in terms of the direct effect of each element on the final elements of *Complete Non-contact* and the *Decision to Co-operate or Refuse*, rather than in terms of the various causal paths in the model. The combined results suggest that ‘neighbourhood characteristics’ from the *Social Context* cell generally did not have a significant direct impact on the final outcomes once household and individual level characteristics were

controlled for. The same was true for 'interviewer continuity' from the *Survey Design* cell and the interviewer's 'socio-demographics' and 'experience' from the *Interviewer* cell. (Others aspects of the interviewer, as captured through random effects models, however, did suggest an influence of the interviewer.) In contrast, *Household Factors* and 'socio-demographic and economic' characteristics from the *Respondent* cell were excellent predictors of both outcomes, as were the various indicators of co-operation and contactability from the past and current *Respondent-Interviewer Interaction* cells.

Although estimating the path analytic coefficients for the reduced theoretical model was not a goal of this thesis, this would be useful to consider for future research. Given the large number of explanatory variables available in all of the cells (with the exception of the *Survey Design* and *Interviewer* cells) and the fact that there were actually 6 outcome variables rather than two, such a task is not straightforward. One approach would be to construct some type of combined indices for each cell which could then be used in the path analysis. It could also be worth considering some type of structural equation modelling (see Long, 1983) as it combines the best aspects of confirmatory factor analysis and regression analysis (i.e., the path coefficients can be estimated among the error-free latent concepts rather than the error-prone original survey measures).

11.1.4 Interviewers and Response Error

Interviewers versus Areas. With respect to response error, the assumption underlying most statistical software is that the observations are independent and identically distributed (*iid*). This is certainly not appropriate for most sample survey data. Variances computed on this assumption do not take into account the effects of survey design (e.g., inflation due to

clustering). Chapter 9 shows the importance of also taking into account the effects of the execution of the survey design (e.g., inflation due to correlated interviewer effects).

The 820 variables or categories examined from Wave 2 of the BHPS showed evidence of a significant impact of both the population clustering and the clustering of households/individuals in interviewer workloads. Sample design effects and interviewer effects were found to be comparable in impact, with the median being an 80 percent increase in the variance. The magnitude of the intra-interviewer correlation coefficients was comparable across question types, though the most susceptible items tended to be the interviewer check items. There was a tendency for variables that were subject to large design effects to be sensitive also to large interviewer effects and a possible interpretation of this correlation is offered in Section 9.3.1.

The explicit modelling of area and interviewer variance was made possible through software developed for cross-classified multilevel analysis. The examples shown in Chapter 9 suggest that significant interviewer effects persisted when controlled for inequalities in the interviewers' workloads and various background characteristics of the interviewers and their years of experience. In addition, there was some support (across a wider selection of items) for interviewers' age as a possible explanation for the interviewer effects. The conclusions also suggested that the influence of the various background characteristics and years of experience of interviewers would have differed in many cases, had only the standard non-hierarchical models rather than hierarchical ones been used.

Interviewer Continuity. Chapter 10 investigates the conventional survey wisdom that maintaining interviewer continuity in a panel study also improves the quality of the data collected. Differences between the group of respondents receiving the same interviewer and

those receiving a different interviewer were explored. Across the 111 household variables from Wave 2 of the BHPS there were no significant effects of interviewer continuity. The one exception was the month of interview which suggested that interviewers preferred to visit their respondents from the previous year first. Across the 584 individual variables from Wave 2 of the BHPS, only 33 (5.5 percent) showed significant effects. Although these were clustered among a much smaller subset of items, there were no patterns in the results with respect to question type, question sensitivity, or consistency from the previous wave. Although this suggests little evidence of an interviewer continuity effect with respect to the quality of the data, it does suggest that the same and different interviewer sub-samples are indeed equivalent. Examining the counts of various types of missing data also suggested little or no significant impact of interviewer continuity on substantive results.

11.2 Implications for Survey Research Practice

11.2.1 Implications for Fieldwork Strategies

Targeting Respondents and Areas to Improve Response Rates. For longitudinal studies, the analyses clearly suggest groups of respondents with particular characteristics who can be identified well in advance of Wave 2 or subsequent waves who could benefit from special targeted fieldwork strategies at Wave 2 or subsequent waves in order to improve response rates (see Chapter 5).

It should be remembered that some of the best indicators are interviewers' subjective ratings of the person's co-operation at the initial wave and variables constructed from the previous call-records and outcomes. Thus, these variables should be routinely included in the Wave 1 data collection.

It could also be possible to develop targeted strategies for one-off surveys, if one could identify the characteristics of respondents in advance. One possibility, as discussed by King (1996), would be to link a geography-based sampling frame (e.g., the Postcode Address File) to a standard area classification code which identifies the predominant characteristics of people in the area (take, for example, ACORN which was developed by CACI and consists of 38 neighbourhood types). Such a strategy, however, does have its limitations as area classifications are often based on Census small area data and as was seen in Chapter 5, these are only an indirect indicator. Also the effects of areas are mainly manifested in terms of problems with making initial contact, rather than with refusals.

Improved Data for Nonresponse Weighting. For a one-off survey, the analyses in Chapter 5 suggest the types of socio-demographic and economic variables that ideally should be sought for nonresponse weighting. This is true whether the nonresponse weighting proceeds through comparison to a more complete data source such as the Census, through information on respondents and nonrespondents present on the sampling frame, or through the use of a special data collection form for the nonresponse cases which is typically completed through observation by the interviewer.

This can be illustrated with the case of a nonresponse form. What should one contain? Results from Chapter 5 suggest that the key items which could be gathered by interviewer observation are approximate age, gender, and race. Lynn (1996) suggests the usefulness other items such as type of housing, the physical state of repair of the housing in the area, whether there are entry phones or other security devices, the floor of the building on which the flat is located, and perhaps a question on whether the sampled address is better or worse than others in the area. Another possibility is household size.

Lynn (1996) also describes a unique approach to nonresponse forms used for the 1996 British Crime Survey in which nonrespondents were asked a very small set of key questions from the survey. This strategy actually proved to be very successful. If questions are to be asked of the nonrespondents for weighting purposes then Chapter 5 suggests that items that are economically-related should be of value such as whether the occupants own or rent; the number of cars they have access to; their social grade; whether they are employed, unemployed, or not in the labour force, etc.

Other Issues with Respect to Nonresponse. The results from Chapter 7 suggests that there are indeed difficult areas in which to achieved good response rates and that often these suffer from both higher numbers of refusals and higher numbers of non-contacts.

Yet, the results with respect to the indicators of co-operation and contactability (see Chapter 5) suggest that reluctant respondents none-the-less do participate. Other research also confirms the situational rather than ideological nature of most refusals (see Campanelli, Sturgis, and Purdon, 1997). Thus field organisations should keep in mind the possibility of always revisiting past refusals and non-contacts at future waves of a panel study as these households and individuals may now participate.

But it must be kept in mind that there is a point of diminishing returns. Pressurising the unco-operative into an interview may boost response rates, may also lower the quality of the data.

Implications for Interviewer Selection. Do the results from this research suggest that certain types of people make better doorstep interviewers and should this influence interviewer recruitment decisions?

A perennial concern of field departments is whether the visible socio-demographic and economic characteristics of interviewers affect survey results. Several studies have looked at the role of interviewer characteristics and response quality, but very little is known about the role of interviewer characteristics in nonresponse.

Chapter 5 suggests that for a general population survey without sensitive items such as the BHPS, the characteristics of the interviewer are not important for nonresponse. This also includes the possibility of interactions between the characteristics of the interviewer and those of the respondent. For example, there is sometimes a concern about response rates for male interviewers, particularly among female respondents. However, no significant problems of this type were detected in the BHPS data. This is heartening given that many more men are becoming interviewers. The day is approaching when the typical UK female interviewing force will be half male. As is the case for response quality, however, we suspect that both of these findings would not hold for surveys whose topics are directly related to the visible characteristics of the interviewer (such as a survey on sexual practice).

Interviewer Continuity. The fact that no net interviewer continuity effect was found in the current study with respect to response rates (Chapters 7 and 8) or the quality of the data (Chapter 10) is somewhat reassuring. Attrition of the interviewing workforce is a perennial problem in most survey organisations. However, this should not mean that interviewer attrition should be ignored!

Although the interviewer continuity results of this study should generalise to other types of longitudinal surveys as well as other classic panel study designs, we feel it would be unwise to ignore interviewer continuity issues. These need to be considered in light of the particular topic of the survey, a given interviewing staff and their level of resourcefulness. (As seen from the qualitative debriefing material, BHPS interviewers were particularly resourceful in any interviewing situation.) In addition, survey organisations should keep in mind that even if interviewer continuity does not affect response rates directly, it could have an impact on interviewers' feelings of vulnerability and morale (for example, the data in Chapter 10 suggested that interviewers, on average, preferred to contact their former respondents first). Chapter 8 also described the essential difficulties in trying to measure interviewer continuity effects with respect to a complete indicator of nonresponse.

Implications for Interviewer Training. The analysis in Chapter 6 suggests that it is not the easily measured characteristics of the interviewer that are having an impact on survey response rates. Possible sources of this variation are interviewer expectations and interviewer behaviour on the doorstep. Past research (such as Morton-Williams, 1993; Groves and Couper, 1994, 1996; and Campanelli, Sturgis, and Purdon, 1997) have suggested the key importance of this latter area, which implies the need to train interviewer explicitly with respect to their doorstep introductions.

From Chapter 8 it was seen that BHPS interviewers offered many resourceful and creative strategies for minimising any potential strain due to interviewer discontinuity. These strategies could be discussed with panel study interviewers. For example, as a minimum interviewers could be given the name of the previous interviewer, instructed to mention

the name, and instructed to give regards from the previous interviewer and explain why the previous interviewer is not there.

Measuring Interviewer Effects. The large values of ρ_i on particular items and the fact that ρ_i is of the same order of magnitude as ρ_s suggests that survey organisations should attempt to incorporate measurement of ρ_i into their designs, when possible. (On long-term continuous surveys, the measurement of interviewer variance should be a definite consideration). As a minimum, organisations should at least try to minimise the impact of interviewer variability on estimates by reducing interviewers' workloads. Current practice tends to favour smaller dedicated interviewer forces with large assignments; in the presence of substantial interviewer effects this is a misguided policy.

11.2.2 Implications for Analysis Strategies

A wide number of analysis techniques were used in this thesis. These ranged from chi-square tests of independence applied to simple two-way cross-tabulations to loglinear modelling for multi-way tables, hierarchical analysis of variance, and multiple logistic regression. However, essential to many of the analyses was the use of cross-classified multilevel logistic regression models and cross-classified multilevel multinomial regression models. For example, although hierarchical analyses of variance can be used to estimate interviewer variability, the use of cross-classified multilevel logistic regression models are essential for the accurate estimation of the random effects of interviewers and areas (see Chapters 6, 7, and 9). Multilevel multinomial regression is essential for looking at the covariance between categories of the dependent variable (e.g., refusals and non-contacts) across random terms (e.g., within interviewers or within areas) (see Chapter 7). Multilevel modelling allows for the incorporation of the complex variance-covariance structure (present in almost all survey data due to design and

implementation) directly into the substantive analyses (see Chapters 6, 7, and 9). Different conclusions can be reached when this complex structure is ignored (take, for example, the difference in conclusions about the significance of fixed effects in single-level versus cross-classified multilevel multinomial models (see Chapter 7), or the difference in conclusions about the impact of interviewer characteristics in explaining response variance under single level versus cross-classified multilevel logistic models (see Chapter 9). Lastly, multilevel models facilitate the direct incorporation of explanatory variables to help explain any observed random effects (see Chapters 6 and 9).

11.3 Limitations of this Research and Suggestions for Future Research

- The majority of the work on interviewer variance in the literature and in this thesis is dependant on ρ , the intraclass correlation coefficient. Yet, the values of ρ produced in any study are themselves estimates. We are reminded of this by the negative values of ρ which often appear. Ideally, the standard errors for ρ should be calculated (e.g., through a jack-knife approach) to aid the interpretation of the obtained values. This is true for this thesis as well as for future research.
- How best to estimate ρ in the case of a multilevel logistic model needs to be determined.
- The random effects of interviewers and areas at the household level did not reach significance. However, random effects for individual level terms of a comparable size did reach significance. This suggests that future research should be conducted with a larger sample of households. The magnitude of random effects found for interviewers and areas in the individual level models should be treated with some caution as household level

variation was not included. Thus, given these concerns, the findings in this thesis be considered as indicative rather than definite.

- No explicit cost models have been considered. A few guideline points are considered here. Real costs are involved with the implementation of an interpenetrated design, but these are relatively small compared with most survey budgets. The two main costs are researcher time and interviewer travel costs:
 - Researcher time is needed to design the interpenetrated sample, to convince the interviewers and other staff to participate, and to actually implement the interviewer/PSU allocations. In Wave 2 of a complex survey such as the BHPS this took approximately 7 to 10 researcher days.
 - Under a design in which households are assigned to interviewers within geographic pools, extra travel costs can be greatly minimised. For the BHPS, geographic pools were only formed of PSUs whose centroids were no more than 10 kilometres apart. Thus, depending on where the interviewer lives in relation to the PSUs, little or no extra travel costs may occur. The BHPS costs of implementing the interpenetrated sample design at Wave 2 were only marginally greater than the costs of a non-interpenetrated design. It should also be remembered that in an interpenetrated telephone survey travel costs do not apply.
- In Chapter 9, this thesis only considered the influence of interviewers on individual items, and not on scales as was done by (O’Muircheartaigh and Wiggins, 1981). Future analysis of the BHPS data could consider this.

- The results of this thesis only generalise to the 153 PSUs of the BHPS whose centroid was no more than 10 kilometres away from any other PSU. As noted in Section 1.4.11, there is a slight urban bias in the 153 PSUs as compared to the full sample of 250 PSUs.
- The interviewers in this study are assumed to be from a ‘superpopulation’ of interviewers who work on social surveys in the UK. This is not a bad assumption given interviewers in the UK are typically free-lance agents, many of whom work for several companies (such as NOP Social and Political, Social and Community Planning Research, and the Office of National Statistics, among others). However, this is just one study and some of the unexpected findings with respect to interviewer characteristics (such as age and experience) are probably best explained by the particular interviewers who were included in the study. Thus a replication would be useful.
- Given the results from this thesis about the typical profiles of persons who become refusals and non-contacts, specific targeted strategies for such individuals need to be researched and tested.
- With respect to interviewer continuity, the current data could be analysed further to understand the interviewer continuity patterns of those households and individuals who were adamant nonresponse cases and not reissued in later stages of fieldwork. Further experimental research on the impact of interviewer continuity on both response rates and response quality would be advisable to determine if the interviewer continuity findings in this thesis replicate to other interviewers, other types of surveys, and other types of populations. Such research would need to carefully document interviewer identification numbers and refusal conversion processes.

