University of Southampton

FACULTY OF SOCIAL SCIENCES

School of Economic, Social and Political Science

Department of Social Statistics and Demography

An Investigation of methods for Improving Survey Quality

by

Eliud Muriithi Kibuchi

Thesis for the degree of Doctor of Philosophy

December 2018

University of Southampton

Abstract

Faculty of SOCIAL SCIENCES

School of Economic, Social and Political Science

Social Statistics

Thesis for the degree of Doctor of Philosophy

An Investigation of methods for Improving Survey Data Quality

Eliud Muriithi Kibuchi

Survey data can reduce the risk of making poor public policies and business decisions. It is therefore essential that we continually seek to understand how survey practices affect data quality. The quality of survey data is affected by how well survey questions measure constructs of interest as well as how generalisable such data is to the target population. This thesis consists of three papers, and each addresses the issues of how survey data quality is affected by different methodological choices.

The first paper provides an assessment of the effectiveness of a Bayesian framework to improve predictions of survey nonresponse using response propensity models. Generally, response propensity models exhibit low predictive power for survey nonresponse. This limits their effective application in monitoring and controlling the performance of the survey processes which, in turn, affect survey data quality. This paper explores the utility of a Bayesian approach in improving the predictions of response propensities by using informative priors derived from historical response data. The estimates from the response propensity models fitted to existing data are used as a source for specifying prior distributions in subsequent data collection rounds. The results show that informative priors only lead to a slight improvement in predictions and discriminative ability of response propensity models.

The second paper investigates whether interviewers moderate the effect of monetary incentives on response and cooperation rates in household interview surveys. Incentives play an important role in maintaining response rates and interviewers are the key conduit of information about the existence and level of incentives offered. This paper uses multilevel models to assess whether some interviewers are more successful than others in the deployment of incentives to leverage survey response and cooperation. This paper also investigates whether interviewer variability on incentives is systematically related to interviewer characteristics. The results show significant and substantial variability between interviewers in the effectiveness of monetary incentives on the probability of response and cooperation, but no observed characteristics of interviewers are related to this tendency.

The third paper focuses on whether low response rate online probability surveys provide data of comparable quality than high response rate face-to-face interviews. Declining response rates and increasing survey costs have promoted many surveys to switch from face-to-face interviews to online administration. The available evidence on data quality between face-to-face and online surveys is mixed. This paper examines measurement differences in online and face-to-face surveys while adjusting for selection effects using propensity score matching. In addition, different methods of handling survey weights in propensity score models and outcome analyses are evaluated. The results show that measurement effects contribute the majority of mode differences with sample compositional differences playing a secondary role. However, propensity score matching had only a minimal effect on the magnitude of mode effects for surveys considered.

Table of Contents

List of Tables

List of Figures

Research Thesis: Declaration of Authorship

Print name: Eliud Muriithi Kibuchi

Title of thesis: An Investigation of Methods for Improving Survey Quality

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:

Kibuchi, E., Sturgis, P., Durrant, G. B., & Maslovskaya, O. (2018). Do interviewers moderate the effect of monetary incentives on response rates in household interview survey. J*ournal of Survey Statistics and Methodology*. [Paper 2]

Signature: Date:

Acknowledgements

This PhD would not have been possible without the PhD studentship from the United Kingdom's Economic and Social Research Council (ESRC) (ES/J500161/1).

I am profoundly grateful to my supervisors Professor Gabriele B. Durrant, Professor Patrick Sturgis and Dr. Olga Maslovskaya for their valuable guidance, support, patience and presence throughout the entire process of my PhD. I am extremely grateful for their meticulous and timely suggestions, comments, and ideas despite their busy schedules. None of this work would have been possible without their help and motivation.

I would like to thank my upgrade examiners Professor Jakub Bijak and Professor Paul Smith for their useful comments and suggestions during the upgrade exam.

A big thank you to my family, you have always supported and encouraged my dreams, even if that means being far away from you. Mum and late dad, I cannot thank you enough for the unwavering belief in my abilities. In particular, you taught me that with fear of God, hard work, dedication, enthusiasm and integrity, I can achieve anything I dream of. I love you very much. Thank you, sis Rose, for always being there for encouragement and support. Your countless calls and texts made it possible to finish this PhD.

I am also grateful to the National Survey for Wales, Office of National Statistics, UK Data Archive and Kantar Public-UK for providing access to the various datasets used in my thesis for the analysis. I would also like to thank The Administrative Data Research Centre-England (ADRC-E) based at University of Southampton for allowing me to use their secure Lab for data analysis.

I am also deeply grateful to all the people in the Department of Social Statistics and Demography and National Centre for Research Methods (NCRM) for making me feel welcome over the last 3 years. I would like to say thank you to my friends and fellow PhD students for their emotional support, care and companionship. Special thanks to my late friend Fredrick Ibinda, for you always brought out the best in me.

I am also grateful to Ann Bytheway for proofreading this thesis.

I am very grateful to have this team of people in my life throughout my PhD and I dedicate this thesis to all of you.

Definitions and Abbreviations

American Association for Public Opinion Research (AAPOR)

Address Based Online Surveying (ABOS)

Audio Computer Assisted Self Interviews (ACASI)

Absolute Percentage Differences (APD)

Area under the Curve (AUC)

Bayesian inference Using Gibbs Sampling (BUGS)

British Household Panel Survey (BHPS)

Computer Assisted Interviewing (CAI)

Computer Assisted Personal Interviewing (CAPI)

Computer Assisted Telephone Interviewing (CATI)

Computer Assisted Self Interviews (CASI)

Community Life Survey (CLS)

Cooperation Rate (CR)

Coefficients of Variation (CV)

Deviance Information Criterion (DIC)

Ethnic Minority Boost Sample (EMBS)

Economic Social Survey (ESS)

Great Britain (GB)

'Greedy' Nearest Neighbour and Calliper Matching (G-NNCM)

Government Office Region (GOR)

General Population sample (GP)

Intraclass Correlation Coefficient (ICC)

Innovation Panel (IP)

Integrated Nested Laplace Approximations (INLA)

Inverse Probability Weighting (IPW)

Local Authority (LA)

Latent Gaussian Models (LGM)

Leverage Saliency Theory (LST)

Markov Chain Monte Carlo (MCMC)

Maximum Likelihood (ML)

Multiple Membership Models (MMM)

Multiple Membership Multiple Classification (MMMC)

Mean-Squared Error (MSE)

Middle Layer Super Output Areas (MSOA)

Non-Numeric Value (NaN)

Northern Ireland (NI)

Nearest Neighbour Matching (NNM)

National Survey for Wales (NSW)

Negative Predicted Value (NPV)

Office of National Statistics (ONS)

Penalised-Quasi Likelihood (PQL)

Postcode Address File (PAF)

Positive Predicted Value (PPV)

Primary Sampling Unit (PSU)

Propensity Scores (PS)

Propensity Score Matching (PSM)

Primary Sampling Units (PSU)

Restricted Maximum Likelihood (REML)

Response Propensity (RP)

Response Rate (RR)

Receiver Operating Curves (ROC)

Social Exchange Theory (SET)

Survey of Income and Program Participation (SIPP)

Sequential Bayesian Updating (SBU)

Standardised Mean Differences (SMD)

Total Survey Error (TSE)

United Kingdom (UK)

United Kingdom Household Longitudinal Survey Innovation Panel (UKHLS-IP)

Watanabe-Akaike Information Criterion (WAIC)

Variance Partition Coefficient (VPC)

Chapter 1 Introduction

Survey research is essential for understanding issues affecting societies and providing guidance on policy. Survey methodological research aims to ensure that survey data are accurate, timely, and accessible to the intended users within the budgeted costs. However, the quality of data from surveys is under threat due to increasing nonresponse rates and survey costs. Despite this, surveys remain the bedrock through which key public policies and business decisions are made. Therefore, a clear understanding of survey quality is of paramount importance because it affects both the accuracy of estimates and the conclusions based on the results obtained. The primary aim of this thesis is to investigate ways of understanding and improving survey data quality by studying the factors that influence survey errors. It comprises three papers. The first paper explores the utility of a Bayesian approach in improving the predictions of response propensity in general population surveys. The second paper investigates the role of interviewers in determining whether incentives are effective in improving response and cooperation rates in household surveys. The third paper compares data quality between online probability surveys with low response rates and a high response rate face-to-face interview survey, while adjusting for selection effects using the propensity score matching approach.

In this first chapter, the Total Survey Error (TSE) is introduced as a framework for understanding the statistical properties of survey estimates while accounting for a range of different error sources (Biemer, 2010; Groves, 1989, pp. 1-47). TSE refers to the accumulation of all errors that arise in the design, collection, processing, and analysis of survey data (Biemer, 2010). The ultimate aim of any survey research is to measure accurately the constructs of interest within budgeted costs. However, survey measures may deviate from the true values leading to bias and noise in survey estimates (Biemer & Lyberg, 2003, pp. 26-62). This is due to survey errors which are classified into five error categories by Groves (1989): nonresponse error, measurement error, processing error, coverage error, and sampling error. These survey errors are interrelated which makes the process of minimising their impact on the TSE difficult, expensive and time consuming. Therefore, survey designers concentrate their efforts in reducing errors depending on their relative impact on survey estimates and the costs associated with reducing these effects.

The initial focus of this chapter is on the nonresponse and measurement error components of TSE because these are the primary focus of the three empirical chapters. Nonresponse errors arise when data is not collected on all persons in the sample and is influenced by respondents, interviewers, mode of data collection and survey design features such as incentives and sponsorship of the study (Bethlehem, Cobben, & Schouten, 2011; Groves, 1989; Groves et al., 2011). On the other hand, measurement errors arise from inaccuracies recorded in the survey instruments due to the effects of respondents, interviewers, questionnaires and mode of data collection (Groves, 1989; Lessler & Kalsbeek, 1992). Therefore, a clear understanding of survey nonresponse and measurement errors and factors that influence their occurrence is of crucial importance.

The later sections of this chapter discuss the literature around the concept of Total Survey Error (TSE) and its influence on data quality. Specifically, nonresponse and measurement errors are considered. Following this, factors that influence TSE are described, and in particular those factors that are the focus of the three papers in this thesis: interviewers, incentives, and mode of data collection. This is followed by a section that describes the methodologies used in the thesis: response propensity models, Bayesian modelling, multilevel modelling, and propensity score matching. Next, an overview and summary of the three papers is presented.

1.1 Survey Errors

The TSE framework was developed by Groves (1989) and consists of a set of principles, methods and processes that minimise TSE for key estimates within the budget allocated. The application of the TSE paradigm starts by identification of the major sources of errors at each stage of the survey process. Survey resources are then allocated to reduce these errors to the extent possible within budgetary and time constraints (Groves, 1989). The TSE framework defines survey quality as the estimation and reduction of the mean squared error (MSE) of statistics of interest. MSE is the expected squared differences between an estimate of the population parameter $\hat{\theta}$ and the actual value of the population parameter θ and is defined as:

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 \qquad (1.1)$$

which decomposes into the sum of variances and squared bias

$$MSE(\hat{\theta}) = B^2(\hat{\theta}) + var(\hat{\theta}) \qquad (1.2)$$

A small MSE indicates an adequacy of survey quality (Biemer, 2010; Groves, 1989). However, the computation and application of MSE is complicated because of the different sources of survey errors which are difficult to distinguish and separate (Biemer, 2010; Groves, 1989; Vehovar, Slavec, & Berzelak, 2012). Also the true scores used in bias estimation are often unknown and they need to be estimated from a census or from ‘gold standard’ criterion , which are not always available (Vehovar et al., 2012). Lastly, the application of MSE is made difficult by many parameters that are often calculated differently across different surveys (Vehovar et al., 2012). Despite these challenges of applying MSE, the TSE approach has been shown to be a useful framework for understanding and evaluating survey error sources and their relative magnitude.

The development of the TSE approach has taken more than 50 years. First, Neyman, (1934) elucidated the sampling theory positing that one could represent a larger population with a probability sample. Neyman proved that sampling error could be measured by calculating the variance of the estimator. Then, Deming (1944) showed that surveys contain multiple sources of error and not only sampling errors. Kish (1965, pp. 514-524) provided the first representation of survey errors in terms of both sampling and non-sampling error. According to Kish, total error in surveys can be obtained by combining the variable errors (VE) and bias. This can be defined as:

$$Total\ Error = \sqrt{VE^2 + Bias^2} \qquad (1.3)$$

where $Bias$ is the deviation of the average survey value from the true population values and arise mostly from nonsampling sources (i.e. measurement biases). On the other hand, variable errors are assumed to be random and are mostly caused by sampling errors. The Kish formulation usually focuses on biases as illustrated in the Figure 1-1:

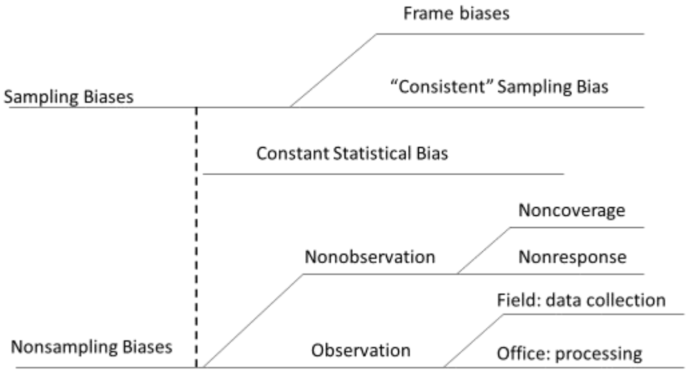


Figure 1-1: Schematic Presentation in Kish of Biases in Surveys, adapted from Kish (1965)

Kish notes that "frame biases" are caused by the unequal selections of the units into a sample and can be adjusted using selection weighting. By "constant statistical bias" Kish meant biases which arise in statistical estimation such as using mean ratio as an estimator of population mean and use of median to estimate the mean of a skewed distribution. The "Constant statistical bias" affects samples of any size and population values based on complete coverage. Finally, Kish notes that nonsampling biases are caused by observation and nonobservation errors. However, Kish fails to note that nonobservation errors are basically sampling errors. Therefore, Kish formulation focused mainly on biases caused by sampling error because they can be reduced using selection weighting in the probability samples.

Dalenius (1974) worked on further development in the theory of survey errors and introduced the term “total survey design”. The “total survey design” refers to essential survey conditions that define the fixed properties of the data collection over all possible implementations. Five years later, Anderson, Kasper, & Frankel (1979), provided an enhanced decomposition of TSE based on the variance and bias, then by sampling and non-sampling, and lastly by observational and non-observational errors. However, Anderson et al. (1979) were not successful in accounting for the consistent statistical bias arising from the inherent properties of the estimate.