- Ideally one would like a research design in which all the elements of all the cells in the theoretical model in Figure 14a are present, including data on all aspects of *Respondents*, all aspects of *Interviewers* and explicit doorstep behaviour data for the *Respondent/Interviewer Interaction* cells, in addition to considering all factors in the *Survey Design* cell and all factors in the *Survey Context* cell. Construction of such a design would be both a challenging and expensive task. One challenge is presented by the *Survey Design* cell. Replications (or random portions) of the survey would be needed for each of the various design elements to be tested. Thus designs with large lists of *Survey Design* elements become unpractical and one could argue that perhaps the best way to obtain a ‘complete’ picture of the impact of the elements of this cell is with a meta-analysis of other research. As described in Section 5.2, the *Survey Context* cell presents even a bigger challenge as its elements usually cannot be manipulated experimentally. One is left with replicating the survey over various locations or time-frames known to vary by the elements of interest. Unfortunately, as no experimental manipulation is possible various confounding factors would need to be considered.
- In contrast, it should be relatively feasible to implement a simultaneous design which included an interpenetrated design, all the elements of the *Respondent’s* cell, all the elements of the *Interviewer’s* cell, area information from Census small area data, keyed call record information, and a sample of tape-recorded doorstep introductions. Although not obtained from a single simultaneous design, all of these elements were present in the ESRC project summarised in the report by Campanelli, Sturgis, and Purdon (1997).

- Once data from all the cells of either a full or reduced theoretical model are present, then one could investigate how best to calculate the coefficients of the paths in the model (see sub-section entitled, “Summarising the Results with Respect to the Reduced Theoretical Model” in Section 11.1.3).
- Many other elements could also be added to Figure 14a, such as the presence of absence of an advanced letter in the *Survey Design* cell, having interviewers reporting on their calling strategies in the *Interviewer* cell, and so on.
- More importantly future research with respect to the interviewer, should always be motivated from the perspective of total survey error, as the interviewer is just one component of a highly interactive survey process. In particular, trade-offs between different sources of error should be kept in mind. For example, looking at ways to train the interviewer to reduce nonresponse error can be misguided if respondents are pressurised or angered by the process and therefore create lots of response errors. Or, advocating the use of flexible interviewing (see Footnote 5) may increase response quality from the standpoint of respondents but at the same time increase interviewer variability. Or testing a design which minimises interviewer error by minimising the presence of the interviewer can also increase response errors and item nonresponse.

REFERENCES

- Agresti, A. (1990), *Categorical Data Analysis*, New York: John Wiley & Sons.
- Ajzen, I., and Fishbein, M. (1974), Attitudes Toward Objects as Predictors of Single and Multiple Behavioural Criteria, *Psychological Review*, 81, 59-74.
- Andersen, R., Kasper, J. and Frankel, M.R. (1979), A Model of Survey Error, Chapter 1, in: R. Andersen, J. Kasper, M.R. Frankel and associations (eds), *Total Survey Error*, San Francisco: Jossey-Bass.
- Anderson, B.A., Silver, B.D., and Abramson, P. (1988a), The Effects of Race of Interviewer on Measures of Electoral Participation by Blacks in SRC National Election Studies, *Public Opinion Quarterly*, 52(1), 53-83.
- Anderson, B.A., Silver, B.D., and Abramson, P. (1988b), The Effects of the Race of Interviewer on Race-Related Attitudes of Black Respondents in SRC/CPS National Election Studies, *Public Opinion Quarterly*, 52(3), 289-324.
- Aldrich, J.H. and Nelson, F.D. (1984), *Linear Probability, Logit, and Probit Models*, Newbury Park, CA: Sage.
- Alvarez, R.M. and Nagler, J. (1998), Estimating Models of Multiparty Elections, *American Journal of Political Science*, 42(1), 55-96.
- Argyle, M. (1973), *Social Interaction*, London: Tavistock.

Bailer, B. and Rothwell, N. (1984), Measuring Employment and Unemployment, in: C.F. Turner and E. Martin (eds), *Surveying Subjective Phenomena*, New York: Russell Sage Foundation.

Bailey, K.D. (1978), *Methods of Social Research*, New York: The Free Press, Macmillan.

Bailey, L., Moore, T.F., and Bailer, B.A. (1978), An Interviewer Variance Study for the Eight Impact Cities of the National Crime Survey Cities Sample, *Journal of the American Statistical Association*, 73, 16-23.

Barber, T.X. and Silver, M.J. (1972), Fact, Fiction, and the Experimenter Bias Effect, in: A.G. Miller (ed), *The Social Psychology of Psychological Experiments*, New York: Free Press.

Barr, A. (1957), Differences Between Experienced Interviewers, *Applied Statistics*, 6, 180-188.

Beatty, P. (1995), Understanding the Standardized/Non-Standardized Interviewing Controversy, *Journal of Official Statistics*, 11, 147-160.

Bebbington, A. (1970), The Effect of Nonresponse in the Sample Survey with an Example, *Human Relations*, 23, 169-180.

- Begg, C.B. and Gray, R. (1984), Calculation of Polychotomous Logistic Regression Parameters using Individualized Regressions, *Biometrika*, 71(1), 11-18.
- Belson, W. (1981), *The Design and Understanding of Survey Questions*, London: Gower.
- Bergman, L., Hanve, R., and Rapp, J. (1978), Why Do Some People Refuse to Participate in Interview Surveys?, *Sartryck ur Statistisk Tidskrift*, 5, 341-356.
- Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N., Sudman, S. (eds), (1991), *Measurement Errors in Surveys*, New York: John Wiley & Sons.
- Biemer, P., and Stokes, S.L. (1991), Approaches to the Modeling of Measurement Error, in: P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman (eds), *Measurement Errors in Surveys*, New York: John Wiley & Sons.
- Blankenship, A.B. (1940), The Effect of the Interviewer Upon the Response in a Public Opinion Poll, *Journal of Consulting Psychology*, 4, 134-136.
- Bohrstedt, G.W. (1983), Measurement, in: Rossi *et al* (eds), *Handbook of Survey Research*, New York: Academic Press.
- Bradburn, N. (1992), A Response to the Nonresponse Problem, 1992 AAPOR Presidential Address, in: *Public Opinion Quarterly*, 56(3), 391-397.

Bradburn, N.M. and Sudman, S. (1979), *Improving Interviewing Methods and Questionnaire Design*, San Francisco: Jossey-Bass, 1979.

Bradburn, N.M. and Sudman, S. (1991), The Current Status of Questionnaire Design, in: P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman (eds), *Measurement Errors in Surveys*, New York: John Wiley & Sons.

Briggs, C.L. (1986), *Learning How to Ask: Sociolinguistic Appraisal of the Role of the Interview in Social Science Research*, Cambridge: Cambridge University Press.

Bryk, A.S., Raudenbush, S.W., Congdon, R. and Seltzer, M. (1986), *An Introduction to HLM: Computer Program and User's guide*, Chicago: University of Chicago, Department of Education.

Cahalan, D., Tamulonis, V., and Verner, H.W. (1947), Interviewer Bias Involved in Certain Types of Attitude Questions, *International Journal of Opinion and Attitude Research*, 1, 63-77.

Campanelli, P.C., Martin, E.A., and Rothgeb, J.M. (1991), The Use of Respondent and Interviewer Debriefing Studies as a Way to Study Response Error in Survey Data, *The Statistician*, 40, 253-264.

Campanelli, P., Sturgis, P. and Purdon, S. (1997), *Can You Hear Me Knocking: An Investigation into the Impact of Interviewers on Survey Response Rates*, London: SCPR.

Campanelli, P. and O'Muircheartaigh, C. (1999), Interviewers, Interviewer Continuity, and Panel Survey Nonresponse, *Quality and Quantity*, 33, 59-76.

Campbell, D.T. and Fiske, D.W. (1959), Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix, *Psychological Bulletin*, 56, 81-105.

Cannell, C.F., Marquis, K.H., and Laurent, A. (1977), A Summary of Studies of Interviewing Methodology: 1959-1970, *Vital and Health Statistics*, Series 2, 69, i-78.

Cannell, C.F. (1954), *A Study of the Effects of Interviewers' Expectations Upon Interviewing Results*, PhD thesis, Ohio State University.

Cannell, C.F., Fowler, F.J., and Marquis, K.H. (1968), The Influence of Interviewer and Respondent Psychological and Behavioral Variables on the Reporting in Household Interviews, *Vital and Health Statistics*, Series 2, 26, i-65.

Cannell, C.F., Miller, P.V., and Oksenberg, L. (1981), Research on Interviewing Techniques, in: S. Leinhardt (ed), *Sociological Methodology*, San Francisco: Jossey-Bass.

Carmines, E.G. and Zeller, R.A. (1979), *Reliability and Validity Assessment*, Beverly Hills, CA: Sage.

CASRO Task Force on Completion Rates (1982), On the Definition of Response Rates, Special Report, New York: Council of American Survey Research Organizations.

Cialdini, R.B., Braver, S.L., and Wolf, W.S. (1991), A New Paradigm for Experiments on the Causes of Survey Nonresponse, paper presented at the 2nd International Workshop on Household Survey Nonresponse, Washington, DC.

Cialdini, R.B., Braver, S.L., and Wolf, W.S. (1993), Predictors of Nonresponse in Government and Commercial Surveys, paper presented at the 4th International Workshop on Household Survey Nonresponse, Bath, UK.

Cochran, W.G. (1963), *Sampling Techniques*, 2nd Ed., New York: John Wiley & Sons.

Collins, M. (1980), Interviewer Variability: A Review of the Problem, *Journal of the Market Research Society*, 22, 77-95.

Collins, M. and Butcher, B. (1982), Interviewer and Clustering Effects in an Attitude Survey, *Journal of the Market Research Society*, 25(1), 39-58.

Colombo, R. (1983), Patterns of Non-Response, in: G. Hoinville (ed), *SSRC Survey Methods Seminar Series, 1980-83*, London: SCPR.

Converse, J., and Presser, S. (1986), *Survey Questions: Handcrafting the Standardized Questionnaire*, Sage Series No 63, Thousand Oaks, CA: Sage.

Converse, J., and Schuman, H. (1974), The Role of the Interviewer, Chapter 2, in: *Conversations at Random: Survey Research as Interviewers See It*, New York: John Wiley & Sons.

Couper, M. and Groves, R.M. (1992), The Role of the Interviewer in Survey Participation, *Survey Methodology*, 18, 263-277.

Couper, M. and Groves, R.M. (1993), Interviewer Questionnaire on Participation, presented at the 4th International Workshop on Household Survey Nonresponse, Bath, UK.

Couper, M. and Groves, R.M. (1996), Introductory Interactions in Telephone Surveys and Nonresponse, paper presented at the annual meeting of the American Association for Public Opinion Research, Salt Lake City, Utah.

Couper, M., Groves, R.M., and Raghunathan, T. (1996), Nonresponse in the Second Wave of a Longitudinal Survey, paper presented at the 7th International Workshop on Household Survey Nonresponse, Rome, Italy.

de Leeuw, E. (1992), *Data Quality in Mail, Telephone, and Face to Face Surveys*, Amsterdam: TT-Publikaties.

de Leeuw, E., and Hox, J. (1996), The Effect of the Interviewer on the Decision to Cooperate in a Survey of the Elderly, in: S. Laaksonen (ed), *International Perspectives on Nonresponse*; Proceedings of the 6th International Workshop on Household Survey Nonresponse, Helsinki: Statistics Finland.

de Leeuw, E., Hox, J., Snijders, G., and de Heer, W. (1997), Interviewer Opinions, Attitudes and Strategies Regarding Survey Participation and Their Effect on Response, paper

presented at the 8th International Workshop on Household Survey Nonresponse, ZUMA, Mannheim, Germany.

DeMaio, T. (1980), Refusals: Who, Where, and Why, *Public Opinion Quarterly*, 44, 223-233.

Deming, W.E. (1953), On a Probability Mechanism to Attain an Economic Balance Between the Resultant Error of Response and the Bias of Nonresponse, *Journal of the American Statistical Association*, 48, 743-772.

Dijkstra, W. and van der Zouwen, J. (1982), *Response Behaviour in the Survey-Interview*, London: Academic Press.

Dillman, D., Gallegos, T.G., and Frey, J.H. (1976), Reducing Refusal Rates for Telephone Interviews, *Public Opinion Quarterly*, 40(1), 66-78.

Duncan, G.J., and Kalton, G. (1987), Issues of Design and Analysis of Surveys Across Time, *International Statistical Review*, 55, 97-117.

Dunkelberg, W., and Day, G. (1973), Nonresponse Bias and Callbacks in Sample Surveys, *Journal of Marketing Research*, 10, 160-168.

Dunn, O.J. and Clark, V.A. (1974), *Applied Statistics: Analysis of Variance and Regression*, New York: John Wiley & Sons.

Durbin, J., and Stuart, A. (1951), Difference in Response Rates of Experienced and Inexperienced Interviewers, *Journal of the American Statistical Association*, 114, 163-195.

Ecob, R. and Jamieson, B. (1992), A Multilevel Analysis of Interviewer Effects on a Health Survey, in: A. Westlake, R. Banks, C. Payne, and T. Orchard (eds), *Survey and Statistical Computing*, Amsterdam: North-Holland.

Eisenhower, D., Mathiowetz, N.A., and Morganstein, D. (1991), Recall Error: Sources and Bias Reduction Techniques, in: P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman (eds), *Measurement Errors in Surveys*, New York: John Wiley & Sons.

Esbensen, F. and Menard, S. (1991), Interviewer-Related Measurement Error in Attitudinal Research: A Nonexperimental Study, *Quality and Quantity*, 25, 151-165.

Esposito, J. and Jobe, J. (1991), A General Model of the Survey Interaction Process, in: *Proceedings of the 1991 Annual Research Conference*, Washington DC: U.S. Bureau of the Census.

Erlich, J. and Reisman, D. (1961), Age and Authority with Interview, *Public Opinion Quarterly*, 25, 39-56.

Farrant, G., and O'Muircheartaigh, C. (1991), Components of Nonresponse Bias in the British Election Surveys, p. 235-249, in: A. Heath, J. Curtice, R. Jowell, S. Evans, J. Field, and S. Witherspoon (eds), *Understanding Political Change*, London: Pergamon Press.

Feather, J. (1973), *A Study of Interviewer Variance*, Saskatoon, Canada: Department of Social and Preventive Medicine, University of Saskatchewan.

Fellegi, I. P. (1964), Response Variance and Its Estimation, *Journal of the American Statistical Association*, 59, 1016-1041.

Fellegi, I.P. (1974), An Improved Method of Estimating the Correlated Response Variance, *Journal of the American Statistical Association*, 69, 496-501.

Ferber, R., and Wales, H. (1952), Detection and Correction of Interviewer Bias, *Public Opinion Quarterly*, 16, 107-127.

Fienberg, S.E. (1985), *The Analysis of Cross-Classified Categorical Data*, Second Edition, Cambridge, Massachusetts: The MIT Press.

Fishbein, M., and Ajzen, I. (1975), *Belief, Attitude, Intention and Behaviour: An Introduction to Theory and Research*, Reading, MA: Addison-Wesley.

Foster, K. (1994), Report on the 5th International Workshop on Household Survey Nonresponse, *Survey Methodology Bulletin*, 34, 28-30, London: ONS.

Foster, K. (1996), A Comparison of the Census Characteristics of Respondents and Nonrespondents to the 1991 Family Expenditure Survey (FES), *Survey Methodology Bulletin*, 38, 9-17, London: ONS.

Fowler, F.J. Jr., (1995), *Improving Survey Questions: Design and Evaluation*, Applied Social Research Methods Series Volume 38, Thousand Oaks, CA: SAGE Publications.

Fowler, F.J., Jr., and Mangione, T.W., (1990), *Standardized Survey Interviewing: Minimizing Interviewer-Related Error*, Newbury Park, CA: SAGE Publications, Inc.

Frankel, L.R. and Dutka, S. (1983), Survey Design in Anticipation of Nonresponse and Imputation, in: W.G. Madow and I. Olkin (eds), *Incomplete Data in Sample Surveys*, Vol. 3, *Proceedings of the Symposium*, New York: Academic Press.