Groves (1989) produced a more complete treatment of survey errors and the corresponding cost implications of attempting to reduce them. He proposed that the costs for different survey designs vary and the aim of the survey methodologist is to identify the one with optimal characteristics within the resources available. Groves presented an enhanced nested structure of total survey errors within the MSE based on the conceptual framework of Kish (1965) and Anderson et al. (1979). Groves (1989) defined the MSE as the sum of variance and squared bias components. The variance component comprises sampling errors arising from differences between the recorded value of a survey variable and a “true” value; while squared bias consists of non-sampling errors that arise during the implementation of survey design. Additionally, Groves provided a clear distinction between errors of observation that are caused by coverage, nonresponse, sampling and measurement errors

Lessler & Kalsbeek (1992) advanced the concept of “total survey design” initially introduced by Dalenius (1974) and suggested the need to incorporate frame errors, sampling errors, nonresponse errors and measurement errors when designing surveys. Biemer & Lyberg (2003) extended the list of survey errors by Groves (1989) and included specification error. They defined specification error as the difference that occurs when the concept implied by the survey question and the concept that should be measured differ. In addition, Biemer & Lyberg (2003) integrated the concept of ‘process quality’ within the total survey error framework. The process quality concept involves the strategies adopted during the survey aimed at improving the quality of the survey data and minimising inefficiencies in a survey process. These strategies include the use of paradata (Groves & Couper, 1998), responsive designs (Groves & Heeringa, 2006), and adaptive designs (Schouten, Calinescu, & Luiten, 2012).

Weisberg (2005) extended the survey error approach to include survey related effects that cannot be minimised in any way because of their context-dependent property. For example, questions that appear first in a questionnaire may influence answers to subsequent questions. Naturally, it is hard to remove such question order effects in a survey regardless of the amount of resources spent on them. Besides, the TSE framework has become even more complex due to new error structures as new modes of data collection are introduced (Biemer, 2010; de Leeuw, 2018). These changes are mostly driven by increasing survey costs coupled with limited budgetary allocations making the use of costly modes associated with high quality data almost unsustainable (de Leeuw, 2018).

The TSE, as an indicator of data quality was later extended by Biemer & Lyberg (2003) to incorporate data accuracy. Survey data accuracy is defined in two dimensions: statistical and non-statistical (Biemer, 2016). The statistical dimension explains data quality in the context of accuracy of estimates, which is defined as the difference between the estimate and the true parameter value. The non-statistical indicators can be viewed as constraints and they include relevance, timeliness, accessibility, coherence, completeness, credibility, interpretability, confidentiality protection and comparability. The survey quality framework that incorporates both statistical and non-statistical dimensions is referred to as Total Survey Quality (TSQ) (Biemer & Lyberg, 2003). The TSQ approach underlines the need to consider usability of the survey results when designing and conducting surveys.

Although the TSE framework provides a good representation of survey errors, it is very difficult to implement in practice. Therefore, survey methodologists must decide which errors to prioritise when reducing TSE because concentrating on one error implies fewer resources are available to minimise other errors. Also, a reduction in one source may increase other survey errors and a trade-off is required. The next section covers the components of the TSE.

1.2 Components of Total Survey Error

The goal of an optimal survey design is minimising TSE subject to budgetary costs and timeliness constraints dependent on the survey quality requirements (Groves, 1989). This requires careful planning to ensure an optimal allocation of resources to the various stages of survey designs (Biemer & Lyberg, 2003, pp. 351-376). This ensures that major sources of survey errors are controlled to acceptable levels. It is practically impossible to have an error free survey even under the best circumstances. Therefore, trade-offs must be made when deciding which errors to control. For example, an intention to increase response rates by providing monetary incentives, means that the sample size has to be reduced to remain within budget. This results in a trade-off of bias against precision. Also, the costs allocated to other aspects of survey, such as the training of interviewers, have to be reduced which in turn may impact survey quality negatively. To make optimal designs intended to reduce the overall TSE requires an understanding of the sources and drivers of the survey errors.

Figure 1.2 presents sampling (i.e. representation) and non-sampling (i.e. measurement) errors which constitute TSE. The green ellipses highlight the nonresponse and measurement errors that are the main focus in this section. Sampling errors include coverage error, sampling error, and nonresponse error. Coverage error arises from the failure of the target population to coincide with the population sampled. The unrepresentative nature of the sample taken results in a sampling error, while a nonresponse error is when sample members do not respond to survey. Sampling errors can be controlled in surveys by adjusting sample sizes.

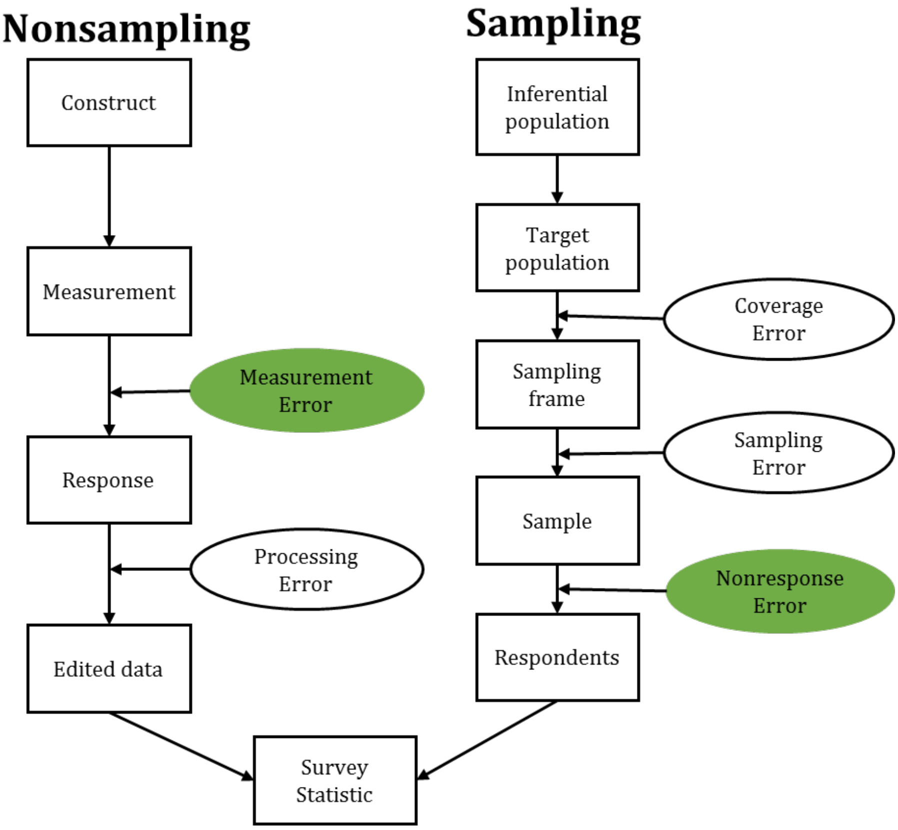


Figure 1-2: Total Survey Error framework, adapted from Groves (1989)

Nonsampling errors are a product of data collection, data processing and estimation processes. They are comprised of measurement and processing errors. Measurement error arises from differences in responses from the true value in a survey process. Differences in measurements may be caused by interviewers, respondents, questionnaires, and modes of data collection. The process of editing, entering, coding and tabulating survey data results in processing error. In this thesis, the review will be limited to nonresponse and measurement errors.

1.2.1 Nonresponse

Survey nonresponse occurs when a sampled unit fails to provide an interview at all (i.e. unit nonresponse) or does not provide answers to some of the items in the questionnaire (i.e. item nonresponse) (Bethlehem et al., 2011; Groves, Dillman, Eltinge, & Little, 2002; Särndal &

Lundström, 2005). Over the last two decades, survey nonresponse has been increasing in most developed countries (de Leeuw & de Heer, 2002; Levy, Lemeshow, Groves, Kalton, & Rao, 2008). The causes of unit nonresponse include noncontacts, refusals, inability to locate sample units, and inability of sample units to respond due to language barriers, ill health or absence. The causes of item nonresponse include, refusal of sample units to provide answers to questions they are not comfortable with , poor survey design, or failure of interviewers to ask or record questions in an adequate manner (de Leeuw, Hox, & Huisman, 2003; Groves & Couper, 1998; Groves et al., 2002). Unit and item nonresponse in surveys have a negative relationship (Dixon, 2002; Yan & Curtin, 2010). That is, a survey with higher item nonresponse tends to have a lower unit nonresponse and vice-versa. This is because respondents with a lower propensity to participate in surveys may transfer their resistance by answering as few questions as possible when interviewers insist on their participation (Yan & Curtin, 2010).

The main objective of random sample surveys is to estimate population characteristics of interest from the samples generated (Groves & Couper, 1998). Survey nonresponse may distort this requirement in samples leading to lack of representativeness. This is because nonresponse error leads to biased estimates when the values of the statistics computed based only on respondent data differ from those based on the entire sample data (Groves, 2006; Groves et al., 2009). Nonresponse bias is defined as the product of nonresponse rate and the difference between the mean of respondents and nonrespondents and is expressed as:

$$Bias(\bar{Y}_r) = \frac{M}{N}(\bar{Y}_r - \bar{Y}_m) \qquad (1.4)$$

where

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

However, Equation (1.4) assumes a “deterministic” view of survey nonresponse because it assumes that there is a fixed number of respondents and nonrespondents in the population (Groves et al., 2009, pp. 189). However, for a given survey, a sample member can be assigned an unobservable propensity of being potentially a respondent or a nonrespondent, which can be represented by ρ_i (Groves et al., 2009; Lessler & Kalsbeek, 1992). This approach assumes that the decision to participate in a survey follows a stochastic process and can be expressed as:

$$Bias(\bar{Y}_r) = \bar{Y}_m + \frac{\sigma_{y\rho}}{\bar{\rho}} \qquad (1.5)$$

where

$\sigma_{y\rho}$ = the covariance between, y, the variable of interest in survey, and ρ, the propensity to respond, among units of the population;

$\bar{\rho}$ is the mean propensity in the target population and over the sample realisations, given the sample design, recruitment realisations, and recruitment protocol design.

Equation (1.5) is suitable when applied at the design stage of a survey because it treats the likelihood of responding as a random variable which varies over different recruitment protocols (Bethlehem, 2002; Groves, 2006). It is crucial to note that low response rates do not necessarily lead to nonresponse bias (Fricker & Tourangeau, 2010; Groves, 2006; Groves & Peytcheva, 2008; Merkle & Edelman, 2002). Nonresponse bias only occurs when there is a systematic difference in characteristics between respondents and nonrespondents (Groves, 1989, 2006). Therefore, instead of focusing only on response rates to reduce bias, survey researchers should focus on whether response propensity and the survey variable are correlated (Groves et al., 2009).

However, maximising response rate may minimise the chances of respondents being systematically different from nonrespondents and in turn reduce nonresponse bias. Improved response rates also lead to accurate survey estimates of variance (Särndal & Lundström, 2005). Survey methodologists employ a variety of approaches all with an aim of increasing response rates. Some of the strategies applied include: offering incentives, training of interviewers, and use of different modes of data collection designs (Campanelli, Sturgis, & Purdon, 1997; de Leeuw, 2005; Groves & Couper, 1998; Singer, 2002).

The respondents' decision to either participate or not participate in surveys can be explained using three main theories namely: social capital theory (i.e. social context theory) (Putnam, 1995b, 1995a), leverage saliency theory (Groves, Singer, & Corning, 2000), and social exchange theory (Blau, 1964). According to Putnam (1995), social capital refers to those attributes that people gain from community organisations through productive interactions that lead to coordination and cooperation for common benefit. For example, communities with good social interactions tend to have higher levels of trust and cooperation that in turn improve willingness of sample persons to participate in surveys for the common good of community.

The social capital is influenced by characteristics at an individual level such as education level, socioeconomic status, marital status, tenure, and number of children (Heyneman, 2006; Letki, 2006). Heyneman (2006) notes that individuals who are highly educated tend to have wide networks in a community. This leads to overall improved cooperation levels, compared to those achieved with individual who are less educated. Individuals with lower socioeconomic status and crime risky neighbours, tend to have lower response and cooperation rates because of reduced trust in neighbourhoods (Letki, 2006).

On the other hand, Brick & Williams (2012) note that the influence of social capital theory on survey nonresponse is a collective (i.e. community) rather than an individual attribute. They suggest that any loss or gain of social capital may be due to the influence of generational changes over time. This theory has been supported by Tourangeau & Plewes (2013) who note that the decline in associational memberships over time may be attributed to reduced public confidence, which may partly explain the lower response rates experienced in surveys. Putnam (1995) also notes that changes in family structure, whereby most people live alone, and the reduction in community engagements, may have resulted in a decline in trust. This may explain why older people are more likely to participate in surveys compared to young ones. Therefore, social capital theory may provide a possible explanation for the declining response rates in the developed world where community engagement is declining.

Leverage-saliency theory (LST) formulated by Groves et al. (2000) explains how different attributes that influence survey participation may help potential respondents in making decisions about survey requests. According to this theory, decisions of individuals to either participate or not in a survey are influenced by their own characteristics, survey characteristics (i.e. reputation of the organisation conducting the survey and the survey topic), and a chance to receive a monetary reward (i.e. incentive). A potential participant usually accords different weights to these components of influence (i.e. leverage) based on their view of the individual importance of the survey request (i.e. saliency).

A sample unit decides to participate in a survey when the expected leverage and saliency of the survey request yields a net positive utility. One clear application of leverage saliency theory is in the use of incentives to promote survey participation. Offering a monetary incentive has been found to have a positive effect on response rates (Dijkstra & Smit, 2002; Singer, 2002). However, any observed positive effect of a factor diminishes when a survey participant places more weight on other factors (Groves et al., 2000). For instance, incentive salience may diminish when a given sample unit places more emphasis on other factors such as community involvement and interest in the survey topic.

Groves et al. (2000) also found that individuals who are more interested in the survey topic have a positive leverage and apply a greater weight to their participation in survey requests when compared to those who are not interested. Sampled persons also tend to experience a positive leverage on any government and academic sponsored surveys in comparison to surveys sponsored by commercial entities. In summary, LST posits that sample persons make their decision to participate in surveys based only on a few attributes of the survey. LST also provides a framework for how survey organisations and interviewers are supposed to design survey features that are attractive to different subgroups.

Social exchange theory (SET) is based on how people behave in their interactions with one another and how various social norms influence these interactions. The SET developed by Blau (1964) proposes that the decision on whether or not to respond to a survey depends on the belief and trust that the perceived benefits for complying with the survey request exceed the costs in the end. Under SET, an individual only expects a flexible positive return from a survey and this is based purely on trust without any reliance on monetary reward (Stafford 2008).

The norms in communities and organisations hugely influence the flexibility of SET in survey response. For example, changing technology has greatly enhanced communication across the globe that has in turn influenced human social interactions both positively and negatively (Drago, 2015). Survey design practices such as offering incentives may create a sense of obligation for future survey participation, an aspect that reinforces the importance of trust as underlined under social exchange theory (Dillman 2007). Interviewers are supposed to build trust with sample units by clearly communicating to them the nature of any expected benefits accruing from survey participation (Groves & Couper, 1998). It is against the backdrop of this information that survey participants evaluate survey benefits and costs and make their decision either to participate in a survey or not.