Freeman, J. and Butler, E.W. (1976), Some Sources of Interviewer Variance in Surveys, *Public Opinion Quarterly*, 40, 79-91.

Gales, K.E., and Kendall, M.G. (1957), An Inquiry Concerning Interviewer Variability, *Journal of the Royal Statistical Society, Series A*, 120, 121-147.

Glenn, N.D. (1981), The Utility and Logic of Cohort Analysis, *Journal of Applied Behavioral Science*, 17, 247-57.

Goldstein, H. (1995), *Multilevel Statistical Models*, 2nd edition, London: Edward Arnold.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G., and Healy, M. (1998), *A User's Guide to MLwiN*, London: Multilevel Models Project, Institute of Education, University of London.

Goudy, W.J. (1976), Nonresponse Effects on Relationships Between Variables, *Public Opinion Quarterly*, 40, 360-369.

Goyder, J. (1987), *The Silent Minority*, Boulder, Colorado: Westview Press.

Gray, P.G. (1956), Examples of Interviewer Variability taken from Two Sample Surveys, *Applied Statistics*, 5, 73-85.

Gray, R., Campanelli, P., Deepchand, K., and Prescott-Clarke, P. (1996), Seven Years on: The Impact of Attrition on a Follow-up of the 1984/85 Health and Lifestyle Survey Sample, *The Statistician*, 45(2), 1-21.

Greene, W. (1997), *Econometric Analysis*, 3rd Edition, Upper Saddle River: Prentice Hall.

Groves, R.M.(1989), *Survey Errors and Survey Costs*, New York: John Wiley & Sons.

Groves, R.M. (1991), Measurement Error Across the Disciplines, in: P.P. Biemer, R.M.

Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman (eds), *Measurement Errors in Surveys*, New York: John Wiley & Sons.

Groves, R.M. and Cialdini, R. (1991), Toward a Useful Theory of Survey Participation, *Proceeding of the Section on Survey Research Methods*, Joint Statistical Meetings, Alexandria, VA: American Statistical Association.

Groves, R.M., Cialdini, R.B., and Couper, M.P. (1992), Understanding the Decision to Participate in a Survey, *Public Opinion Quarterly*, 56(4), 475-495.

Groves, R.M. and Couper, M. (1992), Correlates of Nonresponse in Personal Visit Surveys, paper presented at the 3rd International Workshop on Household Survey Nonresponse, Voorburg, The Netherlands.

Groves, R.M., and Couper, M. (1994), Householders and Interviewers: The Anatomy of Pre-interview Interactions, *SMP Working Paper No. 11*, Ann Arbor, MI: Survey Research Centre, University of Michigan.

Groves, R.M., and Couper, M. (1995), Theoretical Motivation for Post-Survey Nonresponse Adjustment in Household Surveys, *Journal of Official Statistics*, 11(1), 93-106.

Groves, R.M. and Couper, M. (1996), Contact-Level Influences on Cooperation in Face-to-Face Surveys, *Journal of Official Statistics*, 12(1), 63-83.

Groves, R.M. and Fultz, N. (1985), Gender Effects Among Telephone Interviewers in a Survey of Economic Attitudes, *Sociological Methods and Research*, 14, 31-52.

Groves, R.M. and Magilavy, L. (1981), Increasing Response Rates to Telephone Surveys: A Door in the Face for Foot-in the-Door? *Public Opinion Quarterly*, 45, 346-358.

Groves, R.M. and Magilavy, L. (1986), Measuring and Explaining Interviewer Effects in Centralized Telephone Surveys, *Public Opinion Quarterly*, 50, 251-256.

Groves, R.M. and Robinson, D. (1982), *Report on Callback Algorithms on CATI Systems*, Washington, DC: Report to the US Bureau of the Census.

Hamilton, L.E. (1993), *Statistics With STATA 3*, Belmont, CA: Duxbury.

Hansen, M.H., Hurwitz, W.N., and Bershad, M.A. (1961), Measurement Errors in Censuses and Surveys, *Bulletin of the International Statistical Institute*, 38(2), 359-74.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953), *Sample Survey Methods and Theory*, New York: John Wiley & Sons.

Hansen, M.H., Hurwitz, W.N., Marks, E.S., and Mauldin, W.P. (1951), Response Errors in Surveys, *Journal of the American Statistical Association*, 46, 147-190.

Hanson, R.H. and Marks, E.S. (1958), Influence of the Interviewer on the Accuracy of Survey Results, *Journal of the American Statistical Association*, 53, 635-55.

Harkess, S. and Warren, C. (1993), The Social Relations of Intensive Interviewing: Constellations of Strangeness and Science, *Sociological Methods and Research*, 21(3), 317-339.

Hartley, H.O., and Rao, J.N.K. (1978), Estimation of Nonsampling Variance Components in Sample Surveys, in: N.K. Namboodiri (ed), *Survey Sampling and Measurement*, New York: Academic Press.

Hedeker, D. (1993), MIXOR: A Program for Mixed-Effects Ordinal Probit and Logistic Regression, unpublished manual, Chicago: University of Illinois, Prevention Research Center.

Hole, G. (1993), Response Rate Components for Selected Surveys at Statistics Canada, tables presented at the 4th International Workshop on Household Survey Nonresponse, Bath, England.

Hosmer, D.W., Jr. and Lemeshow, S. (1989), *Applied Logistic Regression*, New York, John Wiley & Sons.

House, J., and Wolf, S. (1978), Effects of Urban Residence and Interpersonal Trust and Helping Behaviour, *Journal of Personality and Social Psychology*, 36(9), 1029-1043.

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

Hox, J., de Leeuw, E., and Kreft, I. (1991), The Effect of Interviewer and Respondent Characteristics on the Quality of Survey Data: A Multilevel Model, in: P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman (eds), *Measurement Errors in Surveys*, New York: John Wiley & Sons.

Hox, J., de Leeuw, E., and Vorst, H. (1995), Survey Participation as Reasoned Action; A Behavioural Paradigm for Survey Nonresponse?, paper presented at the 6th International Workshop on Household Survey Nonresponse, Helsinki, Finland.

Hyman, H., with Cobb, W.J., Feldman, J., Hart, C.W., and Stember, C.H. (1954, 1975), *Interviewing in Social Research*, Chicago: University of Chicago Press.

Imai, M. (1986), *Kaizen: The Key to Japan's Competitive Success*, New York: McGraw-Hill.

Kahn, R.L., and Cannell, C.F. (1957), *The Dynamics of Interviewing*, New York: John Wiley & Sons.

Kalton, G. (1979), Ultimate Cluster Sampling, *Journal of the Royal Statistical Society*, 142(2), 210-222.

Kalton, G. (1993), Panel Surveys: Adding the Fourth Dimension, in: *Symposium 92: Design and Analysis of Longitudinal Surveys*, Ottawa: Statistics Canada.

Kalton, G., Lepkowski, J., Heeringa, S., Lin, T., and Miller, M. (1987), The Treatment of Person-Wave Nonresponse in Longitudinal Surveys, Survey of Income and Program Participation Working Paper Number 8704, Washington, DC: U.S. Bureau of the Census.

Kalton, G., Lepkowski, J., Montanari, G.E., and Maligalig, D. (1990), Characteristics of Second Wave Nonrespondents in a Panel Survey, *Proceedings of the Section on Survey Research Methods*, Alexandria, VA: American Statistical Association.

Kalton, G. and Schuman, H. (1982), The Effect of the Question on Survey Responses: A Review, *Journal of the Royal Statistical Society, Series A*, 145(1), 42-73.

Kasprzyk, D., Duncan, G., Kalton, G. and Singh, M. (1989), *Panel Surveys*, New York: John Wiley & Sons.

Katz, D. (1942), Do Interviewers Bias Poll Results?, *Public Opinion Quarterly*, 6, 248-268.

Kemsley, W.F.F. (1965), Interviewer Variability in Expenditure Surveys, *Journal of the Royal Statistical Society, Series A*, 128, 118-139.

Kendall, M. and Buckland, W. (1982), *A Dictionary of Statistical Terms*, 4th Edition, London: Lungman Group Ltd.

Kish, L. (1962), Studies of Interviewer Variance for Attitudinal Variables, *Journal of the American Statistical Association*, 57, 92-115.

Kish, L. (1965), *Survey Sampling*, New York: John Wiley & Sons.

King, J. (1996), Use of Geo-demographic Coding Schemes for Understanding Nonresponse, paper presented at the 7th International Workshop on Household Survey Nonresponse, Rome, Italy.

Kulka, R.A. and Weeks, M.F. (1988), Towards the Development of Optimal Calling Protocols for Telephone Surveys: A Conditional Probabilities Approach, *Journal of Official Statistics*, 4(4), 319-332.

Laurie, H., Smith, R., and Scott, L. (1997). *Strategies for Reducing Nonresponse in a Longitudinal Panel Survey*, Working Paper 97-12, Wivenhoe Park, UK: ESRC Research Centre on Micro-Social Change, University of Essex.

Lee, R. M. (1993), *Doing Research on Sensitive Topics*, London: Sage

Lehtonen, R. (1995), Interviewer Attitudes and Unit Nonresponse in Two Different Interviewing Schemes, paper presented at the 6th International Workshop on Household Survey Nonresponse, Helsinki, Finland.

Lessler, J. and Kalsbeek, W. (1992), *Nonsampling Error in Surveys*, New York: John Wiley & Sons

Lievesley, D. (1986), Unit Nonresponse in Interview Surveys, JCSM unpublished paper, London: SCPR.

Lindstrom, H. (1983), Nonresponse Errors in Sample Surveys, *Urv. Statist. Centbyr.*, 16.

Long, J.S. (1983), *Covariance Structure Models: An Introduction to LISREL*, Volume 34 in the Quantitative Applications in the Social Sciences series, Thousand Oaks, CA: SAGE Publications.

Longford, N.T. (1988), *VARCL Manual*, Princeton, New Jersey: Educational Testing Service.

Lord, F.M. and Novick, M.R. (1968), *Statistical Theories of Mental Test Scores*, Reading, Massachusetts: Addison-Wesley.

Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., and Trewin, D. (1997), *Survey Measurement and Process Quality*, New York: John Wiley & Sons.

Lynn, P. (1995), Evidence of the Inaccuracy of Quota Samples, Report of the First Cathie Marsh Memorial Seminar, November 1994, on Quota Versus Probability Sampling, *Survey Methods Centre Newsletter*, 15(1), 20-23.

Lynn, P. (1996), Who Responds to the British Crime Survey?, paper presented at the Fourth International ISA Conference on Social Science Methodology, University of Essex, UK.

Lynn, P. and Lievesley, D. (1991), *Drawing General Population Samples in Great Britain*, London: Social and Community Planning Research.

Lynn, P. and Purdon, S. (1994). *An Analysis of Factors Related to Nonresponse in a Postal Survey of Young People*, paper presented at the Royal Statistical Society's International Conference, Newcastle upon Tyne, UK.

Magnusson, D. and Bergman, L.R. (1990), *Data Quality In Longitudinal Research*, Cambridge: Cambridge University Press.

Mahalanobis, P.C. (1946), Recent Experiments in Statistical Sampling in the Indian Statistical Institute, *Journal of the Royal Statistical Society*, 109, 325-70.

Marquis, K.H. (1977), Survey Response Rates: Some Causes and Correlates, background paper prepared for a session on Response Rates at the Biennial Conference on Health Survey Research Methods, Williamsburg, Virginia.

Marquis, K.H. and Cannell, C.F. (1969), *A Study of Interviewer-Respondent Interaction in the Urban Employment Survey*, Research Report, Ann Arbor, Michigan: Survey Research Center, University of Michigan.

Mathiowetz, N.A. (1992), A Behavioural Paradigm for Understanding Nonresponse to the 1990 Census, paper presented at the Annual Meeting of the American Association for Public Opinion Research, St. Petersburg, Florida.

Mathiowetz, N.A. and Duncan, G.J. (1988), Out of Work, Out of Mind: Response Errors in Retrospective Reports of Unemployment, *Journal of Business and Economic Statistics*, 6.

Maynard, D.W., Schaeffer, N.C., and Cradock, R.M. (1993), *Declinations of the Request to Participate in the Survey Interview*, report of Joint Statistical Agreement 90-45 with the U.S. Bureau of the Census. Washington, DC: US Bureau of the Census.

McCrossan, L. (1991), *A Handbook for Interviewers*, Chapters 4 & 5, London: HMSO.

McCrossan, L. (1993), Respondent-Interviewer Interactions in Survey Introductions, paper presented at the 4th International Workshop on Household Survey Nonresponse, Bath, UK.

McCullagh, P. and Nelder, J.A. (1983), *Generalized Linear Models*, London: Chapman and Hall.

McKenzie, J.R. (1977), An Investigation into Interviewer Effects in Market Research, *Journal of Marketing Research*, 14, 330-336.

Menard, S. (1995), *Applied Logistic Regression Analysis*: Thousand Oaks, CA: Sage.

Mishler, E.G. (1986), *Research Interviewing: Context and Narrative*, Cambridge, MA: Harvard University Press.

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

Neter, J. Wasserman, W., and Kutner, M. (1985), *Applied Linear Statistical Models*, Homewood, Illinois: Richard D. Irwin, Inc.

Office of Population Censuses and Surveys (1990), *Standard Occupational Classification*, Volumes 1 and 2, London: HMSO.

Oksenberg, L. and Cannell, C.F. (1977), Some Factors Underlying the Validity of Response in Self-Report, *International Statistical Bulletin*, 48, 324-346.

Oksenberg, L., Coleman, L., and Cannell, C. (1986), Interviewer Voices and Refusal Rates in Telephone Surveys, *Public Opinion Quarterly*, 50, 97-111.

Oksenberg, L., Vinokur, A. and Cannell, C.F. (1979), The Effects of Instructions, Commitment and Feedback on Reporting in Personal Interviews, in: C.F. Cannell, L. Oksenberg, and J.M. Converse (eds), *Experiments in Interview Techniques*, Ann Arbor, Michigan: Survey Research Center, University of Michigan.

O'Malley, P.M., Bachman, J.G., and Johnston, L.D. (1984), Period, Age and Cohort Effects on Substance Use Among American Youth, 1976-82, *American Journal of Public Health*, 74, 682-688.

O'Muircheartaigh, C.A. (1976), Response errors in an Attitudinal Sample Survey, *Quality and Quantity*, 10, 97-115.

O'Muircheartaigh, C.A. (1977), Response Errors, Chapter 7, in: C.A. O'Muircheartaigh, and C. Payne (eds), *The Analysis of Survey Data*, Vol 2.

O'Muircheartaigh, C.A. (1982), *Methodology of the Response Errors Project*, Scientific Report Number 28, Voorburg, Netherlands: International Statistical Institute.

O'Muircheartaigh, C.A. (1984a), *The Magnitude and Pattern of Response Variance in the Peru Fertility Survey*, WFS Scientific Report No. 45, The Hague: International Statistical Institute.

O'Muircheartaigh, C.A. (1984b), *The Magnitude and Pattern of Response Variance in the Lesotho Fertility Survey*, WFS Scientific Report No. 70, The Hague: International Statistical Institute.

O'Muircheartaigh, C. and Campanelli, P. (1998), Interviewer Effects and Sample Design Effects: Relative Magnitude and Treatment, *Journal of the Royal Statistical Society*, 161(1), 63-78.

O'Muircheartaigh, C. and Campanelli, P. (1999), A Multilevel Exploration of the Role of Interviewers in Survey Nonresponse, submitted to the *Journal of the Royal Statistical Society*, Series A, November of 1998.

O'Muircheartaigh, C.A. and Wiggins, R.D. (1981), The Impact of Interviewer Variability in an Epidemiological Survey, *Psychological Medicine*, 11, 817-824.

O'Neil, M.J. (1979), Estimating the Nonresponse Bias Due to Refusals in Telephone Surveys, *Public Opinion Quarterly*, 43, 218-232.

O'Neil, M.J. (1980), Telephone Interview Introductions and Refusal Rates: Experiments in Increasing Respondent Cooperation, proceeding of the Survey Research Methods Section of the American Statistical Association, Alexandria, VA: American Statistical Association.

Pavalko, R., and Lutterman, K. (1973), Characteristics of Willing and Reluctant Respondents, *Pacific Sociological Review*, 16(4), 463-476.

Purdon, S., Campanelli, P., and Sturgis, P. (1998), Interviewers' Calling Strategies on Face-to-Face Interview Surveys, *Journal of Official Statistics*, forthcoming.

Rasbash, J., Woodhouse, G., Goldstein, H., Yang, M., Howarth, J., and Plewis, I. (1995), *MLn Software*, London: Multilevel Models Project, Institute of Education, University of London.

Rendtel, U. (1990), Teilnahmebereitschaft in Panelstudien: Zwischen Beeinflussung, Vertrauen and Sozialer Selektion, *Kolner Zeitschrift fur Soziologie und Sozialpsychologie*, 42(2), 280-299.

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

Rope, D. (1993), Preliminary Longitudinal Nonresponse Research with the CPS and CE, paper presented at the 4th International Workshop on Household Survey Nonresponse, Bath, UK.

Rose, D., Buck, N., Busfield, J., Corti, L., Crewe, I., Keen, M., Kemp, G., Laurie, H., Marsden, D., Price, S., Scarbrough, E., Scott, A., Scott, J., Shorrocks, T., and Sullivan, O. (1991), *Micro-Social Change in Britain: An Outline of the Role and Objectives of the British Household Panel Study*, Working Paper 1, Colchester: ESRC Research Centre on Micro-Social Change.

Rosenthal, R. (1966), *Experimenter Effects in Behavioral Research*, New York: Appleton-Century-Crofts.

Schaeffer, N.C. (1991), Conversation with a Purpose – or Conversation? Interaction in the Standardised Interview, in: P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman (eds), *Measurement Errors in Surveys*, New York: John Wiley & Sons.

Schleiffer, S. (1986), Trends in Attitudes Toward Survey Participation in Survey Research, *Public Opinion Quarterly*, 50, 17-26.

Schober, M.F. and Conrad, F.G. (1997), Does Conversational Interviewing Reduce Survey Measurement Error?, *Public Opinion Quarterly*, 61(4), 576-602.

Schuman, H. and Converse, J.M. (1971), The Effects of Black and White Interviewers on Black Responses in 1968, *Public Opinion Quarterly*, 35, 44-68.

Schuman, H. and Presser, S. (1977), Question Wording as an Independent Variable in Survey Analysis. *Sociological Methods and Research*, 6, 27-46.

Schuman, H. and Presser, S. (1981), *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*, New York: Academic Press.

Shapiro, M.J. (1970), Discovering Interviewer Bias in Open-Ended Survey Responses, *Public Opinion Quarterly*, 34, 412-415.

Silberstein, A. (1994), CE-Census Comparisons of Household Characteristics, paper presented at the 5th International Workshop on Household Survey Nonresponse, Ottawa, Canada.

Simonoff, J. (1993), The Relative Importance of Bias and Variability in the Estimation of the Variance of a Statistic, *The Statistician*, 42, 3-7.

Singer, E., Frankel, M. and Glassman, M. (1983), The Effect of Interviewer Characteristics and Expectations on Response, *Public Opinion Quarterly*, 47, 68-83.

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

Smith, W. and Krosnick, J. (1995), Temporal Inconsistency and Cross-Sectional Inconsistency in Surveys: Tests of Satisficing Hypotheses, unpublished report, available from J. Krosnick at the Ohio State University.

Snyder, M. (1979), Self-Monitoring Processes, in: Berkowitz (ed), *Advances in Experimental Social Psychology*, Volume 12, New York: Academic Press.

Social and Community Planning Research (1995), *Interviewers' Manual*, London: Social and Community Planning Research.

Spaeth, M.A. (1992), Response Rates at Academic Survey Research Organizations, *Survey Research*, 23(3-4), 18-20.

Spiro, R.L., and Weitz, B.A. (1990), Adaptive Selling: Conceptualization, Measurement, and Nomological Validity, *Journal of Marketing Research*, XXVII, 61-69.

Steeth, C.G. (1981), Trends in Nonresponse Rates 1952-1979, *Public Opinion Quarterly*, 45, 40-57.

Suchman, L. and Jordan, B. (1990), Interactional Troubles in Face-to-Face Survey Interviews, *Journal of the American Statistical Association*, 85(409), 232-253.

Sudman, S. (1976), *Applied Sampling*, New York: Academic Press.

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

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

Sudman, S., Bradburn, N.M., and Schwarz, N. (1996), *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*, San Francisco: Jossey-Bass Publishers.

Sukhatme, P.V. (1954), *Sampling Theory of Surveys with Application*, Ames, Iowa: Iowa State College Press.

Survey Research Center (1976), *Interviewer's Manual*, Revised Edition, Ann Arbor, Michigan: Institute for Social Research, University of Michigan.

Sturgis, P. and Campanelli, P. (1998), The Scope for 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.

Swires-Hennessy, E. and Drake, M. (1992), The Optimum Time at Which to Conduct Interviews, *Journal of the Market Research Society*, 34(1), 61-72.

- Taylor, A. (1994), Sample Characteristics, Attrition and Weighting, in: N. Buck, J. Gershuny, D. Rose, and J. Scott (eds), *Changing Households: The British Household Panel Survey, 1990-1992*, Colchester, England: ESRC Centre on Micro-social Change, University of Essex.
- Triplett, T., Blair, J., Hamilton, T., Kang, Y.C. (1996), Initial Cooperators vs. Converted Refusers: Are There Response Behaviour Differences? paper presented at the 51st annual conference of the American Association for Public Opinion Research, Salt Lake City, Utah.
- Tourangeau, R. (1984), Cognitive Sciences and Survey Methods, in: T. Jabine, M. Straf, J. Tanur, and R. Tourangeau (eds), *Cognitive Aspects of Survey Methodology: Building a Bridge Between the Disciplines*, Washington, DC: National Academy Press.
- Turner, C.F. and Martin, E. (eds) (1984), *Surveying Subjective Phenomena*, New York: Russell Sage Foundation.
- van de Pol, F.J.R. (1989), *Issues of Design and Analysis of Panels*, Amsterdam: Sociometric Research Foundation.
- van der Putte, B. (1991), 20 Years of the Theory of Reasoned Action of Fishbein and Ajzen: A meta-analysis, unpublished paper, The Netherlands: University of Amsterdam.
- Vehovar, V. (1996), The Substitution Procedure for the Unit Nonresponse, paper presented at the 5th International Conference on Household Survey Nonresponse, Ottawa, Canada.

Vehovar, V. and Zaletel, M. (1995), The Matching Project in Slovenia: Who are the Nonrespondents? paper presented at the 6th International Workshop on Household Survey Nonresponse, Helsinki, Finland.

Verma, V., Scott, C., and O'Muircheartaigh, C. (1980), Sample Designs and Sampling Errors for the World Fertility Survey, *Journal of the Royal Statistical Society, Series A*, 143(4), 431-473.

Vigerhous, G. (1981), Scheduling Telephone Interviews: A Study of Seasonal Patterns, *Public Opinion Quarterly*, 45, 250-259.

Waterton, J., and Lievesley, D. (1987), Attrition in a Panel Study of Attitudes, *Journal of Official Statistics*, 3, 267-282.

Weaver, C., Holmes, S., and Glenn, G. (1975), Some Characteristics of Inaccessible Respondents in a Telephone Survey, *Journal of Applied Psychology*, 60(2), 260-262.

Weber, D. and Burt, R.C. (1972), *Who's Home When*, Washington, D.C.: US Bureau of the Census.

Weeks, M., Jones, B.L., Folsom, R.E., and Benrud, C.H. (1980), Optimal Times to Contact Sample Households, *Public Opinion Quarterly*, 44(1), 101-114.

- Weeks, M.F., Kulka, F.A., and Pierson, S.A. (1987), Optimal Call Scheduling for a Telephone Survey, *Public Opinion Quarterly*, 51, 540-549
- Weiss, C.H. (1968), Validity of Welfare Mothers' Interview Responses, *Public Opinion Quarterly*, 32, 622-633.
- Wiggins, R.D. (1979), Sample Design for West London Survey of Aircraft Noise, in: A. Tarnopolsky and J. Morton-Williams (eds), *Aircraft Noise and Psychiatric Morbidity*, 12-29, SCPR: London.
- Wiggins, R.D. (1985), *A Replicated Study of the Impact of Interviewer Variability in a Community Survey of Physically Handicapped in an Inner London Borough*, Research Working Paper No. 24, PCL.
- Wiggins, R.D., Longford, N., and O'Muircheartaigh, C.A. (1992), A Variance Components Approach to Interviewer Effects, in: A. Westlake, R. Banks, C. Payne, and T. Orchard (eds), *Survey and Statistical Computing*, Amsterdam: North-Holland.
- Wilcox, J. (1977), The Interaction of Refusal and Not-At-Home Sources of Nonresponse Bias, *Journal of Marketing Research*, 14, 592-597.
- Wilson, J. (1997), Executing Schrödinger's Cat, *Popular Mechanics*, October, 34-35.
- Winer, B.J. (1962), *Statistical Principles in Experimental Design*, New York: McGraw-Hill.

Wiseman, F. and McDonald, P. (1980), *Toward the Development of Industry Standards for Response and Nonresponse Rates*, Report 80-101, Cambridge, MA: Marketing Science Institute.

Woodhouse, G. (1995), *A Guide to MLn for New Users*, London: Multilevel Models Project, Institute of Education, University of London.

Zarkovich, S.S. (1963), *Sampling Methods and Censuses; Vol. II, Quality of Statistical Data*, Rome: FAO.

APPENDIX A: Comparison of Interpenetrated Sub-Sample to Full British Household Panel Study Sample at Wave 2 §

Variable	Category Description	Full Sample	Interpenetrated Sub-Sample
Selected Individual Level Characteristics			
Gender $\chi^2 = 0.31, p = .58$	1=Male	47.0	46.4
	2=Female	53.0	53.6
	<i>n</i>	9845	2433
Age $\chi^2 = 5.57, p = .06$	1=15-25	18.2	19.5
	2=26-59	58.7	59.4
	3=60-97	23.1 †	21.1 †
	<i>n</i>	9845	2433
Marital Status $\chi^2 = 6.55, p = .36$	0=Child under 16	0.3	0.2
	1=Married	56.7	55.1
	2=Living as couple	7.2	7.0
	3=Widowed	8.2	7.7
	4=Divorce	4.7	4.9
	5=Separated	1.7	2.0
	6=Never married	21.2 †	23.1 †
<i>n</i>	9845	2433	
Employment Status $\chi^2 = 7.95, p = .54$	1=Self-employed	7.7	7.9
	2=Employed	48.3	49.4
	3=Unemployed	5.8	5.6
	4=Retired	17.1	16.4
	5=Maternity leave	0.2	0.3
	6=Family care	11.3 †	9.8 †
	7=FT student, school	5.8	6.5
	8=LT sick, disabled	3.1	3.3
	9=Gvt training scheme	0.5	0.5
	10=Other	0.2	0.2
<i>n</i>	9844	2432	
Socio-Economic Grade $\chi^2 = 4.83, p = .09$	1=Employers, managers, professionals, and other non-manual	64.7 †	67.7 †
	2=Manual	27.3	24.9
	3=Other	8.1	7.4
	<i>n</i>	5803	1468

§ Note that the two samples are not independent so the χ^2 test statistics are attenuated and the tests are conservative.

† Adjusted standardised residuals are greater than 2.0 or less than -2.0 for this cell.

Appendix A: (continued) §

Variable	Category Description	Full Sample	Interpenetrated Sub-Sample
Selected Household Level Characteristics			
Tenure $\chi^2 = 5.85, p = .21$	1=Owned or on mortgage	67.7	68.4
	2=Shared ownership	0.2	0.2
	3=Rented	29.7	30.2
	4=Rent free	2.1 †	1.2 †
	5=Other	0.2	0.1
	<i>n</i>	5210	1281
Number of cars in household $\chi^2 = 2.67, p = .44$	0=None	29.7	31.7
	1=One	46.8	46.2
	2=Two	20.1	19.2
	3=Three+	3.5	2.9
	<i>n</i>	5155	1265
Number of pensioners in household $\chi^2 = 3.89, p = .14$	0=None	70.0 †	72.8 †
	1=One	20.8	18.9
	2=Two+	9.2	8.3
	<i>n</i>	5227	1283
Type of accommodation $\chi^2 = 59.48, p = .000$	1=Detached house/bungalow	19.6 †	13.4 †
	2=Semi-detached house	33.8	32.5
	3=End terraced house	7.2	6.4
	4=Terraced house	18.4	19.7
	5=Purpose built flat	14.1 †	20.4 †
	6=Converted flat	4.1 †	5.6 †
	7=Other	2.9	2.0
	<i>n</i>	5192	1278
Net monthly housing costs $\chi^2 = 6.47, p = .09$	0=None	30.9 †	28.0 †
	1=1-125 pounds	21.8	21.0
	2=126-250 pounds ‡	24.9	27.4
	3=251-6725 pounds ‡	22.4	23.7
	<i>n</i>	5204	1279

§ Note that the two samples are not independent so the χ^2 test statistics are attenuated and the tests are conservative.

† Adjusted standardised residuals are greater than 2.0 or less than -2.0 for this cell.

‡ When these two categories are combined, adjusted standardised residuals for the new combined cells are greater than 2.0 and less than -2.0.

APPENDIX B: Bivariate Correlates of Nonresponse: Nonresponse Rates and Logistic Regression Coefficients for Various Categories

Variable	Variable categories	Household Level ϕ				Individual Level								
		Non-respondents		Refusals		Non-respondents		Refusals		Non-contacts				
INTERVIEWER LEVEL VARIABLES														
		n	%	%	%	n	%	%	%	%	%			
GENDER	Interviewer gender													
	1=male	148	18	10	7	301	25	*	12	11				
	2=female	1324	12	7	5	2588	19		10	8				
AGE2	Interviewer age													
	1=26-38 yrs	135	13	10	4	265	21	**	11	9				
	2=39-50 yrs	644	11	6	5	1260	17		9	7				
	3=51-60 yrs	526	16	8	7	1028	23		12	9				
	4=61-72 yrs	167	12	7	5	336	19		10	8				
YRSWNOP2	Years working for NOP													
	1=0-2 yrs	397	11	***	7	3	***	766	18	***	10	**	7	**
	2=3-5 yrs	503	12		7	6		1015	18		10		7	
	3=6-7 yrs	241	15		9	6		466	24		13		11	
	4=8-10 yrs	137	6		3	3		278	15		7		7	
	5=11-13 yrs	85	19		5	14		154	22		7		15	
	6=14-20 yrs	109	21		12	9		210	27		16		10	
INTTYPE	Type of interviewer													
	1=standard interviewer	1154	12	***	6	*	5	2282	19	**	10		8	**
	2=supervisor	59	22		10		10	108	32		16		17	
	3=area manager	259	17		10		6	499	21		11		9	
SAMEINT	Same interviewer as wave 1													
	0=different	885	12		6		6	1497	13		7		6	
	1=same	518	11		6		4	912	13		7		6	
INDICATORS OF RESPONDENT CO-OPERATION AND CONTACTABILITY														
		n	β SE(β)	β SE(β)	β SE(β)		β SE(β)	β SE(β)	β SE(β)	β SE(β)	β SE(β)			
AIVNC	Total number of calls at W1	1472	.076 (.030)*	.006 (.040)	.163 (.042)***		.073 (.019)***	.059 (.024)*		.101 (.026)***				
PRO1NCW1	Prop of calls that were non-contacts at W1	1464	.912 (.268)	.369 (.352)	1.656 (.391)***		.602 (.168)***	.418 (.218)		.855 (.238)***				
		n	%	%	%		%	%		%	%			
TOTBR2	Whether broken appointment at W1													
	0=No broken appointments	1385	13	7	5	2710	19	10		8				
	1=At least one broken appointment	85	17	7	9	176	25	14		11				
		n	β SE(β)	β SE(β)	β SE(β)		β SE(β)	β SE(β)		β SE(β)	β SE(β)			
BIVLNC	Num. of calls at last known address (W2)	1472	-.028 (.038)	.060 (.045)	-.173 (.066)**		.044 (.022)*	.072 (.028)*		.014 (.033)				

ϕ Note that when individual level variables are analysed at the household level, they refer to the characteristics of the 'household reference person', the person in whose name the property is owned or rented.