Several factors ranging from socio-demographic, economic, and political environment are associated with survey nonresponse (Groves & Couper, 1996, 1998; Roose, Waege, & Agneessens, 2003). Since there is a substantial literature on factors that are associated with survey nonresponse, this review will only be limited to main factors: gender, age, education, income and urbanicity. In principle, females are reported to have higher participation rates than men in household surveys because they are more likely to interact with nonhouseholders when compared to men (Groves & Couper, 1998; Smith, 1983). However, other studies have found that gender does not have any impact on response behaviour (DeMaio, 1980; Roose et al., 2003).

Accurate assessment of the impact of age on survey participation is much complicated because of the opposing forces related to age (Goyder, 1987). Most of the empirical evidence shows that response rates tend to decline linearly with increasing age because older respondents tend to be more socially isolated leading to higher non-cooperation rates (DeMaio, 1980; Goyder, 1987; Groves & Couper, 1998). On the other hand, Groves & Couper (1998) also note that older people are easy to contact because of their reduced mobility and lower employment which may impact positively on survey response. In addition, older people are more likely to participate in surveys because they have greater civic and social responsibility when compared to younger people (Groves & Couper, 1998). Therefore, these opposing forces makes it difficult to accurately correlate age and survey response.

People with lower education attainment and in lower social class are often associated with lower survey participation rates (Roose et al., 2003). This is because they feel that surveys are only serving the interests of those people who are well-educated and in higher social class (Groves & Couper, 1998; Roose et al., 2003). In addition, lower educated people tend to feel less qualified to successfully complete surveys. Persons at lower and higher income levels are often associated with lower response rates (Groves et al., 2009; Holbrook, Krosnick, & Pfent, 2007). This is because the people in lower income levels are hard to find and are likely to refuse a survey request due to their suspicions of government and strangers (Holbrook et al., 2007; Schejbal & Lavrakas, 1995). On the hand, persons in higher income levels are socially isolated because their homes are inaccessible due to locked gates (Holbrook et al., 2007). People living in urban areas have lower response rates than those in rural areas (Couper & Groves, 1996; Goyder, Lock, & McNair, 1992). This is because of the higher crime rate and weak community belonging which are often associated with urban areas.

It is usually challenging to predict survey nonresponse robustly because of diverse and temporal changing factors that influence survey participation at sample individual levels. This has encouraged survey methodologists to use aggregate data for estimating response propensities. Naturally, it is possible to predict response propensities and percentage response rates for given groups with common background characteristics. However, it may be challenging to predict changes in individual response propensities, due to lack of personal response data. Additionally, it is inherently hard to make response predictions at an individual level because of the many factors that are not generalisable. This issue has engaged survey methodologists over the years and has promoted extensive research into the ways of improving response predictions using response propensity modelling.

One of the main research areas that is attracting attention involves ways of improving the predictive power of survey response models (Durrant, Maslovskaya, & Smith, 2015, 2017). Also, it is crucial to understand whether measures undertaken by survey organisations to improve response rates such as training of interviewers, offering incentives, and data collection using different modes are paying off (de Leeuw, 2005; Groves & Couper, 1998). The effectiveness of these approaches in reducing survey nonresponse can be assessed using response propensity models (Durrant & Steele, 2009; Särndal & Swensson, 1987). Response propensity models are widely used to explain nonresponse, incentive effects, interviewer effects and mode effects (de Leeuw, 2005; Mcgrath, 2005; Schnell & Trappmann, 2006).

The effectiveness of response propensity models in explaining the drivers behind the survey response process, is hindered by their low predictive power. This is because the available auxiliary variables are not sufficiently correlated with survey response and other key survey variables (Kreuter, Olson, et al., 2010; Olson & Groves, 2012; Peytcheva & Groves, 2009; Sinibaldi, Trappmann, & Kreuter, 2014). Therefore, survey researchers are continually looking for ways of improving the predictive power of response propensity models by collecting new sources of information for both respondents and nonrespondents, such as paradata and by exploring statistical approaches such as the Bayesian approach (Beaumont, 2005; Couper, 1998; Durrant & Kreuter, 2013; Kreuter, Couper, & Lyberg, 2010; Kreuter, Olson, et al., 2010; Schouten, Mushkudiani, Shlomo, & Durrant, 2018; Wagner, 2016).

1.2.2 Measurement Error

Measurement error arises when the obtained survey measure (i.e. response) does not reflect the "true" value of the underlying construct[1](Lessler & Kalsbeek, 1992). Suppose that y_i is the response obtained from the i^{th} respondent and U_i is the value of the characteristic for the i^{th} respondent. Then a measurement error model takes the form:

$$y_i = U + \epsilon_i \tag{1.6}$$

where ϵ_i is the random error for the i^{th} respondent. If the $\epsilon_i's$ are independent from U , then the resulting measurement error model is known as a classical measurement model (Groves, 1989). However, classical measurement models are overly restrictive in surveys because they do not account for possible biases in questions of underlying constructs (Groves, 1989; Pischke, 2007). To overcome the drawback of the classical measurement model, a multiple factor model is used (Groves, 1989). The multiple factor model accounts for biases in questions of underlying constructs and allows questions to be influenced by various methods of measurement (Groves, 1989; Pischke, 2007). The multiple factor model takes the form:

$$y_{ij} = U_i + M_{ij} + \epsilon_{ij} \tag{1.7}$$

where y_{ij} is the response obtained from the i^{th} respondent using j^{th} method, U_i is the true value of the characteristic for the i^{th} respondent and M_{ij} is the effect on response of the i^{th} respondent using j^{th} method and ϵ_{ij} is the deviation for the i^{th} respondent from the average effect of the j^{th} method (Groves, 1989).

Survey measures taken from respondents are subject to both systematic and random measurement errors (Groves, 1989). Systematic measurement errors are correlated across observations and do not have a zero-expected value (i.e. the measurement errors are

[1] Construct are the elements of information sought by researchers during the survey (Groves et al., 2009)

particularly wrong in particular direction). On the other hand, random measurement errors occur when responses varies from true values with no consistent pattern (i.e. independently) and have an expected value of zero. Measurement errors arise from various sources namely: interviewers, respondents, modes of data collection, and the questionnaires (Biemer, Chen, & Wang, 2013; Groves, 1989). The errors arising from the information systems and interviewer settings are also considered as measurement errors by Biemer & Lyberg (2003). The sources of measurement errors are interrelated, and errors contributed by one source may be influenced by changes in other sources. For example, measurement errors arising from the respondents are usually affected by whether the mode is interviewer or self-administered. For that reason, a clear understanding of the sources of measurement errors may facilitate the design of optimal surveys which in turn improves data quality.

First, the questionnaire design causes measurement error because of the differences in length, structure, and the context of the questions (Lyberg & Kasprzyk, 2011; Sirken et al., 1999; Sudman, Bradburn, & Schwarz, 1996). A good questionnaire is one that conveys the meaning of the concepts in such a way that systematic and random errors are minimised within the constraints of data collection (Sudman et al., 1996). Despite this, questionnaires that are well designed may still be susceptible to measurement errors. This has made it necessary for survey designers to conduct questionnaire pre-tests and other evaluations prior to the field work (Sudman et al., 1996). Pre-tests aim to identify problems in the questionnaire that were not noticed during the design stage and which may have a negative impact on the survey process. Converse & Presser (1986) recommend at least two pre-tests for a new survey. The first pre-test aims to test the initial wording of the questionnaire while the second acts as a rehearsal for the field work and assesses whether the changes implemented in the first pre-test were effective.

In face-to-face interviewing, pre-testing is usually carried out using the so called “cognitive interviewing techniques” (Campanelli, 1997; Jobe & Mingay, 1991). Cognitive interviewing techniques usually focus on the cognitive process that respondents use to answer survey questions. The behaviour coding schemes are also used in evaluating the effectiveness of the questionnaire (Goldenberg et al., 1997). Behaviour coding scheme may be able to reveal the questions that the interviewers and respondents might have difficulties with during the response process for both interviewer mediated and self-administered surveys. Based on the responses received from the pre-test, the survey designers can improve the questionnaire to assist the respondents in comprehending the researcher’s intended meaning. The approaches adopted in developing effective questions to solicit information, depend on the survey topic. Some methods used include shortening the questions and reducing the number of response options with an aim of reducing response burden. This is because reduced response burden is associated with an increase in data quality (Diehr, Chen, Patrick, Feng, & Yasui, 2005; Sahlqvist et al., 2011).

Measurement errors arising from respondents are identifiable through the four distinct cognitive stages originally proposed by Tourangeau, Rips, & Rasinski (2000, pp. 8-16). These stages include: (1) comprehending the question, (3) recalling information, (4) judging the appropriate answer to the question, and (5) editing and communicating the answer. In the first step, the respondent is expected to have some previous relevant knowledge to the survey question for the response process to start. This enables the respondent to assign meaning to the question with respect to each of the words in the question (i.e. comprehension) and instructions contained in the questionnaires. Usually, previous interactions with questions in questionnaires by researchers, interviewers, and respondents may influence the comprehension of the questions. During the second stage, the respondent searches for specific memories of events relevant to the question to retrieve the required information. At the third stage, the respondent determines the most appropriate response to the question based on their judgements regarding the completeness of the retrieved information. During this stage the respondent also takes into account other factors such as social desirability when formulating the response (Cannell, Miller, & Oksenberg, 1981). Finally, the respondent communicates the response to the question to the interviewer or records the response in a self-administered questionnaire.

At each stage of the cognitive process there is a potential for measurement errors to arise depending on the motivation of the respondent, the survey topic, and the difficulty of the questions (Groves, 1989; L. E. Lyberg & Kasprzyk, 2011). It is possible, for example, for a respondent to incorrectly comprehend the question, recall from memory, make a wrong judgement and communicate this wrong judgment as an answer to the question. In some instances, respondents may revise their answers at the judgment stage after considering the risk of answering accurately and honestly due to social desirability. During the cognitive process some respondents may be unmotivated, disinterested in the survey topic, and in a hurry resulting in response styles such as acquiescence and item nonresponse. To ensure that respondents provide accurate responses with reduced measurement error the following three approaches are adopted in surveys. First, respondents are reminded of the importance of committing to provide accurate responses. Second, the length of the questionnaire can be increased to deepen the memory retrieval. However, this approach can be counterproductive because respondents may feel overburdened. Lastly, interviewers are encouraged to probe for answers (Biemer & Lyberg, 2003; Tourangeau et al., 2000). However, the extent to which these approaches are adopted is dependent on the available budget, and the budget determines interviewing time, interviewer training and questionnaire designs.

1.2.3 Interviewer Error

In face-to-face and telephone surveys interviewers play a critical role in the survey process. (Campanelli et al., 1997; Groves & Couper, 1998; Morton-Williams, 1993; West & Blom, 2017). First, interviewers are required to physically locate the sampled households and find the sample member in face-to-face interviewers (Groves, 1989). In addition, interviewers are the medium thorough which the aspects of survey design such as the purpose of the study, sponsor of the study, and any incentives offered are communicated to the sample members. After establishing the initial contact with the respondent, an interviewer is supposed to motivate the respondent to participate in the survey and accurately record the respondent's answers and any other required information, such as interviewer observations. Therefore, it is crucial to clearly understand the role interviewers play in the survey process and how they influence TSE. Interviewers may affect the survey process both positively by increasing response rates and negatively by introducing unwanted measurements errors (Groves, 1989).

Interaction between a sample unit and an interviewer determines whether a sample unit will participate in a survey or not (Groves and Couper 1998). The decision of a respondent to partcipate in a survey can be expressed as a function of interviewer, social environment, and survey design characteristics (Groves, 1989; Groves & Couper, 1998). Interviewer characteristics that influence a sample unit's decision to participate in surveys can be classified into three main categories: socio-demographic, attitudinal and behavioural (Durrant et al., 2010; Groves & Couper, 1998; Hansen, 2006; Hox & de Leeuw, 2002; Lavrakas, 2008). Physical attributes of interviewers are directly observable by respondents and include age, gender, and ethnicity. Interviewers' attitudinal and behavioural characteristics are not directly observable by respondents, but they are capable of perceiving them. They include interviewer confidence, social skills (i.e. persuasiveness, probing and friendliness), expectations, knowledge, stereotypes about target population, and attitudes towards survey topic (Schaeffer, Dykema, & Maynard, 2010). The unobservable interviewers' characteristics are affected by features of survey design such as the mode of data collection, the use of incentives, the extent and type of training, and the survey topic (Campanelli, Sturgis, and Purdon 1997; Groves and Couper 1998).

The role of interviewers in survey nonresponse has been examined in several studies (Durrant et al., 2010; Groves & Couper, 1998; Pickery & Loosveldt, 2002; West & Blom, 2017). The interviewer characteristics that influence survey response and that have attracted considerable attention include: age (Campanelli & O'Muircheartaigh, 1999; Durrant, D'Arrigo, & Steele, 2011; Singer, Frankel, & Glassman, 1983), gender (Groves, O'Hare, Gould-Smith, Benkí, & Maher, 2007; Lessler & Kalsbeek, 1992), race (Merkle & Edelman, 2002), experience

(Singer et al., 1983; Snijkers, Hox, & de Leeuw, 1999), and skills (Campanelli et al., 1997; Morton-Williams, 1993). These studies have shown mixed relationships between interviewer characteristics and survey response rates (Groves & Couper, 1998; Schaeffer et al., 2010). Female interviewers are perceived to be more friendly and approachable, and are therefore capable of attaining higher cooperation and response rates than male counterparts (Campanelli & O'Muircheartaigh, 1999; Fowler & Mangione, 1990; Morton-Williams, 1993). However, a literature review by Lessler & Kalsbeek (1992) found that there is little systematic evidence supporting the assertion that females achieve significantly higher response rates than males.

Campanelli & O'Muircheartaigh (1999) noted that older interviewers are more likely to achieve slightly higher response rates than younger ones, although Morton-Williams (1993) found no significant association between interviewer age and survey nonresponse. Studies on the effects of interviewer race on survey nonresponse, show that respondents tend to be more confident and cooperative, on sensitive questions, when interviewed by someone with whom they share the same characteristics (Lavrakas 2008). Durrant et al. (2010) found that matching, based on gender and education tends to reduce survey refusal rates. On the other hand, Merkle and Edelman (2002) found no significant interaction between interviewer race and response rates. To summarise, it is not clear the influence of interviewers' socio-demographic characteristics in survey response because the empirical evidence shows mixed results.

The studies examining the attitudinal and behavioural effects on survey nonresponse show mixed results (Blom & Korbmacher, 2013; Durrant et al., 2010; Groves & Couper, 1998; Hox & de Leeuw, 2002; Jäckle, Lynn, Sinibaldi, & Tipping, 2011). One important thing to note is that implicit assessment of the effects of attitudinal and behavioural characteristics on survey nonresponse across studies is made difficult by the variety of measurements used across surveys. For example, interviewer experience may have two measures. The first one is based on the number of years practised in an organisation, and the second one the number of organisations an interviewer has worked for.