* p < .05

** p < .01

*** p < .001

Variable	Variable categories	Household Level ϕ			Individual Level				
		Non-respondents	Refusals	Non-contacts	Non-respondents	Refusals	Non-contacts		
		n	β SE(β)	β SE(β)	β SE(β)	n	β SE(β)	β SE(β)	β SE(β)
TOTCNTW2	Number of calls until first contact at W2	1472	.225 (.043)***	.124 (.057)*	.270 (.055)***	2889	.132 (.031)***	.079 (.041)	.163 (.040)***
		n	%	%	%	n	%	%	%
AIVFHO3	Wave 1 final outcome 0=no internal HH refusal 1=internal HH refusal	1350 122	11 *** 40	6 *** 25	5 ** 11	2594 295	15 *** 61	6 *** 45	8 *** 14
		n	β SE(β)	β SE(β)	β SE(β)	n	β SE(β)	β SE(β)	β SE(β)
RATENR	Wave 1 PSU response rate		.014 (.007)*	.015 (.009)	.012 (.011)		.008 (.004)	.008 (.006)	.005 (.006)
		n	%	%	%	n	%	%	%
REFCON	W2 refusal conversion by different int 0=no conversion 1=conversion	1457 15	13 *** 40	7 *** 33	5 7	2861 28	19 ** 43	10 *** 32	8 7
AIV42	Co-operation of respondent 0=missing 1=very good 2=good/fair 3=poor/very poor	35 1092 234 15	14 *** 8 23 53	3 *** 5 13 33	11 *** 4 9 7	61 1971 396 18	13 *** 11 24 61	3 *** 6 13 44	10 *** 5 10 17
AIV52B	Completion of tracking schedule 0=missing 1=yes, completed 2=no, refusal/other	73 1189 114	12 *** 10 23	6 *** 6 15	7 5 6	122 2139 185	14 *** 13 27	7 *** 7 15	7 ** 5 11
AIV6A2	Problems affecting interview: eyesight 0=missing 1=yes 2=no	46 35 1295	9 * 26 11	2 ** 20 6	7 3 5	74 43 2329	7 *** 33 14	2 *** 21 7	5 2 6
AIV6B2	Problems affecting interview: hearing 0=missing 1=yes 2=no	47 41 1288	9 20 11	2 ** 17 6	6 2 5	75 51 2320	7 * 24 14	2 * 14 7	5 4 6
AIV6C2	Problems affecting interview: reading 0=missing 1=yes 2=no	47 30 1299	9 * 27 11	2 10 7	6 10 5	75 52 2319	7 *** 31 14	2 12 7	5 14 6
AIV6D2	Problems affecting interview: English 0=missing 1=yes 2=no	46 22 1308	9 *** 36 11	2 14 6	7 *** 23 5	74 47 2325	7 *** 36 14	2 *** 21 7	5 * 15 6

ϕ Note that when individual level variables are analysed at the household level, they refer to the characteristics of the 'household reference person', the person in whose name the property is owned or rented.

* $p < .05$

** $p < .01$

*** $p < .001$

Variable	Variable categories	Household Level ϕ			Individual Level				
		Non-respondents	Refusals	Non-contacts	Non-respondents	Refusals	Non-contacts		
HOUSEHOLD LEVEL									
		n	%	%	%	n	%	%	%
AHSTYPE2	Type of accommodation								
	0=missing	41	7 *	7	0 **	75	19	12	3 ***
	1=det. hse/bungalow	189	12	9	3	410	17	11	5
	2=semi/bung/end of terrace	535	11	7	4	1134	19	11	7
	3=terraced	306	12	6	7	617	19	10	8
	4=other	401	18	8	9	653	23	10	12
AHSPRBB2	Problems with home: damp								
	0=not a problem	1258	13	7	5	2472	19	10	8 *
	1=small/big problem	209	13	5	8	405	22	9	11
APHONE2	HH has telephone								
	0=no phone/missing	240	24 ***	10	13 ***	413	28 ***	12	14 ***
	1=phone	1232	11	7	4	2476	18	10	7
AHSROOM4	Number of rooms in accommodation								
	0=3+	1269	11 ***	7 *	4 ***	2602	19 ***	10	8 ***
	1=1-2	197	23	11	12	273	28	10	14
ANCARS2	Car/van available for private use								
	0=none	480	17 **	9	7	779	23 ***	11	10
	1=one	662	12	7	5	1311	20	11	8
	2=2+	324	8	5	3	785	15	8	7
ACD2USE	VCR in accommodation								
	0=no	405	17 **	9	7	624	21	11	8
	1=yes	1062	11	6	5	2253	19	10	8
ACD3USE	Freezer in accommodation								
	0=no	242	21 ***	12 **	8 *	366	25 **	13	10
	1=yes	1224	11	6	5	2510	19	10	8
ACD4USE	Washing machine in accommodation								
	0=no	205	22 ***	12 **	9 **	313	26 **	13	11 *
	1=yes	1261	11	6	5	2563	19	10	8
ACD5USE	Tumble dryer in accommodation								
	0=no	799	14	7	7 *	1459	19	10	8
	1=yes	661	11	7	4	1407	20	11	8

ϕ Note that when individual level variables are analysed at the household level, they refer to the characteristics of the 'household reference person', the person in whose name the property is owned or rented.

* p < .05

** p < .01

*** p < .001

Variable	Variable categories	Household Level ϕ			Individual Level				
		Non-respondents	Refusals	Non-contacts	Non-respondents	Refusals	Non-contacts		
		n	%	%	%	n	%	%	%
ACD6USE	Dish washer in accommodation								
	0=no	1252	13	7	6	2397	20 **	11 *	8
	1=yes	214	10	5	5	479	14	7	7
ACD8USE	Home computer in accommodation								
	0=no	1132	14 **	8 **	6	2141	22 ***	12 ***	9 *
	1=yes	334	8	4	4	735	12	6	6
ACD9USE	CD player in accommodation								
	0=no	1044	14	8	6	1925	21 *	11 *	8
	1=yes	422	10	6	5	951	17	9	8
AHSOWND3	House owned or rented								
	0=owned	955	10 ***	7	3 ***	1972	18 **	10	7 ***
	1=rented	512	18	8	9	906	22	10	11
AXPHSN3	Net monthly housing costs								
	0=none	389	11 *	8 ***	3	715	20 ***	11 ***	7
	1=1-125 pounds	318	16	10	5	594	24	16	7
	2=126-250 pounds	390	14	8	6	779	21	11	9
	3=251-1650 pounds	367	10	3	7	787	14	5	8
AIHTYPE3	Household type								
	1=all other households	834	14 *	7 *	6	1395	19 ***	9 ***	8
	2=households with dependent children	448	10	5	5	984	17	9	8
	3=h/holds with non-dependent children	190	16	11	5	510	28	17	10
ANCOUPL2	A couple present in HH								
	0=no	574	14	7	7	821	19	8 *	9
	1=yes	898	12	7	5	2068	20	11	8
ANKID2	Number of children in HH								
	0=none	1039	14 ***	8 *	6	1956	22 ***	12 **	9
	1=one or more	433	9	5	4	933	15	8	8
ANWAGE2	Number in HH of working age								
	0=0-3 people	1365	12 ***	7	5 *	2538	18 ***	9 ***	8 *
	1=4+ people	107	22	11	10	351	29	17	11
ANUE2	Number unemployed in HH								
	0=none	1282	12 *	7	4 ***	2462	18 ***	10 **	8 ***
	1=1+	189	19	5	12	424	28	14	14
ANPENS2	Number over pensionable age in HH								
	0=none	1084	13	6 ***	7 ***	2242	19	9 ***	9 ***
	1=1+	388	13	11	1	647	22	14	5

ϕ Note that when individual level variables are analysed at the household level, they refer to the characteristics of the 'household reference person', the person in whose name the property is owned or rented.

* p < .05

** p < .01

*** p < .001

Variable	Variable categories	Household Level ϕ			Individual Level					
		Non-respondents	Refusals	Non-contacts	Non-respondents	Refusals	Non-contacts			
INDIVIDUAL LEVEL										
		n	%	%	%	n	%	%	%	%
ASEX	Respondent gender									
	1= male	851	11	6	5	1198	19 ***	8	10	***
	2= female	566	13	7	5	1362	14	7	6	
AAGE2	Age at date of interview									
	1=15-25	152	24 ***	4 **	20 ***	476	22 ***	7 *	16	***
	2=26-59	868	10	5	4	1531	14	7	6	
	3=60-93	397	12	10	1	553	18	10	4	
ARACE3	Ethnic group membership									
	0=white	1299	11 ***	6	4 ***	2296	13 ***	7	5	***
	1=non-white	73	25	11	14	145	25	10	15	
AMASTAT2	Marital status									
	1=married/living as a couple	830	10	7	4 ***	1630	15 *	8	6	***
	2=widowed	191	12	9	2	209	18	9	5	
	3=divorced/separated	146	13	7	6	163	15	7	6	
	4=never married	250	16	4	11	557	20	7	13	
AQFEDH13	Highest educational qualification									
	0=higher, first, nursing, other higher	352	7 **	4 *	4	582	9 ***	4 **	5	
	1= A/O level, commercial, apprentice	1022	13	7	5	1861	15	8	6	
AOPRI.G13	Religion									
	1=no religion	544	11	4	7 **	935	14	6	7	**
	2=c of e /anglican	429	10	8	2	801	13	9	4	
	3=other	397	12	7	6	704	14	7	7	
ANORGM2	Number of organisations belonged to									
	0=none	582	14 **	8	6	1071	16 **	8	7	*
	1=1-8	791	9	5	4	1371	12	6	5	
		n	β SE(β)	β SE(β)	β SE(β)	n	β SE(β)	β SE(β)	β SE(β)	β SE(β)
GHQ	GHQ scale, higher values = more psychiatric symptoms	1313	.03 (.02)	.02 (.02)	.03 (.02)	2355	.02 (.01)*	.02 (.02)	.01 (.02)	
		n	%	%	%	n	%	%	%	%
AHGEST2	Employment status									
	1=working	826	11 *	5	5 ***	1489	15 ***	7	8	***
	2=unemployed	103	20	5	14	172	27	10	16	
	3=other	488	11	9	3	899	16	8	6	

ϕ Note that when individual level variables are analysed at the household level, they refer to the characteristics of the 'household reference person', the person in whose name the property is owned or rented.

* $p < .05$

** $p < .01$

*** $p < .001$

Variable	Variable categories	Household Level ϕ			Individual Level				
		Non-respondents		Refusals	Non-contacts		Non-respondents		Refusals
		n	%	%	%	n	%	%	%
AJBSOC2	0=not working	569	13	8	5	999	18 ***	8 **	7 *
	1=managers and administrators	130	9	5	4	193	12	5	8
	2=professional occupations	109	6	6	0	162	7	4	4
	3=associate professional & technical occupations	90	9	2	7	160	14	4	10
	4=clerical and secretarial occupations	112	12	5	6	298	10	6	4
	5=craft and related occupations	144	13	4	8	209	21	9	12
	6=personal and protective service occupations	70	6	1	4	169	14	6	8
	7=sales occupations	37	11	8	3	98	16	8	8
	8=plant and machine operatives	98	15	9	6	151	23	13	9
	9=other occupations	57	21	14	7	119	23	13	9
AJBGOLD2	Goldthorpe social class								
	0=NA	569	13 *	8	5	999	18 ***	8	7 **
	1=service class, higher	172	6	5	1	243	9	4	5
	2=service class, lower	176	9	3	6	312	13	5	8
	3=routine, non-manual	85	8	2	6	237	8	5	3
	4=personal service	27	11	11	0	104	14	10	5
	5=sml props w employ	27	19	7	11	39	21	5	15
	6=sml props w/out employ	65	12	6	6	99	20	9	11
	8=foreman, technical	78	6	4	3	117	15	8	7
	9=skilled manual worker	88	14	6	8	136	25	11	14
	10=semi, unskilled manual worker	120	20	11	9	255	21	10	10
AFIMN2	Total income last month								
	1=0-333	291	16	9	6	799	19 *	9 *	9
	2=334-666	371	13	8	5	626	15	8	6
	3=667-1000	244	11	6	5	426	17	9	8
	4=1001-1333	198	9	4	5	287	15	6	8
	5=1334+	307	9	5	4	410	12	4	7
AFIYRD12	Income last year from dividends/interest								
	1=nothing	601	16 ***	7 *	9 *	1088	17 ***	8	9 ***
	2=under 100 pounds	334	7	5	2	627	11	6	4
	3=100-1000 pounds	314	11	9	2	546	12	8	3
	4=1000+ pounds	127	5	3	2	185	8	4	3
AF1322	In receipt of income support								
	0=not mentioned	1194	11	7	4 ***	2190	13 *	7	5 ***
	1=income support	178	15	5	10	249	19	7	11

ϕ Note that when individual level variables are analysed at the household level, they refer to the characteristics of the 'household reference person', the person in whose name the property is owned or rented.