Starting with interviewer experience, it has been found that interviewers with more experience tend to have higher cooperation and response rates compared to less experienced ones (Groves & Couper, 1998; Jäckle et al., 2011). However, Durrant et al. (2011) found that interviewer experience is not that important when establishing contacts with respondents, after controlling for any other socio-demographic characteristics of interviewers. Groves & Couper (1998) found a negative relationship between the number of organisations an interviewer has worked for and the survey response rate achieved. Merkle & Edelman (2002) found that no relationship exists between the number of surveys an interviewer had worked for and survey response rates.

Interviewer skills coupled with positive attitudes and expectations are associated with higher response rates (Campanelli et al., 1997; Durrant et al., 2010; Groves & Couper, 1998; Singer et al., 1983). Groves & Couper (1998) and Durrant et al. (2010) note that interviewers who are more confident when interacting with respondents , and are persuasive and persistent in terms of asking for an answer tend to have higher response rates. Campanelli et al. (1997) also reported that interviewers who are persistent in making follow up calls tend to have higher response rates. de Leeuw, Hox, Snijkers, & de Heer (1998) also found that interviewers who are more inclined to persuading survey members to participate in surveys tend to have relatively higher response rates. Hox & de Leeuw (2002) note that interviewer personalities tend to be better predictors of survey response than socio-demographic characteristics. This assertion was supported by Jäckle et al. (2011) and Yu, Liu, & Yang (2014) who found that interviewers who are extrovert and assertive tend to have higher response rates. Interviewers with better tailoring ability and friendlier introductions also tend to have higher response rates (Cialdini, 1984; Lemay & Durand, 2002). For example, interviewers may tailor their introductions in such a way that they make incentives very clear to respondents leading to improved survey cooperation (Cialdini, 1984; Groves & Couper, 1996).

One of the approaches used by survey organisations to improve response rates obtained by interviewers involves offering training. Mayer & Brien (2001) found that offering extra training for interviewers may lead to a reduced number of survey refusals. Groves & Mcgonagle (2001) found that the training of interviewers not only increased response rates but also reduced variations between interviewers. This is an important aspect because it leads to data of better quality. However, a critical knowledge gap between interviewer characteristics and the use of incentives in improving survey response still exists.

The role of survey interviewer as a source of measurement error has been studied extensively over the years (Boyd Jr. & Westfall, 1955; Groves, 1989; Hansen, Hurwitz, & Bershad, 1961; Kish, 1962; O'Muircheartaigh & Campanelli, 1998; West & Blom, 2017). The main factors that influence interviewer effects on measurement errors include: (1) socio-demographic characteristics of interviewers and respondents, (2) interviewer expectations, (3) design of the questionnaires and question types, and (4) survey settings (Groves, 1989). The socio-demographic characteristics of the interviewers and respondents yield a greater influence on the measurement errors through the cognitive response process than other effects (Fowler & Mangione, 1990; Tourangeau et al., 2000). The response pattern regarding the interactions of interviewer and respondent characteristics vary, they depend on the questions and topics, and cannot be generalised across all questions in a survey (Dykema, Lepkowski, & Blixt, 2012; Schaeffer et al., 2010). The differences also occur when the subject matter is related to the respondents' characteristics. For example, the gender of the interviewer may influence the response patterns , these responses may differ between females and males on questions about gender roles (Ballou & DelBoca, 1980; Huddy, Billig, Bracciodieta, Moynihan, & Pugliani, 1997).

The interviewer expectations regarding answers and reactions of the respondents to given questions may lead to measurement errors as interviewers may try either rephrasing the question or skipping it (Biemer & Lyberg, 2003; Groves, 1989). Naturally, interviewers especially experienced ones, expect respondents to react negatively to sensitive questions. Consequently, they may either skip these questions or accept the 'don't know' responses and refusals quickly, without further probing. Third, the design of the questionnaire influences the measurement errors because interviewers vary in the way they ask questions with different levels of complexity (Mangione, Jr Fowler, & Thomas A., 1992). Usually the decision whether or not an interviewer is expected to probe for clarification and provide feedback on respondents' responses, depends on the questionnaire design, the survey questions and associated instructions (Groves, 1989).

To reduce interviewer effects on measurement errors most surveys follow standardised interviewing techniques (Fowler & Mangione, 1990). One pitfall associated with standardised interviewing is the possibility of a reduction in response accuracy (Suchman & Jordan, 1990). This is caused by the limited conversations with which interviewers can engage with respondents, especially on questions about attitudes, sensitive, open-ended and those with difficult items (Fowler & Mangione, 1990). Alternatively, interviewers may use conversational interviewing where they can deviate from the standardised script and engage respondents in a conversation (Suchman & Jordan, 1990). This ensures that respondents are guided to correct and consistent interpretation of questions leading to improved response accuracy (Dykema et al., 2012). The drawback associated with flexible interviewing is the varying probing ability of interviewers, which may in itself contribute to measurement errors (Groves, 1989). In summary, interviewers play a significant role in ensuring that response quality is realised, and they need to be provided with proper training to reduce measurement errors. It is crucial to note that both standard and flexible interviewing will only produce data of high quality when respondents can accurately understand and map the concepts of the questions into their own particular situations.

1.2.4 Incentives

Incentives are used in surveys to motivate sample members to participate (Mizes, Fleece, & Roos, 1984; Singer, 2002; Singer, Hoewyk, Gebler, Raghunathan, & Mcgonagle, 1999). The role of incentives in motivating response has been emphasised in three theories: leverage saliency theory (LST) (Groves et al., 2000), social exchange theory (SET) (Blau, 1964), and economic exchange theory (Biner & Kidd, 1994). The three theories have been discussed earlier in the section of survey nonresponse. Incentives are either non-monetary or monetary payments. Non-monetary incentives include gifts (i.e. pens, calendars, or diaries), lotteries, and summaries of survey results (Lavrakas 2008). Monetary incentives, either prepaid or promised, tend to yield higher response rates than non-monetary gifts (Cantor, O'Hare, & O'Connor, 2008; Church, 1993; Singer, Groves, & Corning, 1999). Prepaid incentives are more effective in increasing survey response than promised incentives (Church, 1993; Singer, Hoewyk, et al., 1999; Singer & Ye, 2013; J. Yu & Cooper, 1983). The magnitude of the effect of the incentive on response rates increases with the size of the incentive (Singer, Hoewyk, et al., 1999). However, this relationship is curvilinear, with the size of the increase in the response rate declining with additional increases in the value of the monetary incentive (Cantor et al., 2008; Mercer, Caporaso, Cantor, & Townsend, 2015).

The existing literature attributes the positive effects of incentives to the behaviour and attributes of respondents (Currivan, 2005; Patrick, Singer, Boyd, Cranford, & Mccabe, 2013; Singer, 2002; Singer, Hoewyk, et al., 1999). For example, Currivan (2005) investigated the impact of using refusal conversion incentives on the composition of the sample using data from the New York Adult Tobacco Survey (NYATS). In this survey, respondents were offered an incentive of $20 if they initially refused to participate. It was found that these refusal conversion incentives increased the proportions of respondent who were older, did not have a college degree, and were unemployed. Berlin et al. (1992) and Petrolia & Bhattacharjee (2009) found that sample members with higher levels of education tend to be overrepresented in non-incentive groups compared to incentive groups.

Incentives have also proved successful when used to draw in particular units with specific characteristics from the sample, who would otherwise have refused to participate (Shettle & Mooney, 1999; Singer, Hoewyk, & Maher, 2000). For example, Shettle & Mooney (1999) found that incentives are effective at converting refusals from minority ethnic, lower levels of education, and lower income groups in longitudinal studies. However, Cantor et al. (2008) found that pre-paid incentives have no effect on sample composition of the participants after reviewing 23 Random Digit dialling (RDD) experiments. On respondent behavioural aspect, Singer et al. (1999) investigated the effect of the sample members reaction to differential incentives offered in surveys using the Detroit Area Study (DAS). They found respondents to be sensitive about the fairness of using differential incentives, although this sensitivity had no significant influence on the willingness of the respondents to participate in future surveys.

Many studies have put their focus on investigating the interactions between incentives and the behaviour of respondents. Despite this, none of the studies have focused on the effects interviewers may have on the effectiveness of incentives in interviewer-mediated surveys. Normally, interviewer's attitudes and behaviour towards a sample member may be influenced by the knowledge of whether they have received an incentive or not. This may in turn influence the likelihood that he/she will secure survey cooperation or not. The effects of incentives on interviewers may be either positive or negative (Singer, 2002). Interviewers may be more confident in approaching sample members if they know that they have been or will be offered incentives. This may lead to improved response rates because confident interviewers have been found to have higher response rates (Durrant et al., 2010; Groves & Couper, 1998; Hox & de Leeuw, 2002). Interviewers also expect respondents who have received incentives to be more cooperative because they are being rewarded for their efforts (Singer et al., 2000; Singer & Maher, 2000).

Singer et al. (2000) carried out an experiment to determine whether the effect of prepaid incentives on survey response is influenced by the interviewers' knowledge that incentives have been delivered to sample members. The sample members were randomly divided into three groups in a RDD survey. One group was sent an advance letter and $5 with interviewers being kept blind (i.e. unaware) of the incentive offered. The second group of sample members also received an advance letter and an incentive of $5, while the third group received only an advance letter. The incentive condition in the second group was known by interviewers through the information presented on CATI screens. Singer et al. found that sample members who were offered advance letters and $5 incentive had higher response rates compared to those who received an advance letter only. In addition, they found that interviewers blinded of the incentive offered (i.e. group 1) had higher cooperation rates of 85% compared to 81% of those who were aware of the incentive offered (i.e. group 2). This shows that interviewer knowledge about the incentive offered to sample members does not lead to higher cooperation rates. Probably interviewers do not feel the same need to motivate incentivised sample members to participate in a survey because they are being 'paid' for their efforts.

Lynn (2001) investigated interviewer expectations and attitudes towards incentive effects in an experimental study. The study involved offering a conditional incentive of $10 to any member in the household who completed two diaries and an interview. The interviewers were then allocated an equal number of incentivised and non-incentivised households. The number of interviewers involved in the study was 20. These interviewers were then questioned at the end of the study period about the survey experience. They also provided feedback on their perceptions about the use of incentives, using a structured questionnaire. Lynn found that half of the interviewers felt incentives have little or no effects on the improvement of cooperation and response rates. The other half of the interviewers had an impression that incentives had a negative effect on cooperation and response rates. However, the joint influence of interviewer and incentives on survey participation has not yet been investigated. This gap in knowledge will be addressed in the second paper.

1.2.5 Mixed-Mode

Over the last thirty years the use of different modes of data collection has been on the rise, which in turn has affected both who responds and how they answer (de Leeuw, 2005, 2018; Dillman, Smyth, & Christian, 2009). This is mostly driven by technological advancements and societal changes. The motive to offer an alternative mode of data collection come from consideration of data quality and cost due to the increased costs of traditional methods and cuts in survey budgetary allocations (Couper, 2011; de Leeuw, 2005, 2009; Klausch & Schouten, 2015). Additionally, there has been an increase in cross-national surveys and countries tend to have differences in survey traditions and characteristics (de Leeuw, 2018). The use of different modes of data collection together has been shown to lead to improved coverage and response (de Leeuw, 2005, 2018; Dillman et al., 2009). However, there is a hidden price to the use of different modes of data collection in terms of data quality, especially in the reporting of sensitive questions. For example, response rates and data quality differ substantially when self-administered and interviewer administered modes are compared (Burkill et al., 2016; de Leeuw, Hox, & Kef, 2003; Newman et al., 2002; Roberts, 2007).

The benefits attributed to the use of different modes of data collection depend on the choice of the modes used. The first comprehensive study discussing mixed-mode designs was by Dillman & Tarnao (1989). They noted that using different modes of data collection may improve coverage and response rates in face-to-face interviews, mail, and telephone surveys. However, they also noticed that using different modes may lead to data comparability issues. Since then the use of mixed-modes for data collection has increased in surveys, and has become a norm (Biemer & Lyberg, 2003; de Leeuw, 2005, 2018; Tourangeau, 2017). The application of mixed modes of data collection usually takes three forms, namely:(1) contact by different modes, (2) different modes for specific questions, and (3) different response modes for different respondents (de Leeuw, 2018; Dillman et al., 2009).

Different modes are used to contact the respondents with the aim of obtaining a good representative sample. For example, the recruitment of probability based online surveys sample units involves sending advance letters to the listed addresses informing the recipients of the survey and communicating any special features of the design (Blom, Gathmann, & Krieger, 2015; Dillman, 2007). In addition, many studies using face-to-face interviewing usually send advance letters ahead of the time to the sampled addresses detailing the various aspects of survey such as sponsor, topic, any incentives offered, and the expected dates of interviews (Lavrakas, 2008).

The different response modes for different respondents may be implemented using two different ways: concurrent and sequential designs (de Leeuw, 2005; Dillman et al., 2009). In concurrent mixed-mode design, different modes are offered at the same time during the survey. The aim of using concurrent design is to overcome any coverage problems and allow for data collection in different countries which have different traditional main modes (de Leeuw, 2018). Concurrent designs are therefore mostly implemented in surveys conducted across countries and among special groups. The sequential design involves following up nonrespondents using a different mode from the one in which they were initially requested to provide a response. For example, a survey may start data collection using a cost effective mode and then follow up the nonrespondents with a more expensive mode to reduce nonresponse (Revilla, 2010; Sakshaug & Eckman, 2017; Ziegenfuss, Burmeister, Harris, Holubar, & Beebe, 2012).

Data collection modes can be classified into two main categories: interviewer mediated and self-administered modes (P. P. Biemer & Lyberg, 2003; Groves, 1989; Wolf, Joye, Smith, & Fu, 2016). For interviewer mediated surveys, interviewers are involved in administering the survey questions either face-to- face or by telephone. On the other hand, self-administered surveys such as online and mail surveys, are designed in such a way that respondents 'are able to complete questionnaires without any interviewer involvement. There is substantial literature on how different methods of data collection influence survey data quality in the context of selection and measurement effects (de Leeuw, 2005, 2018; Dillman, 2002; Jäckle, Roberts, Lynn, Robert, & Lynn, 2010). In this thesis, the focus will be limited to face-to-face interviews and online probability surveys. For online probability surveys, sample units are usually selected randomly from a list of addresses obtained from postcode address files or pre-recruited from a panel survey (Toepoel, 2012). A pre-recruited panel survey involves pre-recruiting survey participants from other existing surveys conducted in other modes such as face-to-face interviews selected via probability-based sampling.

The effect of face-to-face and online probability surveys, on data quality has been assessed in numerous studies (de Leeuw, 2005, 2018; Dillman, 2002; Jäckle, Robert, & Lynn, 2010). Online surveys are less costly, enable fast data processing, and are flexible in terms of providing more complex displays such as videos (Beebe, Mika, Harrison, Anderson, & Fulkerson, 1997; Bethlehem & Biffignandi, 2011; Tourangeau, Conrad, & Couper, 2013). On the other hand, face-to-face interviews have higher response rates than online probability samples. This is because interviewers can motivate and easily convince respondents about the legitimacy of the study by highlighting key survey features such as survey sponsor and incentives. The higher response rates in face-to-face interview comes with significantly higher costs (de Leeuw, 2005). However, the fact that lower response rates do not always lead to nonresponse bias makes less costly online surveys a feasible alternative to higher response face-to-face interviews (Fricker & Tourangeau, 2010; Groves, 2006; Groves & Peytcheva, 2008).

The presence of interviewers in face-to-face surveys also leads to lower item-nonresponse rates than in online surveys (Heerwegh & Loosveldt, 2008; Jäckle, Lynn, & Burton, 2015; Lesser, Newton, & Yang, 2012). The presence of interviewers during a survey process keeps survey participants motivated and engaged, ensuring that questions are answered correctly and by the intended persons (Couper, 2011; Holbrook, Green, & Krosnick, 2003; Szolnoki & Hoffmann, 2013). Contrarily, online surveys are completed in a less controlled environment than face-to-face interviews making respondents prone to incidences of item nonresponse and 'don't know' responses. Additionally, the use of the internet is associated with multi-tasking which may distract some of the respondents (Lozar Manfreda & Vehovar, 2002).

Generally, respondents interviewed by interviewers tend to provide answers which they perceive will agree with other members of society (de Leeuw, 2005). Additionally, the presence of interviewers may make respondents take social norms into account when providing answers, resulting in a social desirability bias (Burkill et al., 2016; Heerwegh, 2009; Klausch & Schouten, 2015; Kreuter et al., 2010; Revilla & Saris, 2013; Williams, 2017b). This results in more positive and socially desirable answers by respondents in face-to-face surveys than online surveys (Burkill et al., 2016; Klausch & Schouten, 2015; Schouten, van den Brakel, Buelens, van der Laan, & Klausch, 2013). On the other hand, online surveys are prone to less social desirability bias because they have a higher degree of privacy and respondent is in full control of survey process (Couper, 2011; Dillman et al., 2009).

Currently, many surveys are using mixed-mode designs where different modes of data collection are used in the same study. Combining different modes may have a beneficial effect on survey measurement by exploiting the key strengths of each mode (de Leeuw, 2018). Some surveys are also changing the mode of data collection from the traditional expensive modes to cheaper alternatives. This raises the question of whether using alternate modes of data collection which are cheaper compared to traditional modes such as face-to-face interviews, results in data of equal or better quality. For example, does changing from an interviewer mediated to a self-administered mode lead to data of equal or better quality? (Tourangeau et al., 2013). The literature shows that there is a lower risk of measurement errors when modes that are either self-administered (i.e. online and mail modes;) or interviewer mediated (face-to-face interviewing and telephone) are used together (Couper, 2011; de Leeuw, 1992; Jäckle, Roberts, et al., 2010). However, mixing a self-administered mode and an interviewer mediated mode (i.e. online and face-to-face interviewing) may result in higher measurement differences (Couper, 2011). Therefore, it becomes important to have a clear understanding of measurement differences between different modes, as this will enable well designed surveys.

1.3 Methodology

This section provides an overview of the methodological approaches used in the three papers.

1.3.1 Response Propensity Models

Survey response behaviour is often explored by researchers using response adjustment models (Särndal & Swensson, 1987). Response adjustment models are either classified as deterministic or stochastic models (Särndal & Swensson, 1987). The deterministic model treats survey response as a fixed outcome that can be defined in terms of two non-overlapping strata consisting of respondents and nonrespondents (Särndal & Lundström, 2005). However, a deterministic model is limited by its assumption that each sample unit in the response stratum will definitely participate in a survey and those in the nonresponse stratum will have a zero probability of participating in surveys.

The stochastic model overcomes this limitation of the deterministic model by assigning an unknown response probability of participating in a survey to each survey unit (Särndal & Swensson, 1987). The probability of survey participation is estimated using response propensity models by making use of all available and relevant auxiliary data for the sample units (Pfeffermann & Rao, 2009). Several studies have applied response propensity models for predicting survey response (Durrant et al., 2011, 2015, 2016; Plewis et al., 2012; West & Groves, 2011). They have also been used for developing nonresponse weights (Biemer et al., 2013; Kreuter & Olson, 2011; Little, 1986), for calculating representativeness indicators such as R-indicators and coefficients of variation (CV) (Moore, Durrant, & Smith, 2018; Schouten & Cobben, 2007; Schouten, Shlomo, & Skinner, 2011), and for providing guidance on adaptive and responsive survey designs (Durrant et al., 2011; Groves & Heeringa, 2006).

Response propensities are estimated based on the socio-demographic characteristics of the sampled units which are obtained from a sampling frame, administrative data, and paradata (Bethlehem et al., 2011; Groves & Couper, 1998; Kreuter, Couper, et al., 2010). The auxiliary variables are then used in prediction models of survey response (Blom, Jäckle, & Lynn, 2010; Durrant & Steele, 2009; Groves & Couper, 1998; Hox & de Leeuw, 2002; Pickery, Loosveldt, &

Carton, 2001; Vassallo, Durrant, & Smith, 2016; West & Elliott, 2014; West & Kreuter, 2015). Survey design features such as mode of data collection, use of incentives, and the organisation sponsoring the study are also sometimes included as predictors (Bethlehem et al., 2011; Groves & Couper, 1998). Response propensity is formally defined as the probability that a sample unit responds to a survey request, given the characteristics of such a unit (Bethlehem et al., 2011). Response propensity (RP) models have been mostly employed to investigate how household, interviewer, and survey design characteristics influence survey response (Durrant, D'Arrigo, & Steele, 2013; Durrant et al., 2015, 2017; Durrant & Steele, 2009; Kreuter, 2013; Sinibaldi & Eckman, 2015; Vassallo, Durrant, Smith, & Goldstein, 2015).

Despite the widespread use of RP models for investigating survey nonresponse they tend to have low predictive power in terms of Pseudo R^2 (Groves & Couper, 1996, 1998; Kreuter, Couper, et al., 2010; Olson & Groves, 2012; Olson, Smyth, & Wood, 2012; West & Groves, 2011). Pseudo R^2 is a common measure of the predictive strength of model with either binary or multinomial outcomes to some explanatory variables (Hu, Shao, & Palta, 2006; McKelvey & Zavoina, 1975). Pseudo R^2 is a corresponding indicator for coefficient of determination R^2 in a linear regression model. The values of Pseudo R^2 ranges from zero to one, with zero indicating a model with no predictive power and one indicating a perfect fit. Generally, the low predictive power of RP models may be explained by auxiliary variables which are not sufficiently correlated with key survey variables (Kreuter & Olson, 2011; Olson & Groves, 2012; Olson et al., 2012; Plewis et al., 2012). For example, Olson & Groves (2012) found that the predictive power of RP models investigating within person variations over the data collection period ranged between 2% and 4% in terms of pseudo R^2 . They used the National Survey of Family Growth (NSFG), conducted by the University of Michigan for the National Centre for Health Statistics, and the Wisconsin Divorce Study (WDS), conducted by the University of Wisconsin-Madison.

Olson et al. (2012) also found that the predictive power of the RP models ranged between 3.2% and 7.7% in terms of pseudo R^2 in a study that investigated the effects of mode preference on response, contact, and cooperation rates. They used data from the 2008 Nebraska Annual Social Indicators Survey (2008 NASIS) and the Quality of Life in a Changing Nebraska survey (QLCN). West & Groves (2011) used the National Fertility Survey data from the United States to evaluate an interviewer performance indicator. The models used in their study to predict the likelihood of an interviewee completing a main interview on the next visit or call attempt, had predictive power that ranged between 3.3% and 7.4%. In summary, the predictive power of a RP model ranges between 2% and 8%, which is substantially low.

This limitation of RP models has motivated survey researchers to investigate strategies of improving their predictive power. One of the main strategies adopted to improve predictive power of RP models is the use of survey process data known as paradata (Durrant, D'Arrigo, & Müller, 2013; Durrant et al., 2011, 2015; Sinibaldi & Eckman, 2015; West, 2013; West & Sinibaldi, 2013). Paradata are broadly classified either as system generated or interviewer generated (Smith, 2011). The system generated paradata include call records (i.e. dates, times, and counts of call attempts) and keystrokes (i.e. audit trails), device type, question navigation (i.e. breaks offs, mouse clicks, change of answers, typing and keystrokes) (Callegaro, 2013). The interviewer generated data is related to observation information about the demographic of respondents and the conditions of the neighbourhoods (Kreuter, Couper, et al., 2010; Kreuter & Olson, 2013; West, 2013). They include variables such as condition of the houses in the surrounding area, number of cars, and presence of locked gate among other features of households. For instance, Durrant et al. (2015) used the Understanding Society Survey conducted in the UK to investigate whether using previous call information in RP models improved their predictive power. They found that the predictive power of RP models increased by 18 percentage points from 8% to 26%. Sinibaldi & Eckman (2015) used an experimental telephone survey conducted in Germany to investigate whether the interviewer ratings, call record data and interviewer characteristics improved the fit and discriminative power of RP models. They found that predictive power improved by 4 percentage points from 6% to 10% in terms of pseudo R^2.

Durrant et al. (2017) used longitudinal data from the Understanding Society and found that conditioning on previous wave paradata including call records, interviewer observation, and indicators of change improved the pseudo R^2 from l2% to 36%. However, they also noted that significant improvements in the predictive power were observed when conditioned on the most recent call record information. In summary, substantial progress has been achieved in improving the predictive power of RP models using paradata. However, the predictive power of RP models is still generally weak, limiting their utility in improving survey data quality. In paper 1, an assessment of whether a Bayesian framework for fitting RP models improves their predictive power by incorporating existing information as informative priors is conducted (Fearn, Gelman, Carlin, Stern, & Rubin, 1996). In the following section, further detail on Bayesian modelling is provided.

1.3.2 Bayesian Estimation

The Bayesian framework offers an attractive way of modelling response data due to its ability to allow incorporation of prior information on quantities of interest, with flexibility in modelling of complex data structures, exact inferences rather than asymptotic inference (e.g. asymptotic p-values calculated using an approximation to the true distribution especially in large sample sizes (Grendár, 2012)), and with more accurate estimates of parameter uncertainty (Fearn et al., 1996; Gill, 2014). In the Bayesian approach, model parameters are treated as random quantities while observed data are assumed to be fixed quantities (Fearn et al., 1996; Gill, 2014; Kruschke, 2011; Zyphur & Oswald, 2015). The Bayesian approach assigns a probability distribution of the possible values to the uncertainty attributed to a model parameter.

Bayesian analysis is based on Bayes' theorem (Bayes & Price, 1763), and incorporates existing knowledge and the joint distribution of observed data via mathematical relationships which are based on conditional probabilities (Fearn et al., 1996). Let $\Theta = (\theta_1, \theta_2, \dots, \theta_n)^T$ denote the vector of all the unknown parameters of the model, and $Y = (y_1, \dots, y_n)^T$ denote the vector of observed data. Applying Bayes' theorem, the posterior probability distribution of Θ for the observed data Y is:

$$\pi(\Theta|Y) = \frac{\pi(Y|\Theta)\pi(\theta)}{\pi(Y)} \qquad (1.8)$$

where $\pi(Y|\Theta)$ is the likelihood function which specifies the distribution of data Y given the parameters Θ , $\pi(\theta)$ is the prior probability distribution which represents all the relevant information available before observing the current data Y (prior belief on Θ), and $\pi(Y)$ is the marginal probability of the data. The $\pi(Y)$ is a normalising constant (i.e. a constant that makes the posterior density integrate to one) and does not depend on the model parameters about which inference is made. Ignoring the constant $\pi(Y)$, Bayes' theorem can be defined as:

$$\pi(\Theta|Y) \propto \pi(Y|\Theta)\pi(\theta) \qquad (1.9)$$

The posterior distribution can therefore be defined as a probabilistic combination of the information contained in the data (likelihood) and the prior distribution (Gill, 2014). The posterior distribution can be used for future inferences and decisions involving θ .This condition makes the Bayesian inference intuitively appealing for statistical inference.

The prior distribution is an intrinsic part of the Bayesian approach and relates to any information already known about the parameters of interest (Gelman, 2002; Gill, 2014). The priors can be broadly classified either as vague or informative. Vague priors, also referred to as reference, diffuse, flat and uninformative, tend to have a minimal influence on the posterior distribution of parameters Θ (Ghosh, 2011; Gill, 2014; Zhu & Lu, 2004). Vague priors for fixed effects are mostly assumed to follow a normal distribution with mean zero and a large variance (i.e. $\beta \sim N(0, \sigma^2)$ [2] where $\sigma^2 = 1000000$); while for the variance components they are assumed to follow an inverse gamma distribution and are defined as

[2] $\sim N(0, \sigma^2)$: β and σ^2denote regression coefficient and variance of the fixed effect respectively.

IG(α,1|β) [3] as α→0 and 1/(β) →0 which ensures that it is uninformative (Fong, Rue, & Wakefield, 2010; Gelman, 2006). The definition of inverse gamma parameters: α→0 and 1/(β) →0 ensures that it is uninformative. The posterior estimates obtained using vague priors are approximately equal to those estimated using a frequentist approach (Fearn et al., 2004; Gill, 2014).

Informative priors are priors that incorporate existing knowledge about the parameters of interest (Gill, 2014; Gill & Walker, 2005). Informative priors usually have an impact on the posterior distribution of Θ and may be derived from existing data, expert opinion, pilot studies, and scientific literature (Gill, 2014; Gill & Walker, 2005; Simpson, 1998; Winkler, 1967). For example, informative priors for the model coefficients do not have a mean of zero (i.e. $\beta \sim N(0, \sigma^2)$) but a value obtained either from previous research or theories. This is because a mean of zero in $\beta \sim N(0, \sigma^2)$ assumes that the coefficient for the parameter β is completely unknown. In addition, the corresponding variance component σ^2 for an informative prior does not necessarily take large values because known parameters are expected to be within a narrow range of a bounded integrals (Gill, 2014).

Informative priors for fixed effects, based on the previous or historical data can be formulated using various estimation methods such as: methods of moments, maximum likelihood estimation, maximum entropy estimation, and sequential Bayesian updating (i.e. uninformative pre-prior) (Guikema, 2007). The moments' method involves matching the measures of central tendency and spread to the appropriate moments of the distribution being fitted to the data. This approach is easy to implement and provides consistent estimates although it tends to produce estimates with the highest error covariance of all unbiased estimators compared to other estimation methods (Guikema & Pate-Cornell, 2004).