* p < .05

** p < .01

*** p < .001

Variable	Variable categories	Household Level ϕ			Individual Level			
		Non-respondents	Refusals	Non-contacts	Non-respondents	Refusals	Non-contacts	
GEOGRAPHIC AREA								
		β	SE(β)	β	SE(β)	β	SE(β)	
POPDENS	Population density % population density (persons per square kilometre x 1000)	.038 (.016) *	.016 (.023)	.049 (.021) *		-.027 (.011)*	.007 (.015)	.043 (.014)**
	Housing type							
PERCDET	% detached	-.017 (.006) **	-.008 (.007)	-.026 (.010) **		-.013 (.003)***	-.007 (.004)	-.020 (.005)***
PERCFLAT	% flat	.006 (.003) *	.001 (.004)	.008 (.004) *		-.006 (.003)*	.001 (.003)	.010 (.003)**
PERCOHOU	% other housing	-.000 (.003)	.003 (.004)	-.002 (.005)		.002 (.002)	.005 (.003)	-.002 (.003)
	No. of rooms							
PERRM12	% 1-2 rooms	.008 (.010)	-.005 (.014)	.019 (.014)		.006 (.006)	-.009 (.009)	.021 (.008)**
PERRM34	% 3-4 rooms	.016 (.006) **	.006 (.008)	.022 (.009) *		.011 (.004)**	-.005 (.005)	.017 (.005)**
PERRM5P	% 5+ rooms	-.009 (.004) *	-.002 (.006)	-.013 (.006) *		-.006 (.003)*	-.001 (.003)	-.011 (.004)**
	Car ownership							
PEROCAR	% no car	.016 (.005)**	.004 (.007)	.025 (.007)***		.012 (.003)***	.005 (.004)	.018 (.004)***
PERC1CAR	% 1 car	-.030 (.012)*	-.001 (.016)	-.051 (.017) **		-.024 (.007)**	-.013 (.010)	-.032 (.010)**
PERC2CAR	% 2+ cars	-.021 (.007)**	-.006 (.009)	-.036 (.012) **		-.015 (.004)***	-.005 (.005)	-.026 (.006)**
	Tenure							
PERCOO	% Owners	-.014 (.004)***	-.005 (.006)	-.019 (.006) **		-.010 (.003)***	-.005 (.003)	-.015 (.004)***
PERCLA	% LA	.015 (.005) **	.011 (.006)	.015 (.007) *		.013 (.003)***	.012 (.004)**	.012 (.004)**
PERCORNT	% other renters	.004 (.007)	-.012 (.010)	.018 (.010)		.001 (.004)	-.013 (.006)*	.016 (.006)**
	Household size							
PERC1P	% 1 person h/holds	.010 (.009)	-.009 (.012)	.026 (.012) *		.007 (.005)	-.006 (.007)	.022 (.007)**
PERC2P	% 2 person h/holds	-.068 (.026) *	-.058 (.034)	-.057 (.039)		-.046 (.016)**	-.032 (.020)	-.051 (.023)*
PERC3P	% 3 person h/holds	-.003 (.027)	.051 (.037)	-.056 (.038)		.006 (.017)	.040 (.023)	-.036 (.023)
PERC4P	% 4+ person h/holds	-.004 (.012)	.022 (.016)	-.030 (.018)		-.005 (.007)	.012 (.010)	-.026 (.011)*
	Household type							
PERCFO	% 0 families	.011 (.007)	-.004 (.010)	.024 (.010) *		.007 (.004)	-.003 (.006)	.017 (.006)**
PERCFLDC	% lone parent, dependent children	.075 (.020)***	.048 (.026)	.082 (.029) **		.060 (.013)***	.048 (.016)**	.066 (.017)***
PERCODC	% couple, dependent children	-.025 (.011) *	-.002 (.014)	-.044 (.016) **		-.015 (.007)*	-.003 (.009)	-.030 (.010)**
PERCOFAM	% other families	-.024 (.012) *	-.002 (.015)	-.042 (.017) **		-.018 (.007)**	-.005 (.009)	-.032 (.010)**
	Race							
PERCNWT	% non-white	.029 (.007)***	.015 (.009)	.036 (.009)***		.019 (.004)***	.014 (.005)**	.020 (.006)***
	Health							
PERCILL	% long term illness	.032 (.019)	.003 (.025)	.065 (.026) *		.036 (.011)**	.018 (.015)	.053 (.016)***
	Employment status							
PERCUNEM	% unemployed	.032 (.012)**	.005 (.016)	.054 (.017) **		.024 (.007)***	.010 (.009)	.037 (.010)***
	Social Grade							
PERCC1I	% AB	-.022 (.009) *	-.011 (.011)	-.034 (.014) *		-.016 (.005)**	-.010 (.007)	-.023 (.008)**
PERCC1	% C1	-.003 (.012)	-.006 (.015)	.003 (.017)		-.008 (.007)	-.012 (.009)	-.004 (.010)
PERCC2DE	% C2, DE	.011 (.006)	.007 (.008)	.014 (.009)		.009 (.004)**	.008 (.005)	.012 (.005)*

ϕ Note that when individual level variables are analysed at the household level, they refer to the characteristics of the 'household reference person', the person in whose name the property is owned or rented.

* p < .05

** p < .01

*** p < .001

Variable	Variable categories	Household Level ϕ			Individual Level		
		Non-respondents	Refusals	Non-contacts	Non-respondents	Refusals	Non-contacts
		β SE(β)	β SE(β)	β SE(β)	β SE(β)	β SE(β)	β SE(β)
	Qualifications						
PERCEDHQ	% higher quals	-.015 (.010)	-.018 (.013)	-.011 (.015)	-.013 (.006)*	-.014 (.008)	-.012 (.009)
	Age						
PERC0P	% 0-15	.029 (.020)	.038 (.026)	.008 (.029)	.020 (.012)	.030 (.016)	.004 (.017)
PERC15P	% 15-24	.061 (.030)*	.029 (.039)	.082 (.043)	.019 (.018)	-.009 (.024)	.051 (.026)*
PERC25P	% 25-34	.019 (.017)	-.006 (.023)	.036 (.025)	.008 (.011)	-.016 (.014)	.032 (.015)*
PERC35P	% 35-44	-.063 (.032)*	.009 (.041)	-.150 (.048)**	-.039 (.019)*	-.006 (.025)	-.077 (.028)**
PERC45P	% 45-54	-.084 (.039)*	.012 (.049)	-.191 (.059)**	-.051 (.023)*	.015 (.030)	-.135 (.034)***
PERC55P	% 55-64	-.008 (.041)	.051 (.053)	-.062 (.061)	-.007 (.025)	.065 (.032)*	-.086 (.037)
PERC65PL	% 65+	-.001 (.015)	-.014 (.020)	.021 (.021)	-.008 (.009)	.004 (.012)	.014 (.013)
	Migration						
PERCMIG	% diff. address year before census	.009 (.019)	-.031 (.027)	.043 (.026)	-.007 (.012)	-.037 (.016)*	.032 (.016)*
	Standard Industrial Classification						
PERAGRI	% agriculture	-.031 (.036)	.006 (.039)	-.102 (.082)	-.010 (.020)	.030 (.022)	-.096 (.046)*
PERENWAT	% energy & water supplies	-.139 (.074)	.034 (.088)	-.336 (.123)**	-.053 (.042)	-.018 (.054)	-.105 (.063)
PERMINE	% mining	-.044 (.027)	-.006 (.032)	-.091 (.046)*	-.023 (.015)	-.012 (.019)	-.032 (.022)
PERMANM	% metal, engineering & vehicle	.003 (.017)	.003 (.022)	.002 (.025)	.005 (.010)	.007 (.013)	.001 (.015)
PERANTOT	% other manufacturing	-.010 (.014)	-.006 (.018)	-.014 (.021)	-.006 (.008)	.001 (.010)	-.011 (.012)
PERCONST	% construction	.018 (.032)	.019 (.042)	.021 (.048)	-.009 (.020)	.011 (.026)	-.022 (.029)
PERDISC	% distribution, hotel, catering	.034 (.021)	.042 (.027)	.021 (.031)	.024 (.012)*	.044 (.016)**	.005 (.018)
PERTRANS	% transport & communications	.038 (.027)	.039 (.035)	.027 (.040)	.029 (.017)	.026 (.022)	.030 (.024)
PERFINAN	% finance	-.011 (.012)	-.011 (.016)	-.012 (.018)	-.004 (.007)	-.009 (.009)	-.003 (.010)
PERO SER	% other services	.012 (.012)	-.011 (.016)	.038 (.018)*	.001 (.007)	-.019 (.010)*	.024 (.010)*

ϕ Note that when individual level variables are analysed at the household level, they refer to the characteristics of the 'household reference person', the person in whose name the property is owned or rented.

* p < .05

** p < .01

*** p < .001

APPENDIX C: Background Tables for the Calculation of Predicted Probabilities of Nonresponse by Respondent Co-operation and Contactability, Characteristics of the Respondent, and Location ϕ

	GOOD co-operation and contactability EASY to reach demographic group (Other aspects held neutral)	GOOD co-operation and contactability HARD to reach demographic group (Other aspects held neutral)	POOR co-operation and contactability EASY to reach demographic group (Other aspects held neutral)	POOR co-operation and contactability HARD to reach demographic group (Other aspects held neutral)
Household Level Nonresponse				
Co-operation and contactability	1 call until first contact at W2 0.42 No internal HH refusals in W1 0 Co-operation very good at W1 0	1 call until first contact at W2 0.42 No internal HH refusals in W1 0 Co-operation very good at W1 0	6 calls until first contact at W2 2.52 Internal HH refusal in W1 1.61 Co-operation poor/very poor at W1 2.07	6 calls until first contact at W2 2.52 Internal HH refusal in W1 1.61 Co-operation poor/very poor W1 2.07
Demographic and economic profile	3+ Rooms 0 Head of HH aged 26-59 0 Head of HH in non-manual occ. 0 Head of HH has £1000+ dividends -0.21	1-2 Rooms 0.69 Head of HH aged 15-25 0.99 Head of HH in manual occ. 0.87 Head of HH has £0 dividends 0.71	3+ Rooms 0 Head of HH aged 26-59 0 Head of HH in non-manual occ. 0 Head of HH has £1000+ dividends -0.21	1-2 Rooms 0.69 Head of HH aged 15-25 0.99 Head of HH in manual occ. 0.87 Head of HH has £0 dividends 0.71
Other aspects	Same interviewer as W1 0.01 6 calls at Wave 2 -1.50 Geo. pool in Inner/Outer London (Geo. pool in Scotland) 1.55(0)	Same interviewer as W1 0.01 6 calls at Wave 2 -1.50 Geo. pool in Inner/Outer London (Geo. pool in Scotland) 1.55(0)	Same interviewer as W1 0.01 6 calls at Wave 2 -1.50 Geo. pool in Inner/Outer London (Geo. pool in Scotland) 1.55(0)	Same interviewer as W1 0.01 6 calls at Wave 2 -1.50 Geo. pool in Inner/Outer London (Geo. pool in Scotland) 1.55(0)
	Constant -3.65	Constant -3.65	Constant -3.65	Constant -3.65
Household Level Refusals				
Co-operation and contactability	1 call until first contact at W2 0.18 No internal HH refusals in W1 0 Co-operation very good at W1 0	1 call until first contact at W2 0.18 No internal HH refusals in W1 0 Co-operation very good at W1 0	6 calls until first contact at W2 1.08 Internal HH refusals in W1 2.02 Co-operation poor/very poor at W1 1.55	6 calls until first contact at W2 1.08 Internal HH refusals in W1 2.02 Co-operation poor/very poor at W1 1.55
Demographic and economic profile	No Pensioners in HH 0 Head of HH has £1000+ dividends -0.39	1+ Pensioners in HH 1.22 Head of HH has £0 dividends 0.25	No Pensioners in HH 0 Head of HH has £1000+ dividends -0.39	1+ Pensioners in HH 1.22 Head of HH has £0 dividends 0.25
Other aspects	Same interviewer as W1 -0.18 Area with 20% distribution, hotel, or catering industries 2.40 Geo. pool in Inner/Outer London (Geo. pool in Scotland) 0.93(0)	Same interviewer as W1 -0.18 Area with 20% distribution, hotel, or catering industries 2.40 Geo. pool in Inner/Outer London (Geo. pool in Scotland) 0.93(0)	Same interviewer as W1 -0.18 Area with 20% distribution, hotel, or catering industries 2.40 Geo. pool in Inner/Outer London (Geo. pool in Scotland) 0.93(0)	Same interviewer as W1 -0.18 Area with 20% distribution, hotel, or catering industries 2.40 Geo. pool in Inner/Outer London (Geo. pool in Scotland) 0.93(0)
	Constant -6.49	Constant -6.49	Constant -6.49	Constant -6.49
Household Level Non-contacts				
Co-operation and contactability	1 call at W1 0.23 1 call until first contact at W2 0.76 Co-operation very good at W1 0 No problems with English 0	1 call at W1 0.23 1 call until first contact at W2 0.76 Co-operation very good at W1 0 No problems with English 0	6 calls at W1 1.38 6 calls until first contact at W2 4.56 Co-operation poor/very poor at W1 0.50 Problems with English 2.13	6 calls at W1 1.38 6 calls until first contact at W2 4.56 Co-operation poor/very poor at W1 0.50 Problems with English 2.13
Demographic and economic profile	Has phone 0 Head of HH aged 60-93 -1.50	No phone 0.86 Head of HH aged 15-25 1.81	Has phone 0 Head of HH aged 60-93 -1.50	No phone 0.86 Head of HH aged 15-25 1.81
Other aspects	Same interviewer as W1 -0.01 6 calls at Wave 2 -3.66 Area with 30% flats -2.40 Area with 38% no cars 2.28 Area with 22% couples, dep children -3.52 Geo. pool in Inner/Outer London (Geo. pool in Scotland) 1.98(0)	Same interviewer as W1 -0.01 6 calls at Wave 2 -3.66 Area with 30% flats -2.40 Area with 38% no cars 2.28 Area with 22% couples, dep children -3.52 Geo. pool in Inner/Outer London (Geo. pool in Scotland) 1.98(0)	Same interviewer as W1 -0.01 6 calls at Wave 2 -3.66 Area with 30% flats -2.40 Area with 38% no cars 2.28 Area with 22% couples, dep children -3.52 Geo. pool in Inner/Outer London (Geo. pool in Scotland) 1.98(0)	Same interviewer as W1 -0.01 6 calls at Wave 2 -3.66 Area with 30% flats -2.40 Area with 38% no cars 2.28 Area with 22% couples, dep children -3.52 Geo. pool in Inner/Outer London (Geo. pool in Scotland) 1.98(0)
	Constant 1.27	Constant 1.27	Constant 1.27	Constant 1.27

ϕ Note that when individual level variables are analysed at the household level, they refer to the characteristics of the 'household reference person', the person in whose name the property is owned or rented.