Maximum likelihood estimation uses maximisation algorithms to estimate coefficient parameters based on the likelihood of the data. The obtained coefficients are then applied as informative priors for the subsequent analysis (Guikema, 2007). The paramater estimates obtained using the maximum likelihood approach tend to be consistent and efficient especially for large datasets (Gill, 2014; Guikema, 2007). The maximum likelihood approach faces the drawback of being computationally intensive for models with complex structure (Guikema, 2007). In addition, this approach does not satisfy the efficiency condition when used for data with a small sample size.

The maximum entropy approach estimates the informative prior by maximising information contained in measures such as mean, mode, and median for a given probability density

[3] IG(α,1/β): β and α denote scale and shape variances respectively

function, while taking into consideration data constraints (Guikema, 2007). This approach is theoretically appealing because it maximises the uncertainty in the prior distribution resulting in a prior that agrees with evidence from data, irrespective of the sample size. The credible interval fitting with bootstrapping approach uses the tails of the parameter distributions to estimate informative priors and is related to the moments approach that uses measures of central tendency. For example, it is possible to estimate a 95% confidence interval for the success rate of an outcome based on the previous data to the 95% credible interval of the distribution being fit. The credible interval fitting approach focuses on the tails of the distribution instead of central moments and assumes that the obtained bootstrap interval is representative of the interval in the previous data. This assumption makes it difficult to know whether the measures of central tendency obtained are near the measures of the past data.

Finally, sequential Bayesian updating (SBU) proposed by Lindley (1972) involves choosing a suitable vague prior known as a pre-prior and updating it with the likelihood of the previous data (Armstrong, 1977; Gill, 2014; Guikema, 2007). The resulting posterior estimates from the previous data are then used as informative priors for the subsequent analysis of the data. The sequential Bayesian updating assumes that the previous data contains relevant information that can update the pre-prior resulting in a posterior estimate that converges asymptotically to the estimate of the observed data as the sample size increases. This approach is applied in this thesis because it assumes a Bayesian approach in all levels of the analysis.

Bayesian inference is implemented using simulation-based inference through Markov Chain Monte Carlo (MCMC) (Brooks, Gelman, Jones, & Meng, 2011; Fearn et al., 1996; Jannink, 2003). The posterior estimates are obtained through simulations where the initial values (i.e. priors) are updated in each iteration using the data (i.e. likelihood) (Kruschke, 2011). The final posterior estimates are obtained when the distribution of the posterior samples generated by Markov Chain converges to a stationary distribution. The rate of convergence is one of the key analytical factors that is used to determine the efficiency of MCMC and varies considerably depending on the target distributions. Convergence rate is influenced by the choice of starting values (i.e. priors), data transformation, thinning (retaining of the k^{th} value in the chain), blocking of parameters, and over-parameterisation of models (Brooks et al., 2011). The bias which may be introduced by the choice of starting values is reduced by discarding a defined number of first iterations within a burn -in period (Fearn et al., 1996). MCMC techniques are based on Gibbs Sampling, and Metropolis-Hastings sampling algorithms (Damlen, Wakefield, & Walker, 1999; Fearn et al., 1996; Jannink, 2003).

The Gibbs sampling technique generates posterior samples by sweeping through each parameter (or block of parameters) to sample from its conditional distribution with the remaining parameters fixed to their current values ,until convergence is achieved (Damlen et al., 1999; Lebanon, 2006). On the other hand, the Metropolis-Hastings algorithm starts with initial values for parameters of interest and generates new values from a proposal distribution that determines how to choose a new parameter value, given the current parameter value (Lebanon, 2006). For detailed derivations of the Gibbs and Metropolis-Hastings algorithms please see Lebanon (2006).

Bayesian inferences based on MCMC estimation are implemented in various statistical software packages such as MLwiN (Browne, Kelly, Charlton, & Pillinger, 2016) and WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), Stat-JR (Charlton et al., 2013). However, MCMC faces some issues when fitting models with large sample sizes and complex structures because it is computationally demanding (Rue, Martino, & Chopin, 2009; Taylor & Diggle, 2014). This makes it difficult to attain convergence because the sampled values do not end up having the same distribution as they would if they were sampled from the true posterior joint distribution. Currently, MCMC estimation has been extended to include algorithms that are computationally effective both in handling complex models and large datasets. One of the approaches involves using MCMC techniques that are based on Hamiltonian Monte Carlo (HMC) which is a more efficient and robust sampler than Gibbs sampling or Metropolis-Hastings for models with complex posteriors. This approach is implemented in the STAN package (Carpenter et al., 2016).

To counter computational drawbacks associated with MCMC, Rue et al. (2009) introduced Integrated Nested Laplace Approximations (INLA). The INLA approach is based on the multiple use of Laplace approximations combined with numerical integration, to obtain posterior estimates (Ferkingstad & Rue, 2015; Grilli, Metelli, & Rampichini, 2014; Rue et al., 2016). The INLA approach tends to be both faster and more accurate than MCMC alternatives (Blangiardo & Cameletti, 2015; Held, Schrodle, & Rue, 2010). However, INLA is restricted to the class of latent Gaussian models (LGMs) that represent a very useful generalisation of a large class of statistical models. Detailed formulation of the INLA approach can be found in Rue, Martino, & Chopin (2010).

The use of the Bayesian approach based on informative priors as a way of improving the predictions of RP models, has started attracting the attention of survey methodologists in recent years (Schouten et al., 2018; Wagner, 2016). For instance, Wagner (2016) used a Bayesian approach to predict survey response during data collection. Wagner specified the informative priors of fixed coefficients using the data collected in the last 21 days of the previous quarter of a survey. The results showed that using prior information in RP models improved classification power from a low of 40% to a high of 64%. Wagner (2016) also noted that prior information is more valuable in the early stages of data collection compared to the later stages.

Schouten et al. (2018) used the Bayesian approach to include and update prior knowledge about the survey design parameters, in the context of adaptive survey designs. They found that a correctly specified Bayesian analysis is robust compared to a non-Bayesian analysis when used for smaller sample sizes. This shows that the Bayesian approach may be used to learn and update strategies in adaptive and responsive surveys by using historical survey data. However, both Wagner (2016) and Schouten et al. (2018) noted that careful consideration of timeliness and the amount of previous data that is available is needed, when priors are based on previous survey data. Despite Wagner (2016) and Schouten et al. (2018) applying the Bayesian approach, a knowledge gap still exists in the use of informative priors derived from previous wave data in the context of longitudinal studies. This gap in knowledge will be addressed in paper 1 of this thesis.

1.3.3 Multilevel Modelling

In most instances survey data assumes a hierarchical or clustered structure. For instance, in face-to-face interviews, sample units have a natural hierarchy within the interviewers, and area primary sampling units. Usually, respondents interviewed by the same interviewer are more likely to have similar response patterns compared to those interviewed by different interviewers. The hierarchy is grouped at different levels where the lowest level (e.g. a sample units) may be defined as level-1, while a higher level such as interviewers may be defined as level-2. The units at a lower level are clustered or nested within groups of higher units. Survey data are mainly constructed of hierarchical structures and observations are therefore not independent. For this reason, it is crucial to account for hierarchical dependencies when modelling data from complex survey designs.

Multilevel models are an extension of the standard regression models (Goldstein, 2011). They account for correlations in the hierarchical data by including a residual error term for each level in the hierarchical structure (Goldstein, Browne, & Rasbash, 2002; Snijders & Bosker, 2012). This ensures that standard errors for the regression coefficients are not biased (Snijders & Bosker, 2012). Additionally, it becomes possible to explore complexities of variations in an outcome variable (Snijders & Bosker, 2012). The multilevel analysis was first implemented in education research where pupils (level-1) were clustered within schools (level-2), which themselves could be clustered within education authorities (level-3). Since then multilevel modelling has been extended other into disciplines including survey research (Durrant et al., 2010; Vassallo et al., 2015).

The standard response model for the household survey response can be defined as follows. Let y_i denote the binary response for household i $(i = 1, \dots, i)$ where

$$y_i = \begin{cases} 1 & \text{Response} \\ 0 & \text{Nonresponse} \end{cases} \qquad (1.10)$$

y_i is assumed to follow a Bernoulli distribution, with conditional response probabilities $\pi_i = Pr(y_i = 1)$ and $1 - \pi_i = Pr(y_i = 0)$. Then the standard logistic regression model takes the form

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \mathbf{x}_i'\boldsymbol{\beta} \qquad (1.11)$$

where, $\beta_0$is the intercept (i.e. represents the reference group which constitutes those households in the reference level), $\mathbf{x}_i'$ is a vector of household-level characteristics with coefficient vector $\boldsymbol{\beta}$. Now let's assume that y_{ij} denote the binary response for household i $(i = 1, \dots, i)$, interviewed by interviewer j $(j = 1, \dots, j)$ where

$$y_{ij} = \begin{cases} 1 & \text{Response} \\ 0 & \text{Non Response} \end{cases} \qquad (1.12)$$

y_{ij} is assumed to follow a Bernoulli distribution, with conditional response probabilities $\pi_{ij} = Pr(y_{ij} = 1)$ and $1 - \pi_{ij} = Pr(y_{ij} = 0)$. The multilevel logistic regression model accounting for interviewer effects takes the form

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_j'\boldsymbol{\alpha} + \mu_{oj} \qquad (1.13)$$

where β_0 is the intercept, $\mathbf{x}_{ij}'$ is a vector of household-level characteristics with coefficient vector $\boldsymbol{\beta}$, $\mathbf{z}_j'$ is a vector of interviewer-level covariates with coefficient vector $\boldsymbol{\alpha}$ and μ_{0j} is a random intercept. The random intercept is assumed to follow a normal distribution with zero mean and constant variance: $\mu_{oj} \sim N(0,\ \sigma_{\mu 0}^2)$. The required binomial variance in Equation (1.13) is obtained by constraining the level 1 variance (i.e. for households) to be one (Goldstein, 2010, pp. 113). Equation (1.13) represents the random intercepts model which can be extended to include random slope which allows the explanatory variable (i.e. incentive) to have a different effect for each group (i.e. interviewer). The random intercept and slope extended from Equation (1.13) takes the form

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + (\beta_1 + \mu_{1j})x_{1ij} + \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_j'\boldsymbol{\alpha} + \mu_{oj} \qquad (1.14)$$

$$= \beta_0 + \ \beta_{1j}x_{1ij} + \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_j'\boldsymbol{\alpha} + \mu_{1j}x_{1ij} + \mu_{oj} \qquad (1.15)$$

$$\beta_{0ij} = \beta_0 + \mu_{0j}; \beta_{1j} = \beta_1 + \mu_{1j} \qquad (1.16)$$

where β_0 is the intercept, $\beta_1$is the coefficient for x_{1ij} which is a dummy indicator of the household level variable (i.e. incentive) for household *i* within the assignment of interviewer *j*, $\mathbf{x}'_{ij}$ is a vector of household-level characteristics with coefficient vector $\boldsymbol{\beta}$, $\mathbf{z}'_j$ is a vector of interviewer-level covariates with coefficient vector $\boldsymbol{\alpha}$, μ_{0j} is a random intercept and μ_{1j} is a random coefficient for incentive variable. The random intercept and slope, μ_{0j} and μ_{1j}, are assumed to follow a normal distribution with zero mean and variance matrix Ω_μ defined as

$$\begin{bmatrix} \mu_{0j} \\ \mu_{1j} \end{bmatrix} \sim N\left(0, \Omega_\mu\right) \text{ where } \Omega_\mu = \begin{bmatrix} \sigma^2_{\mu 0} & \\ \sigma_{\mu 01} & \sigma^2_{\mu 1} \end{bmatrix} \qquad (1.17)$$

where $\sigma^2_{\mu 0}$ is the intercept variance, $\sigma^2_{\mu 1}$ is the variance in slope and $\sigma_{\mu 01}$ is the covariance between intercepts and slope.

Equation (1.15) expresses the log-odds (i.e. logit of π_{ij}), as a sum of a linear function of explanatory variables and a random group-dependent deviation μ_{0j} and μ_{1j}. The overall intercept in the linear relationship between the log-odds of y_{ij} and the explanatory variables included in the equation is represented by β_{0j}. The explanatory variables may also include interactions between $\mathbf{x}'_{ij}$ and $\mathbf{z}'_j$ variables to determine whether the nature of a lower-level relationship (i.e. household) depends on a higher-level factor (i.e. interviewer level factor). This relationship between low- and high-level variables is referred to as a cross-level interaction effects. The variance components σ^2_{u0} and σ^2_{u1} can be used for the computation of the Intraclass correlation coefficient (ICC) in sample survey (Snijders & Bosker, 2012). The ICC represents the degree of resemblance between variables measured for two randomly drawn individuals in one random group.

The main advantage of accounting for hierarchical structures in survey data is that regression coefficients and standard errors obtained are correctly estimated (Goldstein, 2011; Snijders & Bosker, 2012). Hierarchical structures also make it possible to split residual variation into different components. This enables exploration of the extent to which variability of an outcome variable can be explained by the characteristics associated with different levels (Snijders & Bosker, 2012). For example, the multilevel model framework helps to study the extent to which interviewers' characteristics may influence survey response among sample units. The statistical inference about variations among sample units (i.e. lower level) on the outcome variable is obtained because they are regarded as a random sample from a population of higher level units (Snijders & Bosker, 2012). This enables the derivation of information about relationships at different groups or levels.

Multilevel models are defined by both regression and variance components and are estimated using maximum likelihood (ML) and restricted maximum likelihood (REML) (both frequentist approaches) and Bayesian approaches (Gill 2014; Goldstein 2011; Havard Rue, Martino, and Chopin 2009; Simon 2009). The frequentist approaches (ML and REML) differ little with respect to estimating the regression coefficients, but they do differ with respect to how variances are estimated (Snijders & Bosker, 2012). The REML method considers the loss of degrees of freedom resulting from the estimation of the regression parameters when estimating the variance components, while ML does not take this into account. This makes variance estimates for ML to have a downward bias, a limitation not faced by REML estimates. Although both ML and REML are widely used for estimating parameters in multilevel models, they lack flexibility and tend to underestimate the variance components especially where the number of clusters is small (Joe, 2008). On the other hand, Bayesian approaches are naturally suited to estimate multilevel models because they can robustly account for any uncertainties associated with statistical parameters by the assumed probability distribution (Browne, Draper, & David, 2006).

The computation of ICC in multilevel response propensity models is not straight forward because variance components for household and interviewer are not directly comparable. This is because in logistic regression the random error ϵ_i is assumed to have a logistic cumulative density function given explanatory variables (i.e. in probability scale). In addition, random error is dependent on the expected value of $var(y_{ij}) = \pi_{ij}(1 - \pi_{ij})$. On the other hand, the variance components for random intercept and slope are measured on logistic scales. Therefore, approaches such as linearisation, simulation, binary linear model and a latent variable are used for computing the ICC for binary outcome models (Goldstein, 2010, pp.123-131). For example, the latent variable approach calculates the ICC by assuming that household variance (i.e. random error) is fixed at $\pi^2/3 = 3.29$ (i.e. variance for the standard logistic distribution) and both household and interviewer variances can be expressed on a continuous scale. The ICC is therefore calculated as the ratio of the interviewer variance to the sum of household and interviewer variances. However, it is important to note that latent variable approach is not justifiable for calculating the ICC when the response outcome is truly discrete (i.e. a response that is not derived from the truncation of a continuous variable). Goldstein et al. (2002) propose that linearisation or simulation approaches should be used when the response outcome is discrete.

The multilevel model in Equation (1.13) can be extended to represent a random structure when clustering or nesting is not perfect. For example, an interviewer may be assigned to different households across different primary sampling units (PSU) while some households may be assigned to more than 2 interviewers after re-issues. This may introduce complex hierarchical structures which are handled by a specific class of multilevel models, known as cross-classified models (Goldstein, 2010, pp. 243-254; Rasbash & Goldstein, 1994; Snijders &

Bosker, 2012). The cross-classified models handle data in which a lower level unit (e.g. a household) belongs uniquely to more than two higher levels (e.g. interviewers and areas). The cross-classified model assumes that units (i.e. households) can only be members of one higher level unit (i.e. interviewers) and it is not expected that a given household will be interviewed by more than one interviewer. In this case it becomes crucial to account for cross-classification at level-2 between interviewers and areas ,to produce unbiased variance estimates (Dunn, Richmond, Milliren, & Subramanian, 2015; Meyers & Beretvas, 2006). Due to the complexity of the survey data it is crucial to consider the appropriateness of the multilevel model to avoid misspecification effects of the variance estimates. If a given household is interviewed by more than one interviewer then it assumes a multiple membership structure which is analysed using multiple membership models (Goldstein, 2010, pp. 255).

Multilevel models have been widely used to investigate interviewer effects on survey cooperation and response (Durrant, D'Arrigo, & Steele, 2013; Durrant et al., 2010; Durrant & Steele, 2007, 2009; O'Muircheartaigh & Campanelli, 1998; Vassallo et al., 2016, 2015). For example, O'Muircheartaigh & Campanelli (1998) used multilevel cross classified models for the British Household Panel Study (BHPS) to investigate the relative impact of interviewer effects and sample design effects on survey precision. They concluded that the multilevel framework is naturally designed to analyse survey data that have different levels. Durrant, Groves, & Steele (2010) used a multilevel cross-classified logistic model with random interviewer effects to account for the clustering of households within interviewers, and for the classifications of interviewers within households. They found that matching interviewer characteristics to different subgroups of the population such as age and ethnicity improved cooperation rates. Vassallo, Durrant, & Smith (2015) used data from the UK Family and Children Survey and found that cross-classified multilevel models provide a flexible class of models for the analysis of interviewer effects on survey response. In summary, multilevel models provide a flexible approach for modelling interviewer effects and cross-level interactions in survey data.

1.3.4 Propensity Score Analysis

Propensity score analysis is a statistical approach used to make different groups compositionally equivalent (Lee, 2006; Lugtig, Lensvelt-Mulders, Frerichs, & Greven, 2011; Rosenbaum & Rubin, 1983; Särndal & Lundström, 2005). It was introduced by Rosenbaum & Rubin (1983) to serve as a dimension reduction tool by condensing treatment assignment information into a single score (Rosenbaum & Rubin, 1983). The propensity score is the probability of treatment assignment conditional on observed baseline characteristics. This approach is applied in observational studies where randomisation is not possible or ethical. Nonrandomisation in observational studies makes participants for the treated and control groups probabilistically unequal thereby providing less compelling support for counterfactual inferences because they are susceptible to selection bias (Shadish, Cook, & Campbell, 2002).

Selection bias arises from differential coverage and nonresponse across treatment and control groups (Starks, Diehr, & Curtis, 2009; Voogt & Saris, 2005; Weisberg, 2005). When sample characteristics that influence selection into either treatment or control group are related to an outcome of interest, confounding is introduced. This means that an estimate of the association between an exposure and the outcome of interest is distorted by selection bias. Therefore, ignoring confounding in the outcome analysis may lead to estimation of treatment effects that differ from the true values as a result of being falsely attributed to the intervention (Starks et al., 2009).

Propensity score models are intended to correct for the imbalance of different groups such that they mimic the characteristics of randomised studies, in which treated and control groups are probabilistically comparable (Agostino, 1998; Austin, 2011a). The key assumption of propensity scores is the 'ignorability': given a set of observed covariates X, treatment assignment is independent of the potential outcomes. This is defined as

$$(Y^{z=1}, Y^{z=0}) \perp \mathrm{Z}|\mathrm{X} \qquad (1.18)$$

where $Y^{z=1}$ and $Y^{z=0}$ are potential outcomes observed for the treatment and control groups, respectively. This means that conditional on covariates X, the assignment of units to binary treatment conditions (i.e. treatment and control) is independent of the outcome of control ($Y^{z=0}$) and of treatment ($Y^{z=1}$). Therefore, conditional on the propensity score, the distribution of observed baseline characteristics will be similar between the treatment and control groups (Rosenbaum & Rubin, 1983). This property makes it possible for each sample unit to have the same probability of assignment to each group (i.e. treatment and control) as in a randomised experiment (Rosenbaum & Rubin, 1983). The propensity score, as the predicted probability of being in a given group can be estimated using logistic regression, the probit model, and discriminant analysis (Guo & Fraser, 2014). Propensity scores can be implemented in a number of ways including matching, stratification, inverse probability weighting (IPW), and covariate adjustment (Heinze & Jüni, 2011; Lunceford & Davidian, 2004; Rosenbaum & Rubin, 1983; Rubin & Rosenbaum, 1984; Williamson, Morley, Lucas, & Carpenter, 2012).

Propensity score matching (PSM) entails matching treated to control individuals based on their respective estimated propensity scores (Rosenbaum & Rubin, 1985). The stratification approach consists of using the propensity score distribution to divide the sample of treatment and control units into strata that are similar with respect to the distribution of covariates (Agostino, 1998; Rubin & Rosenbaum, 1984). Inverse probability weighting (IPW) uses the inverse of the propensity score as a weight to create a synthetic sample which the distribution of baseline covariates is assumed to be independent of treatment assignment (Lunceford & Davidian, 2004; Rosenbaum, 1987). Lastly, the covariate adjustment approach includes the propensity score as an additional covariate in the outcome regression model (Elze et al., 2017; Kazmi, Obrador, Khan, Pereira, & Kausz, 2004). Usually, the outcome variable is regressed on an estimated propensity score and the indicator variable of the treatment status.

Currently, there is a wealth of literature about how effectively the four different approaches account for selection bias. In this thesis, the review will be limited to focus only on PSM which is applied in paper 3 to adjust for selection effects in the evaluation of measurement effects in online probability and face-to-face surveys. The use of PSM is motivated by the fact that it results in well-balanced groups of comparison (Dehejia & Wahba, 2002; Ertefaie & Stephens, 2010; Hirano, Imbens, & Geert, 2003). This is because after matching, all the unmatched units are discarded and are not directly used in estimating mode effects (Dehejia & Wahba, 2002). This results in estimators with lower Mean-Squared Error (MSE) compared to those obtained using Inverse Probability Weighting (IPW) which is sensitive to extreme observations (Ertefaie & Stephens, 2010; Hahn, 1998; Hirano et al., 2003). Before discussing the PSM approach, it is necessary to understand the potential outcome framework for estimating treatment effects.

Under the potential outcome framework, there can be only two possible treatments on an individual, a sample or population (Rubin, 1978). The key assumption is that individuals selected into treatment and control groups have potential outcomes in two states. These states are the one in which they are observed and the one in which they are not. For example, in the context of data collected using two modes, it is assumed that sample members are either assigned to online or face-to-face interviews. Then, given individuals and treatments (i.e. face-to-face and online modes), each individual may be thought of as having a pair of potential outcomes: $Y_1(0)$ and $Y_1(1)$, the outcomes obtained using either face-to-face or online modes, respectively. In practice, it is only possible to assign each individual to either the face-to-face or the online mode, not both. Therefore, in this context, the potential outcome framework aims to compare what the outcome would be if each individual was assigned to both face-to-face and online modes.

This may be explained as follows: Let Z be an indicator variable denoting the treatment group assigned to an individual, such that that $Z = 0$ and $Z = 1$ indicates being in control and treatment, respectively. Let $Y^{Z=1}$ be the outcome value that would have been observed under treatment group when $Z = 1$, and $Y^{Z=0}$ be the actual outcome value observed under control group. Then the value $Y^{Z=1}$ represents an individual's potential outcome that would have been observed if an individual was potentially assigned in treatment group which he/she was not actually assigned to. Since these potential outcomes would have been observed in situations that did not actually happen they are also known as counterfactual outcomes (i.e. in counter to the fact situations)(Hernan, 2004; Leite, 2017).

Therefore, based on the potential outcome framework the treatment effect for each individual i arises when $Y_i^{z=1} \neq Y_i^{z=0}$ (Hernan, 2004). The treatment effects attributable to a given individual i may be computed as a difference between the potential outcomes such as $Y_i^{z=1} - Y_i^{z=0}$ or as proportions $Y_i^{z=1}/Y_i^{z=0}$. However, this is not possible because only one outcome is observed for each individual i in a given time in what is known as the fundamental problem of the causal inference (Holland, 1986). To overcome this issue the focus of estimating treatment differences moves from each individual to all individuals i $(i = 1, \dots, i)$ in the sample where measures of central tendency are used for evaluation of mode effects. For example, the average mode effect (τ) can be defined as

$$\tau = E[Y^{z=1}] - E[Y^{z=0}] \text{ if } E[Y^{z=1}] \neq E[Y^{z=0}] \qquad (1.19)$$

where $E[Y^{z=1}] = \frac{1}{N}\sum_{i=1}^{N} Y_i^{z=1}$ and $E[Y^{z=0}] = \frac{1}{N}\sum_{i=1}^{N} Y_i^{z=0}$ are the average of the potential outcomes of individuals in the sample. The condition $E[Y^{z=1}] \neq E[Y^{z=0}]$ indicates that exposure has a causal effect (Hernan, 2004).

Propensity Score Matching

PSM is a 5 step analytic procedure (Agostino, 1998). The first step involves the estimation of the propensity scores. The second step comprises choosing a matching algorithm followed by the assessment of common support. The fourth step involves the diagnosis of matches and the final step is the estimation of treatment effects.

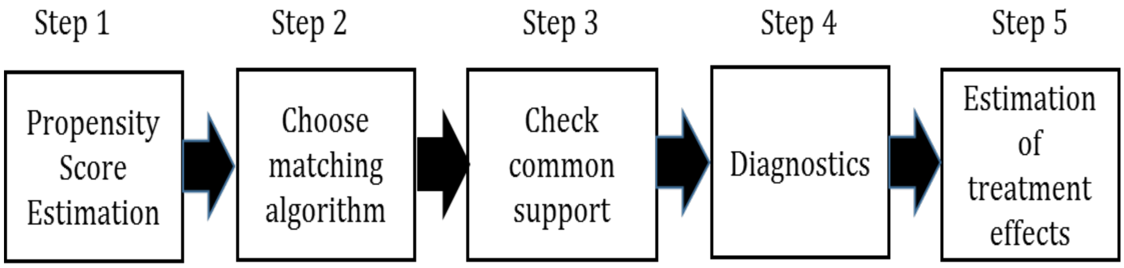

Step 1
Step 2
Step 3
Step 4
Step 5
Propensity Score Estimation
Choose matching algorithm
Check common support
Diagnostics
Estimation of treatment effects

Figure 1-3: PSM implementation steps, adapted from Caliendo & Kopeinig (2008)

Step 1: Estimating propensity scores using logistic regression

The propensity score for respondent i is defined as a conditional probability of treatment assignment $(Z_i = 1)$ versus control $(Z_i = 0)$, given a vector of observed baseline covariates x_i (Rosenbaum & Rubin, 1983). The propensity score takes the form

$$e(x_i) = P(Z_i = 1|X_i = x_i) \tag{1.20}$$

$$= \frac{exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}{1 + exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)} \tag{1.21}$$

where Z represents an indicator variable for treatment and X is a vector of the observed baseline covariates $x_1, x_{2,} \dots., x_n$; and $\beta_0, \beta_{2,} \dots., \beta_n$ are the corresponding regression coefficients. Equation (1.20) assumes that, given $X's$, the Z_i's are independent:

$$pr(Z_1 = z_1, \dots, Z_n = z_n|X_1 = x, \dots, X_n = x_n) = \prod_{i=1}^{N} e(x_i)^{z_i}\{1 - e(x_i)\}^{1-z_i} \tag{1.22}$$

This implies that propensity score reduces all information contained in covariates of a given unit into a single value that lies between 0 and 1.

The choice of the baseline covariates included in the propensity score model is of great importance because the final estimation about treatment effect for the treated is clearly sensitive to this specification (Austin, Grootendorst, and Anderson 2007; Brookhart et al. 2007; Brooks et al. 2011; Smith and Todd 2005). The choice of baseline covariates affect bias, variance, and mean-squared error of the estimated treatment effects. Brookhart et al. (2007) recommend that all covariates that have a direct effect on the probability of treatment assignment and are related to the outcome should be included in the propensity score model. Such covariates are known as true confounders and their inclusion in the propensity score model leads to a reduction in bias and variance of the treatment effect estimates. Brookhart et al. (2007) and Cuong (2013) also recommend the inclusion of variables that are not related to the treatment assignment but are related to the outcome of interest. These variables are known as potential outcomes and their inclusion in the propensity score model leads to a reduction in the variance of treatment effect estimates without increasing bias. Finally, Brookhart et al. (2007) showed that the inclusion of covariates related to treatment (i.e. treatment predictors) in the propensity score model and not the outcome, may increase variance of treatment effect estimates with no reduction in bias. They suggest that those variables that may explain the relationship between treatment and outcome (i.e. mediators) should not to be included in propensity score models because they tend to remove some of the treatment effects.