Appendix C: (continued)

Individual Level Nonresponse								
Co-operation and contactability	1 call until first contact at W2	0.21	1 call until first contact at W2	0.21	6 calls until first contact at W2	1.26	6 calls until first contact at W2	1.26
	No internal HH refusals in W1	0	No internal HH refusals in W1	0	Internal HH refusals in W1	1.50	Internal HH refusals in W1	1.50
	Co-operation very good at W1	0	Co-operation very good at W1	0	Co-operation poor/very poor at W1	2.34	Co-operation poor/very poor at W1	2.34
	No problems with English	0	No problems with English	0	Problems with English	0.73	Problems with English	0.73
Demographic and economic profile	Terraced house	-0.54	Flats	0	Terraced house	-0.54	Flats	0
	2+ cars	0	No cars	0.73	2+ cars	0	No cars	0.73
	Has home computer	0	No home computer	0.51	Has home computer	0	No home computer	0.51
	Net monthly housing costs £251+	0	Net monthly housing costs £1-125	0.64	Net monthly housing costs £251+	0	Net monthly housing costs £1-125	0.64
	0-3 person of working age	0	4+ person of working age	0.53	0-3 person of working age	0	4+ person of working age	0.53
	No pensioners in HH	0	1+ pensioners in HH	0.76	No pensioners in HH	0	1+ pensioners in HH	0.76
	R is female	0	R is male	0.41	R is female	0	R is male	0.41
	R aged 60-93	-0.11	R aged 15-25	0.94	R aged 60-93	-0.11	R aged 15-25	0.94
R is not in the labour force	-0.54	R is unemployed	0.18	R is not in the labour force	-0.54	R is unemployed	0.18	
Other aspects	Same interviewer as W1	-0.01	Same interviewer as W1	-0.01	Same interviewer as W1	-0.01	Same interviewer as W1	-0.01
	Area with 8% non-white residents	-0.24	Area with 8% non-white residents	-0.24	Area with 8% non-white residents	-0.24	Area with 8% non-white residents	-0.24
	Area with 20% distribution, hotel, or catering industries	1.20	Area with 20% distribution, hotel, or catering industries	1.20	Area with 20% distribution, hotel, or catering industries	1.20	Area with 20% distribution, hotel, or catering industries	1.20
	Geo. pool in Inner/Outer London (Geo. pool in Scotland)	1.96(0)	Geo. pool in Inner/Outer London (Geo. pool in Scotland)	1.96(0)	Geo. pool in Inner/Outer London (Geo. pool in Scotland)	1.96(0)	Geo. pool in Inner/Outer London (Geo. pool in Scotland)	1.96(0)
Constant	-5.47	Constant	-5.47	Constant	-5.47	Constant	-5.47	
Individual Level Refusals								
Co-operation and contactability	1 call until first contact at W2	0.16	1 call until first contact at W2	0.16	6 calls until first contact at W2	0.96	6 calls until first contact at W2	0.96
	No internal HH refusals in W1	0	No internal HH refusals in W1	0	No internal HH refusals in W1	2.12	No internal HH refusals in W1	2.12
	Co-operation very good at W1	0	Co-operation very good at W1	0	Co-operation poor/very poor at W1	2.08	Co-operation poor/very poor at W1	2.08
Demographic and economic profile	Has home computer	0	No home computer	0.72	Has home computer	0	No home computer	0.72
	Net monthly housing costs £251+	0	Net monthly housing costs £1-125	1.50	Net monthly housing costs £251+	0	Net monthly housing costs £1-125	1.50
	No couples in household	-0.67	Couple in household	0	No couples in household	-0.67	Couple in household	0
	No pensioners in HH	0	1+ pensioners in HH	1.02	No pensioners in HH	0	1+ pensioners in HH	1.02
R is not in the labour force	-0.59	R is unemployed	0.19	R is not in the labour force	-0.59	R is unemployed	0.19	
Other aspects	Same interviewer as W1	-0.09	Same interviewer as W1	-0.09	Same interviewer as W1	-0.09	Same interviewer as W1	-0.09
	Interviewer aged 61-72	0	Interviewer aged 61-72	0	Interviewer aged 61-72	0	Interviewer aged 61-72	0
	Area with 8% non-white residents	-0.24	Area with 8% non-white residents	-0.24	Area with 8% non-white residents	-0.24	Area with 8% non-white residents	-0.24
	Area with 20% distribution, hotel, or catering industries	2.00	Area with 20% distribution, hotel, or catering industries	2.00	Area with 20% distribution, hotel, or catering industries	2.00	Area with 20% distribution, hotel, or catering industries	2.00
Geo. pool in Inner/Outer London (Geo. pool in Scotland)	1.06(0)	Geo. pool in Inner/Outer London (Geo. pool in Scotland)	1.06(0)	Geo. pool in Inner/Outer London (Geo. pool in Scotland)	1.06(0)	Geo. pool in Inner/Outer London (Geo. pool in Scotland)	1.06(0)	
Constant	-6.89	Constant	-6.89	Constant	-6.89	Constant	-6.89	

φ Note that when individual level variables are analysed at the household level, they refer to the characteristics of the 'household reference person', the person in whose name the property is owned or rented.

Appendix C: (continued)

Individual Level Non-contacts									
Co-operation and contactability	25 of W1 calls were non-contacts	0.21	25 of W1 calls were non-contacts	0.21	1.0 of W1 calls were non-contacts	0.83	1.0 of W1 calls were non-contacts	0.83	
	1 call until first contact at W2	0.27	1 call until first contact at W2	0.27	6 calls until first contact at W2	1.62	6 calls until first contact at W2	1.62	
	Co-operation very good at W1	0	Co-operation very good at W1	0	Co-operation poor/very poor at W1	2.33	Co-operation poor/very poor at W1	2.33	
Demographic and economic profile	Accommodation is owned	0	Accommodation is rented	0.54	Accommodation is owned	0	Accommodation is rented	0.54	
	R is female	0	R is male	0.69	R is female	0	R is male	0.69	
	R aged 60-93	-0.81	R aged 15-25	1.38	R aged 60-93	-0.81	R aged 15-25	1.38	
	R is white	0	R is non-white	0.78	R is white	0	R is non-white	0.78	
Other aspects	Same interviewer as W1	-0.05	Same interviewer as W1	-0.05	Same interviewer as W1	-0.05	Same interviewer as W1	-0.05	
	Interviewer with 0-2 years exper.	0	Interviewer with 0-2 years exper.	0	Interviewer with 0-2 years exper.	0	Interviewer with 0-2 years exper.	0	
	6 calls at Wave 2	-0.66	6 calls at Wave 2	-0.66	6 calls at Wave 2	-0.66	6 calls at Wave 2	-0.66	
	Geo. pool in Inner/Outer London	1.23(0)	Geo. pool in Inner/Outer London	1.23(0)	Geo. pool in Inner/Outer London	1.23(0)	Geo. pool in Inner/Outer London	1.23(0)	
	(Geo. pool in Scotland)		(Geo. pool in Scotland)		(Geo. pool in Scotland)		(Geo. pool in Scotland)		
Constant	-5.02	Constant	-5.02	Constant	-5.02	Constant	-5.02		

φ Note that when individual level variables are analysed at the household level, they refer to the characteristics of the 'household reference person', the person in whose name the property is owned or rented.

⊕ The profiles were constructed as follows:

For **categorical variables** (represented by a series of dummy variables), the categories with coefficients which indicated the biggest reduction in nonresponse were chosen for the 'good/easy' profiles, coefficients which indicated the biggest increase in nonresponse were chosen for the 'poor/hard' profiles, and categories which made no impact were chosen for the neutral 'Other aspects' variables.

For **continuous variables**, a relatively low value of x and a relatively high value of x were chosen depending on the sign of the coefficient and whether or not it was for a 'good/easy' profile or a 'poor/hard' profile. The most extreme values of x were avoided in instances where these rarely occurred in the data. The mean of x , rounded to the nearest integer, was chosen for the neutral 'Other aspects' variables.

There were some **exceptions** to these general rules. When two coefficients in a categorical variable had similar values, the most frequently occurring category in the data was chosen, even if it wasn't the most extreme. For example in terms of AJBGOLD2 (Goldthorpe Social Class), although the coefficient for manual workers was 0.87 and the coefficient for small proprietors was 0.94, the manual workers category was chosen. Other exceptions occurred in order to keep consistency across the profiles. For example with respect to AIV42 (the co-operation of the respondent), the coefficient making the most impact on 5 of the 6 models was the 'poor/very poor' co-operation category. Thus it was used throughout. Other times the most sensible category was chosen, even if it wasn't the most extreme. For example in terms of AFIYRD12 (Income last year from dividends/interest), the 'nothing' category with a coefficient of 0.71 for household nonresponse was chosen instead of the '100-1000 pounds' category with a coefficient of 0.73, as the former was more consistent with the profile of someone not economically well off.

As summarised in Section 7.4.2, the variable BIVLNC (Number of calls at last known address at Wave 2) is not well behaved. As it can have a strong impact on the predicted probabilities depending on the value chosen, it was placed in the 'Other aspects' section and is held constant with a value of 6 across all profiles.

**APPENDIX D: Wave 2 Binary and Multinomial Cross-Classified Multilevel
Regression Models using RIGLS, PQL and 2nd Order Estimation**

Table D1: Variance Components Models

Household Level Nonresponse (n = 1,365) †				
	Model 1 Binary Refusals	Model 2 Binary Non-contacts	Estimate	Model 3 Multinomial Correlation between random terms
Fixed Effects				
Refusals	-2.70 (0.13)		-2.71 (0.13)	
Non-contacts		-3.05 (0.16)	-3.00 (0.15)	
Random Effects				
Geographic pools				
Refusals	0.0		0.0	NA
Non-contacts		0.0	0.0	
PSUs				
Refusals	0.0		0.0	NA
Non-contacts		0.31 (0.23)	0.21 (0.22)	
Interviewers				
Refusals	0.48 (0.20) *		0.50 (0.20) *	0.63
Non-contacts		0.52 (0.26) *	0.47 (0.26)	
Households	0.80 (0.03) * <u>u</u>	0.66 (0.03) * <u>u</u>	0.76 (0.02) * <u>u</u>	
Individual Level Nonresponse (n = 2,421)				
	Model 4 Binary Refusals	Model 5 Binary Non-contacts	Estimate	Model 6 Multinomial Correlation between random terms
Fixed Effects				
Refusals	-2.75 (0.13)		-2.74 (0.13)	
Non-contacts		-2.98 (0.14)	-2.97 (0.14)	
Random Effects				
Geographic pools				
Refusals	0.0		0.0	NA
Non-contacts		0.0	0.0	
PSUs				
Refusals	0.13 (0.13)		0.12 (0.12)	-0.36
Non-contacts		0.35 (0.19)	0.32 (0.18)	
Interviewers				
Refusals	0.52 (0.19) *		0.53 (0.19) *	0.60
Non-contacts		0.27 (0.18)	0.30 (0.18)	
Individuals	0.91 (0.03) * <u>u</u>	0.91 (0.03) * <u>u</u>	0.90 (0.02) * <u>u</u>	

† 2nd order estimation would not converge for this household level multinomial model. RIGLS and PQL with unconstrained level 1 variance were used for the binary household level models to facilitate comparison with the multinomial model.

* Coefficient significant at p < .05 using MLwiN's separate (rather than simultaneous) Wald test (p < .01 and p < .001 not shown).

*u Model is significantly under-dispersed.

Table D2: Covariates Models

Household Level Nonresponse (n = 1,365) ϕ					
	Model 7 Binary	Model 8 Binary	Model 9 Multinomial	Model 9a Multinomial	
	Refusals	Non-contacts	Estimate	Correlation between random terms	Single Level
Fixed Effects					
Refusals (constant)	-3.84 (0.32)		-4.00 (0.31)		-3.67 (0.31)
Non-contacts (constant)		-5.66 (0.46)	-5.37 (0.43)		-4.50 (0.35)
W1 num of calls - Ref	-0.03 (0.06)		-0.02 (0.05)		-0.01 (0.06)
W1 num of calls - Ncon		0.19 (0.06) *	0.17 (0.06) *		0.15 (0.05) *
Tot calls till 1 st contact W2 - R	0.25 (0.08) *		0.27 (0.07) *		0.22 (0.07) *
Tot calls till 1 st contact W2 - N		0.35 (0.07) *	0.34 (0.06) *		0.26 (0.06) *
W1 not comp co-op HH-R	1.98 (0.36) *		2.04 (0.34) *		1.90 (0.35) *
W1 not comp co-op HH-N		0.35 (0.70)	0.50 (0.63)		0.28 (0.64)
W1 co-operation rating (base category = very good)					
missing - Ref	0.55 (1.35)		0.59 (1.24)		0.53 (1.39)
good/fair - Ref	1.15 (0.26) *		1.19 (0.25) *		1.11 (0.26) *
poor/very poor - Ref	1.93 (0.72) *		2.35 (0.64) *		1.97 (0.70) *
missing - Ncon		3.83 (1.22) *	3.59 (1.27) *		3.63 (1.23) *
good/fair - Ncon		1.30 (0.32) *	1.22 (0.32) *		0.97 (0.31) *
poor/very poor - Ncon		1.63 (1.21)	2.28 (0.93) *		0.37 (1.26)
Problems with English (base category = no)					
missing - Ref	-1.34 (1.32)		-1.35 (1.22)		-1.30 (1.36)
yes - Ref	0.66 (0.77)		0.80 (0.64)		1.01 (0.73)
missing - Ncon		-3.57 (1.52) *	-3.28 (1.55) *		-3.56 (1.48) *
yes - Ncon		1.97 (0.61) *	2.06 (0.59) *		2.25 (0.61) *
Any pensioners in HH? - Ref	1.28 (0.58) *		1.22 (0.54) *		1.38 (0.58) *
Any pensioners in HH? - Ncon		0.89 (0.93)	0.78 (0.95)		0.93 (0.91)
Age of Head of Household (base category = 26-59)					
15-25 - Ref	-0.23 (0.51)		-0.18 (0.45)		-0.16 (0.49)
60-93 - Ref	-0.19 (0.58)		-0.12 (0.54)		-0.33 (0.58)
15-25 - Ncon		1.94 (0.31) *	1.86 (0.31) *		1.70 (0.30) *
60-93 - Ncon		-2.04 (1.06)	-1.89 (1.06)		-1.93 (1.00)
Random Effects					
Geographic pools					
Refusals	0.0		0.0	NA	NA
Non-contacts		0.0	0.0		
PSUs					
Refusals	0.0		0.06 (0.19)	0.48	NA
Non-contacts		0.80 (0.49)	0.48 (0.40)		
Interviewers					
Refusals	0.49 (0.25)		0.70 (0.29) *	0.49	NA
Non-contacts		1.23 (0.58) *	1.03 (0.50) *		
Households	0.89 (0.04) * <u>u</u>	0.65 (0.03) * <u>u</u>	0.76 (0.02) * <u>u</u>		1

ϕ Note that when individual level variables are analysed at the household level, they refer to the characteristics of the 'household reference person', the person in whose name the property is owned or rented.

* Coefficient significant at $p < .05$ using MLwiN's separate (rather than simultaneous) Wald test ($p < .01$ and $p < .001$ not shown).

*u Model is significantly under-dispersed.

Table D2: (continued)

Individual Level Nonresponse (n = 2,421)				
	Model 10 Binary	Model 11 Binary	Model 12 Multinomial	Model 12a Multinomial
	Refusals	Non-contacts	Estimate	Correlation between random terms Single Level
Fixed Effects				
Refusals (constant)	-4.72 (0.45)		-4.68 (0.44)	-4.42 (0.36)
Non-contacts (constant)		-4.98 (0.43)	-4.94 (0.40)	-4.50 (0.35)
Interviewer yrs with company (base category = 0-2 years)				
3-5 years - Ref	-0.37 (0.37)		-0.39 (0.36)	-0.29 (0.21)
6-7 years - Ref	0.26 (0.42)		0.26 (0.41)	0.16 (0.25)
8-10 years - Ref	-0.51 (0.58)		-0.54 (0.57)	-0.48 (0.35)
11-13 years - Ref	-0.82 (0.75)		-0.74 (0.73)	-0.75 (0.50)
14-20 years - Ref	-0.35 (0.57)		-0.31 (0.55)	-0.23 (0.35)
3-5 years - Ncon		0.35 (0.34)	0.44 (0.31)	0.20 (0.26)
6-7 years - Ncon		0.51 (0.39)	0.72 (0.34) *	0.31 (0.30)
8-10 years - Ncon		0.68 (0.48)	0.92 (0.44) *	0.29 (0.35)
11-13 years - Ncon		1.10 (0.49) *	1.10 (0.43) *	1.04 (0.35) *
14-20 years - Ncon		0.15 (0.53)	0.15 (0.47)	0.32 (0.40)
Tot calls till 1 st contact W2 - R	0.23 (0.06) *		0.22 (0.06) *	0.21 (0.05) *
Tot calls till 1 st contact W2 - N		0.25 (0.05) *	0.25 (0.05) *	0.26 (0.05) *
W1 not comp co-op HH - R	2.15 (0.25) *		2.14 (0.25) *	1.94 (0.23) *
W1 not comp co-op HH - N		0.63 (0.39)	0.64 (0.36) *	0.61 (0.36)
W1 co-operation rating (base category = very good)				
missing - Ref	-0.56 (0.80)		-0.54 (0.78)	-0.69 (0.74)
good/fair - Ref	0.86 (0.21) *		0.84 (0.21) *	0.80 (0.19) *
poor/very poor - Ref	2.47 (0.70) *		2.74 (0.59) *	2.61 (0.64) *
missing - Ncon		0.79 (0.54)	0.77 (0.53)	0.68 (0.49)
good/fair - Ncon		0.93 (0.24) *	0.91 (0.23) *	0.80 (0.22) *
poor/very poor - Ncon		3.03 (0.84) *	3.04 (0.69) *	2.72 (0.80) *
Home computer in accommodation - R	0.77 (0.28) *		0.76 (0.27) *	0.78 (0.25) *
Home computer in accommodation - N		0.54 (0.25) *	0.56 (0.24) *	0.50 (0.23) *
Net monthly housing costs (base category = 251-1650)				
none - Ref	0.57 (0.34)		0.56 (0.33)	0.61 (0.30) *
1-125 pounds - Ref	1.41 (0.32) *		1.39 (0.32) *	1.30 (0.29) *
126-250 pounds - Ref	0.72 (0.32) *		0.70 (0.32) *	0.69 (0.29) *
none - Ncon		-0.11 (0.31)	-0.12 (0.30)	-0.20 (0.28)
1-125 pounds - Ncon		-0.10 (0.32)	-0.09 (0.30)	-0.19 (0.28)
126-250 pounds - Ncon		0.12 (0.27)	0.07 (0.25)	0.01 (0.23)
A couple in the household - Ref	-0.62 (0.23) *		-0.61 (0.23) *	-0.61 (0.21) *
A couple in the household - Ncon		0.18 (0.22)	0.21 (0.21)	0.26 (0.20)
Any pensioners in HH? - Ref	0.92 (0.34) *		0.92 (0.34) *	0.89 (0.31) *
Any pensioners in HH? - Ncon		0.14 (0.49)	0.17 (0.47)	0.07 (0.46)
Respondent gender, men - Ref	0.13 (0.18)		0.14 (0.18)	0.18 (0.17)
Respondent gender, men - Ncon		0.67 (0.20) *	0.63 (0.19) *	0.62 (0.19) *
Age of Head of Household (base category = 26-59)				
15-25 - Ref	0.11 (0.27)		0.14 (0.26)	0.14 (0.24)
60-93 - Ref	-0.24 (0.35)		-0.25 (0.35)	-0.27 (0.32)
15-25 - Ncon		1.38 (0.21) *	1.35 (0.20) *	1.29 (0.19) *
60-93 - Ncon		-0.84 (0.55)	-0.87 (0.53)	-0.72 (0.51)

Table D2: (continued)

	Model 10 Binary	Model 11 Binary	Model 12 ‡ Multinomial	Correlation between random terms	Model 12a Multinomial
	Refusals	Non-contacts	Estimate		Single Level
Random Effects					
Geographic pools					
Refusals	0.0		0.0	NA	NA
Non-contacts		0.08 (0.22)	0.20 (0.19)		
PSUs					
Refusals	0.06 (0.13)		0.05 (0.12)	-0.24	NA
Non-contacts		0.34 (0.24)	0.27 (0.21)		
Interviewers					
Refusals	0.69 (0.24) *		0.68 (0.23) *	1.0 ‡	NA
Non-contacts		0.11 (0.18)	0.00 (0.13)		
Individuals	0.98 (0.03)	0.98 (0.03)	1.00 (0.02)		1

* Coefficient significant at $p < .05$ using MLwiN's separate (rather than simultaneous) Wald test ($p < .01$ and $p < .001$ not shown).

‡ The interviewers contribution towards non-contacts is essentially zero and as such has distorted this correlation so that it is greater than 1.0.

APPENDIX E: Wave 2 Binary and Multinomial Cross-Classified Multilevel Regression Models using RIGLS and MQL Estimation

Table E1: Variance Components Models

Household Level Nonresponse (n = 1,365)				
	Model 1 Binary Refusals	Model 2 Binary Non-contacts	Estimate	Model 3 Multinomial Correlation between random terms
Fixed Effects				
Refusals	-2.62 (0.13)		-2.62 (0.13)	
Non-contacts		-2.89 (0.16)	-2.90 (0.15)	
Random Effects				
Geographic pools				
Refusals	0.0		0.0	NA
Non-contacts		0.04 (0.26)	0.0	
PSUs				
Refusals	0.0		0.0	NA
Non-contacts		0.15 (0.27)	0.13 (0.23)	
Interviewers				
Refusals	0.38 (0.20)		0.32 (0.19)	0.78
Non-contacts		0.31 (0.31)	0.30 (0.26)	
Households	0.97 (0.04)	0.96 (0.04)	0.97 (0.03)	
Individual Level Nonresponse (n = 2,421)				
	Model 4 Binary Refusals	Model 5 Binary Non-contacts	Estimate	Model 6 Multinomial Correlation between random terms
Fixed Effects				
Refusals	-2.47 (0.12)		-2.47 (0.12)	
Non-contacts		-2.66 (0.14)	-2.67 (0.13)	
Random Effects				
Geographic pools				
Refusals	0.0		0.0	NA
Non-contacts		0.12 (0.17)	0.09 (0.17)	
PSUs				
Refusals	0.07 (0.10)		0.07 (0.10)	-0.76
Non-contacts		0.24 (0.17)	0.23 (0.16)	
Interviewers				
Refusals	0.45 (0.16) *		0.43 (0.15) *	0.55
Non-contacts		0.22 (0.16)	0.21 (0.16)	
Individuals	0.95 (0.03)	0.93 (0.03) * _u	0.95 (0.02) * _u	

* Coefficient significant at $p < .05$ using MLwiN's separate (rather than simultaneous) Wald test ($p < .01$ and $p < .001$ not shown).

*_u Model is significantly under-dispersed.

Table E2: Covariates Models

Household Level Nonresponse (n = 1,365) ϕ					
	Model 7 Binary	Model 8 Binary	Model 9 Multinomial	Model 9a Multinomial	
	Refusals	Non-contacts	Estimate	Correlation between random terms	Single Level
Fixed Effects					
Refusals (constant)	-3.61 (0.31)		-3.58 (0.29)		-3.67 (0.31)
Non-contacts (constant)		-4.48 (0.33)	-4.42 (0.34)		-4.50 (0.35)
W1 num of calls - Ref	-0.02 (0.05)		-0.02 (0.05)		-0.01 (0.06)
W1 num of calls - Ncon		0.15 (0.05) *	0.14 (0.05) *		0.15 (0.05) *
Tot calls till 1 st contact W2 - R	0.23 (0.07) *		0.22 (0.07) *		0.22 (0.07) *
Tot calls till 1 st contact W2 - N		0.27 (0.06) *	0.26 (0.06) *		0.26 (0.06) *
W1 not comp co-op HH-R	1.86 (0.36) *		1.86 (0.33) *		1.90 (0.35) *
W1 not comp co-op HH-N		0.31 (0.55)	0.24 (0.60)		0.28 (0.64)
W1 co-operation rating (base category = very good)					
missing - Ref	0.50 (1.36)		0.55 (1.24)		0.53 (1.39)
good/fair - Ref	1.10 (0.26) *		1.14 (0.24) *		1.11 (0.26) *
poor/very poor - Ref	1.84 (0.73) *		1.91 (0.66) *		1.97 (0.70) *
missing - Ncon		3.29 (1.14) *	3.31 (1.24) *		3.63 (1.23) *
good/fair - Ncon		1.05 (0.26) *	1.01 (0.29) *		0.97 (0.31) *
poor/very poor - Ncon		1.64 (1.05)	1.01 (1.04)		0.37 (1.26)
Problems with English (base category = no)					
missing - Ref	-1.27 (1.33)		-1.31 (1.24)		-1.30 (1.36)
yes - Ref	0.70 (0.78)		0.42 (0.74)		1.01 (0.73)
missing - Ncon		-3.08 (1.33) *	-3.12 (1.44) *		-3.56 (1.48) *
yes - Ncon		1.70 (0.60) *	1.65 (0.62) *		2.25 (0.61) *
Any pensioners in HH? - R	1.27 (0.57) *		1.20 (0.53) *		1.38 (0.58) *
Any pensioners in HH? - N		0.76 (0.77)	0.68 (0.88)		0.93 (0.91)
Age of Head of Household (base category = 26-59)					
15-25 - Ref	-0.21 (0.50)		-0.24 (0.47)		-0.16 (0.49)
60-93 - Ref	-0.22 (0.57)		-0.18 (0.53)		-0.33 (0.58)
15-25 - Ncon		1.64 (0.26) *	1.66 (0.28) *		1.70 (0.30) *
60-93 - Ncon		-1.84 (0.84) *	-1.71 (0.95)		-1.93 (1.00)
Random Effects					
Geographic pools					
Refusals	0.0		0.0	NA	NA
Non-contacts		0.0	0.0		
PSUs					
Refusals	0.0		0.00	NA	NA
Non-contacts		0.55 (0.31)	0.24 (0.27)		
Interviewers					
Refusals	0.35 (0.21)		0.41 (0.20) *	0.49	NA
Non-contacts		0.67 (0.32) *	0.47 (0.31)		
Households	0.97 (0.04)	0.69 (0.03) * _u	0.86 (0.02) * _u		1

ϕ Note that when individual level variables are analysed at the household level, they refer to the characteristics of the 'household reference person', the person in whose name the property is owned or rented.

* Coefficient significant at $p < .05$ using MLwiN's separate (rather than simultaneous) Wald test ($p < .01$ and $p < .001$ not shown).

*_u Model is significantly under-dispersed.

Table E2: (continued)

Individual Level Nonresponse (n = 2,421)				
	Model 10 Binary	Model 11 Binary	Model 12↓ Multinomial	Model 12a Multinomial
	Refusals	Non-contacts	Estimate	Correlation between random terms Single Level
Fixed Effects				
Refusals (constant)	-4.31 (0.40)		-4.31 (0.40)	-4.42 (0.36)
Non-contacts (constant)		-4.67 (0.39)	-4.66 (0.38)	-4.50 (0.35)
Interviewer yrs with company (base category = 0-2 years)				
3-5 years - Ref	-0.31 (0.33)		-0.31 (0.32)	-0.29 (0.21)
6-7 years - Ref	0.23 (0.38)		0.26 (0.37)	0.16 (0.25)
8-10 years - Ref	-0.52 (0.52)		-0.48 (0.50)	-0.48 (0.35)
11-13 years - Ref	-0.82 (0.67)		-0.80 (0.66)	-0.75 (0.50)
14-20 years - Ref	-0.32 (0.53)		-0.35 (0.52)	-0.23 (0.35)
3-5 years - Ncon		0.39 (0.31)	0.37 (0.30)	0.20 (0.26)
6-7 years - Ncon		0.60 (0.34)	0.57 (0.33)	0.31 (0.30)
8-10 years - Ncon		0.84 (0.41) *	0.80 (0.40) *	0.29 (0.35)
11-13 years - Ncon		1.08 (0.44) *	1.11 (0.43) *	1.04 (0.35) *
14-20 years - Ncon		0.17 (0.48)	0.19 (0.48)	0.32 (0.40)
Tot calls till 1 st contact W2 - R	0.21 (0.06) *		0.19 (0.05) *	0.21 (0.05) *
Tot calls till 1 st contact W2 - N		0.23 (0.05) *	0.23 (0.05) *	0.26 (0.05) *
W1 not completely co- operating HH - R	1.94 (0.24) *		1.93 (0.23) *	1.94 (0.23) *
W1 not completely co- operating HH - N		0.59 (0.36)	0.57 (0.34)	0.61 (0.36)
W1 co-operation rating (base category = very good)				
missing - Ref	-0.56 (0.69)		-0.58 (0.69)	-0.69 (0.74)
good/fair - Ref	0.78 (0.19) *		0.77 (0.19) *	0.80 (0.19) *
poor/very poor - Ref	2.12 (0.64) *		2.17 (0.58) *	2.61 (0.64) *
missing - Ncon		0.72 (0.49)	0.72 (0.49)	0.68 (0.49)
good/fair - Ncon		0.87 (0.22) *	0.87 (0.21) *	0.80 (0.22) *
poor/very poor - Ncon		2.76 (0.82) *	2.49 (0.72) *	2.72 (0.80) *
Home computer in accom - R	0.74 (0.24) *		0.75 (0.24) *	0.78 (0.25) *
Home computer in accom - N		0.50 (0.22) *	0.51 (0.22) *	0.50 (0.23) *
Net monthly housing costs (base category = 251-1650)				
none - Ref	0.54 (0.31)		0.54 (0.30)	0.61 (0.30) *
1-125 pounds - Ref	1.31 (0.29) *		1.28 (0.29) *	1.30 (0.29) *
126-250 pounds - Ref	0.68 (0.29) *		0.67 (0.29) *	0.69 (0.29) *
none - Ncon		-0.10 (0.28)	-0.13 (0.38)	-0.20 (0.28)
1-125 pounds - Ncon		-0.08 (0.28)	-0.13 (0.28)	-0.19 (0.28)
126-250 pounds - Ncon		0.11 (0.24)	0.08 (0.24)	0.01 (0.23)
A couple in the household - Ref	-0.60 (0.21) *		-0.59 (0.21) *	-0.61 (0.21) *
A couple in the household - Ncon		0.18 (0.21)	0.19 (0.20)	0.26 (0.20)
Any pensioners in HH? - Ref	0.87 (0.32) *		0.87 (0.31) *	0.89 (0.31) *
Any pensioners in HH? - Ncon		0.13 (0.43)	0.13 (0.43)	0.07 (0.46)
Respondent gender, men - Ref	0.12 (0.17)		0.14 (0.17)	0.18 (0.17)
Respondent gender, men - Ncon		0.64 (0.18) *	0.64 (0.18) *	0.62 (0.19) *
Age of Head of Household (base category = 26-59)				
15-25 - Ref	0.10 (0.24)		0.10 (0.24)	0.14 (0.24)
60-93 - Ref	-0.22 (0.32)		-0.23 (0.32)	-0.27 (0.32)
15-25 - Ncon		1.31 (0.19) *	1.32 (0.19) *	1.29 (0.19) *
60-93 - Ncon		-0.82 (0.48)	-0.78 (0.49)	-0.72 (0.51)

Table E2: (continued)

	Model 10 Binary	Model 11 Binary	Model 12 ‡ Multinomial	Correlation between random terms	Model 12a Multinomial
	Refusals	Non-contacts	Estimate		Single Level
Random Effects					
Geographic pools					
Refusals	0.01 (0.18)		0.0	NA	NA
Non-contacts		0.21 (0.17)	0.16 (0.16)		
PSUs					
Refusals	0.02 (0.11)		0.01 (0.09)	-0.22	NA
Non-contacts		0.24 (0.18)	0.24 (0.18)		
Interviewers					
Refusals	0.62 (0.25) *		0.57 (0.19) *	NA	NA
Non-contacts		0.0	0.0		
Individuals	0.97 (0.03)	0.97 (0.03)	0.98 (0.02)		1

* Coefficient significant at $p < .05$ using MLwiN's separate (rather than simultaneous) Wald test ($p < .01$ and $p < .001$ not shown).

‡ Model 12 did not converge. The un-converged version, however, has been included in this table for comparison purposes as the current and previous estimates agree with respect to the first and second significant digits in the vast majority of cases.