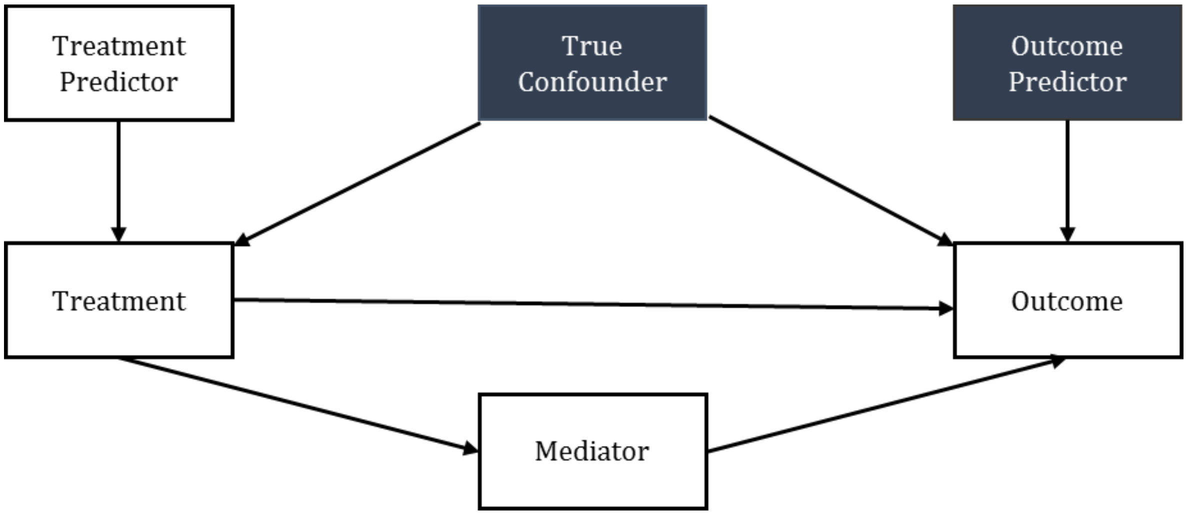


Figure 1-4: Relationship between covariates, treatment assignment, and outcome, with black boxes indicating covariates that should be included in propensity score model (Adapted from Leite (2017)).

Several selection strategies for the inclusion or exclusion of the covariates in the final propensity score model have been proposed (Agostino, 1998; Brookhart et al., 2007; Hirano & Imbens, 2001). Hirano and Imbens (2001) propose that the choice of variables to be included in the final propensity model should be selected based on the significance of the univariate relationships between the covariates and treatment assignment. That is, only significant covariates in the univariate propensity model should be included in the final propensity score model. On the other hand, Agostino (1998) and Brookhart et al. (2007) recommend that the propensity scores model should contain as many variables as possible even if they are not statistically significant. This is because the propensity score model aims to match treatment and control units, while controlling for as much confounding as possible.

Step2: Choose matching algorithm

The common matching algorithms include: greedy matching, optimal matching and genetic matching (Guo & Fraser, 2014; Leite, 2017; Rosenbaum, 2002). Optimal matching is a process of developing matched sets in such a way that the total sample distance of propensity scores is minimised (Rosenbaum, 2002). Genetic matching is a method for multivariate matching that minimises a weighted distance between treated and control groups (Diamond & Sekhon, 2013). Greedy matching consists of choosing each of the treated units and searching for the best available match among the control units (Rosenbaum, 2002). There is a large literature about each of these matching algorithms. In this thesis, the review will be limited to focus on greedy matching which is applied on third paper. Greedy matching was selected because it has shown to have superior performance compared to other matching algorithms in terms of reduced bias for matched samples (Austin, 2009b, 2012).

Greedy matching algorithms involve dividing a large decision problem into a series of smaller and simpler decisions without taking into account earlier decisions when making later decisions (Rosenbaum, 2002). Greedy matching, especially for units matched using the nearest neighbour algorithm, allows the evaluation of causal effects in a similar way to that in randomised experiments. However, greedy matching tends to be sensitive when the distribution of the propensity scores for the units in treatment and control group are not similar (i.e. common support not adequate) (Guo & Fraser, 2014). In PSM, common support implies that for each value of covariates X , there is a positive probability of being in both treatment and control groups (Austin, 2011a, 2011b). In practice, greedy matching requires a large sample size for the matching to be completely effective, because units without matching pairs are discarded after matching (Austin, 2011b; Guo & Fraser, 2014). This may result in an increase in the variance of treatment effects. Greedy matching may be implemented by choosing any of the 3 methods: one-to-one or fixed ratio matching, nearest neighbour matching and within caliper matching (Austin, 2011a; Guo & Fraser, 2014; Leite, 2017; Rosenbaum, 2002).

One-to-one matching is the most common approach for propensity score matching (Austin, 2011a). In this approach, pairs of treated and control units with similar propensity scores are matched. The resulting matched sample is usually homogeneous , leading to a reduction in bias of estimated treatment effects (Cohen, 1988). However, one-to-one matching may result in treatment estimates with higher variance if common support between treatment and control units is not adequate, leading to a matched sample with few units (Bryson, Dorsett, & Purdon, 2002; Leite, 2017). The fixed ratio matching involves matching a single treatment unit to a given number of control units depending on a specified ratio (Austin, 2011a; Leite, 2017). In fixed ratio matching, matching will occur even if there is no adequate common support, leading to an increase in bias of the treatment effects.

Nearest neighbour (NN) matching involves finding the control unit with the closest propensity score to that of treated unit (Rosenbaum & Rubin, 1985; Stuart, 2010). NN matching has two main variants: NN matching "with replacement" and "without replacement". In the NN matching with replacement a control unit can be used more than once as a match while for NN matching without replacement, a control unit can only be matched to one treated unit. NN matching with replacement is preferred in samples where the distribution of propensity scores in the treatment and control groups are not similar, and the number of potential matches between the two group is small (Rosenbaum, 1989). This is because NN matching with replacement increases the average quality of matching and reduces the overall bias. However, the difference between these two NN matching approaches disappears when the number of available matches between control and treatment groups is large.

NN matching can lead to bad matches if the closest neighbour is far way. This drawback of NN matching is avoided by imposing a common support condition known as a calliper between propensity scores for control and treatment units. This ensures that pairs of treated and control units in matched sample are within the specified calliper distance (Austin, 2008a, 2008b). The choice of the width of the calliper is crucial because it reflects an implicit trade-off between the variance and the bias of the estimated treatment effects (Smith & Todd, 2005). The literature proposes calliper widths that range between 0.1 and 0.25 standard deviations of the logit of the propensity score (Austin, 2011a, 2011b). Specification of narrower callipers may result in matching of more similar units leading to a reduction in bias by reducing systematic differences between the treated and control units (Caliendo & Kopeinig, 2008). However, this also leads to a higher number of unmatched units which may result in an increase in the variance of the estimated treatment effects. Specification of wide callipers have the opposite effect. Austin (2011b) recommends using callipers of width equal to 0.2 of the standard deviations of the logit of the propensity score, because callipers of this size tend to have optimal performance for estimating treatment effects.

Nearest neighbour and calliper matching can be combined into one method known as the nearest neighbour with calliper approach (Stuart, 2010). This approach begins with randomly ordering control and treated units, then selecting the first treated unit and finding the control unit with the closest propensity score within a specified calliper width of the propensity scores. Then both units are removed from the matching sample and the next treated unit is selected. PSM using nearest neighbour matching with calliper only uses units which are close to the area of common support leading to a reduction in the overall bias. The units that are out of the range of the area of the common support in terms of propensity scores are discarded and are not used for estimation of treatment effects. When the number of discarded units is large it may result in an increase in the variance of the estimated treatment effects (Bryson et al., 2002).

Step 3: Common Support

The effectiveness of the propensity score as a balancing score is determined by evaluating whether it has been adequately specified. This is done by checking the area of common support (Austin, 2011a; Leite, 2017). The assessment of the area of common support involves determining whether or not the distribution of measured baseline covariates between treated and control groups is similar using histograms and boxplots. This is determined by checking the overlap of propensity scores between treated and control units. It is expected that after conditioning on the propensity scores, all systematic differences between treated and control group will have been removed (Austin, 2011a). Therefore, an adequate common support indicates that an appropriate number of matches for treated and control units will be attained for effective estimation of treatment effects (Austin, 2011a; Hansen, 2008). Inadequacy of the area of common support may result when there is a covariate imbalance between treatment and control groups. This happens when many of the control units are different from most of the treatment units making them inappropriate for estimating treatment effects. PSMs are preferred for estimating treatment effects in observational studies because it is possible to assess the area of common support of the resulting matched sample (Austin, 2011a).

Step 4: Diagnosing matches

It is important to assess the quality of the matched samples. This is done by an assessment of covariate balance between the matched groups. Covariate balance is defined as the similarity of the empirical distributions of the full set of covariates in the matched treated and control groups. Covariate balance is evaluated using graphical, descriptive and inferential procedures (Austin, 2009a; Leite, 2017; Linden, 2015; Stuart, 2010). The main graphical displays to visualise and compare covariate balance is the quantile-quantile plot (QQ plot) for continuous variables and histograms for categorical covariates (Linden, 2015; Stuart, 2010). The QQ plot involves plotting the quantiles of the covariate for the treatment group against those of the control group. The Q-Q plot shows how and where the points deviate from the diagonal line, which represents the perfect correlation between the two distributions. Points that lie far away from the diagonal line show that the covariates are not balanced. For histograms, categories of each covariate, for treated and control groups, can be overlapped, and any nonoverlapping areas indicate a lack of covariate balance.

For descriptive, standardised mean difference (SMD) has been used to quantify differences in means and proportions between the two groups (Austin, 2009a; Leite, 2017). The SMD compares the difference in means in units of the pooled standard deviation. The SMD is not affected by the sample size which makes it an ideal measure to compare balance in measured variables between treated and control units, before and after matching. For continuous variables, the standardised difference is defined by Austin (2009a) as

$$d = \frac{(\bar{x}_{treatment} - \bar{x}_{control})}{\sqrt{\frac{s^2_{treatment+}s^2_{control}}{2}}} \tag{1.23}$$

where $\bar{x}_{treatment}$ and $\bar{x}_{control}$ denote the sample mean of the covariate in treated and control subjects respectively. While $s^2_{treatment}$ and $s^2_{control}$ denote the sample variance of the covariate in treatment and control units respectively. The choice of the pooled standard deviation $\sqrt{\frac{s^2_{treatment+}s^2_{control}}{2}}$ is informed by the assumption that the number of repeated measurements made within each group (i.e. treatment and control) are the same (i.e. one). On the other hand, if the mixed-mode data collection was done with different samples, each measured repeatedly, then the standardised difference is defined as:

$$d = \frac{(\bar{x}_{treatment} - \bar{x}_{control})}{\sqrt{\frac{(n_1 - 1)s^2_{treatment+}(n_2 - 1)s^2_{control}}{n_1 + n_2 - 2}}} \tag{1.24}$$

Where $n_1$and n_2 are the number of measurements made for different samples.

For the dichotomous variables, standardised difference is defined as

$$d = \frac{(\hat{p}_{treatment} - \hat{p}_{control})}{\sqrt{\frac{\hat{p}_{treatment}(1 - \hat{p}_{treatment}) + \hat{p}_{control}(1 - \hat{p}_{control})}{2}}} \tag{1.25}$$

Where $\hat{p}_{treatment}$ and $\hat{p}_{control}$ denote the proportion or mean of the dichotomous variable in treated and control units. An adequate covariate balance is achieved if the absolute standardised mean difference is below 0.1 standard deviations (Austin 211). This is known as a strict criterion threshold. A less strict criterion of less than 0.25 standard deviations is proposed by Stuart (2007). According to Cohen (1988) standardised differences of 0.2, 0.5 and 0.8 represent small, moderate and large effect sizes respectively.

Inferential measures for covariate balance evaluation include t-tests comparing group means and Kolmogorov-Smirnov tests. Significant t-tests are expected before matching, while no significant differences are expected after matching if the covariate balance is adequate. However, inferential measures are not robust measures of evaluating covariate balance for small and large samples. This is because small samples tend to be underpowered and thus fail to indicate any substantial covariate unbalance even if covariate differences between groups are very large. On the other hand, high levels of power for large samples may make it hard to achieve covariate balance, even if the covariate differences between groups are very small. Additionally, covariate balance is a property of the sample (i.e. surveys), and hypothesis tests associated with inferential measures are known to refer to the general population and not a sample.

Step 5: Estimating treatment Effects

The new sample derived after matching is assumed to be comparable across treated and control groups. In this stage, analysis is performed on the new matched sample to compare outcomes between treated and control groups. The differences between the two groups is known as treatment effects. In addition, average treatment effect for either the treated or control groups can be computed. The matched sample can be used to obtain treatment effects from continuous, categorical, ordinal and censored outcomes.

One of the methodological issues that arises around the use of propensity scores in complex surveys is the use of survey weights. Several studies have explored whether survey weights should be included in propensity score models or not (Austin, Jembere, & Chiu, 2018; DuGoff, Schuler, & Stuart, 2014; Ridgeway, Kovalchik, Griffin, & Kabeto, 2015; Zanutto, 2006). For example, Zanutto (2006) used propensity score analyses to investigate the effect of gender on Information Technology (IT) jobs, using data from the 1997 U.S. Scientists and Engineers Statistical (SESTAT) database (NSF 99-337). Zanutto included survey weights in propensity score models with an aim of adjusting for differential selection probabilities, nonresponse and post-stratification adjustments. Zanutto (2006) recommended that survey weights should not be used in the propensity score model but can be used in the outcome analysis because it results in unbiased estimated treatment effects.

Dugoff et al. (2008) used Monte Carlo simulations and the 2008 Medical Expenditure Panel Survey data to illustrate the application of propensity score models in complex survey data. They compared the following methods for estimating the treatment effects: (1) naïve model-where they ignored both survey weights and propensity scores, (2) propensity scores model that ignored both survey weights and survey strata, and (3) propensity scores model that accounted for survey weights. Their findings were consistent with those of Zanutto (2006), that survey weights should not be incorporated in the estimation of the propensity score model, but survey weights can be used in the outcome analysis, especially if the goal is to make inferences on population. This is because including survey weights in the outcome analysis resulted to unbiased treatment effects that are generalisable to the original survey target population. Additionally, they suggested that survey weights should be included as covariates in the propensity score model.

Ridgeway et al. (2015) used simulated data and The 2009 Insights data from the Newest Members of America's Law Enforcement Community survey, to compare the performance of four different methods for estimating propensity scores in complex surveys. They considered the following formulations of propensity score models: (1) an unweighted model, (2) a weighted model, (3) an unweighted model with sampling weights as an additional covariate, and (4) a weighted model with sampling weights as an additional covariate. They found that propensity score models that incorporated survey weights in a weighted model resulted in a better covariate balance compared to those models that incorporated sampling weights as an additional covariate. In addition, Ridgeway et al. (2015) suggested that a product of sampling weights and propensity score weights should be used as the weights in the outcome analysis of treatment effect for matched samples. This is because treatment effects estimated in this way had the lowest root mean squared error (MSE). In summary, Ridgeway et al. (2015) concluded that survey weights should be incorporated during estimation of propensity scores in propensity score models and in the outcome analysis of treatment effects.

The recent work by Lenis et al. (2017) used simulated data and data from Early Childhood Longitudinal Study, Kindergarten class 1998-1999 (ECLS-K) to examine three different formulations of handling survey weights in propensity score models: (1) no survey weights, (2) survey weights in a weighted estimation, and (3) survey weights as a covariate in the estimation. Lenis et al. (2017) found that survey weights incorporated in the propensity score models do not influence the estimation of the population treatment estimates. In addition, Lenis et al. (2017) suggested that matched control units should use inherited weights of the treated units they are matched to as survey weights in the outcome analysis. They found using inherited weights to be beneficial in terms of reducing nonresponse bias under certain nonresponse mechanisms. Austin et al. (2018) used both simulated data and CCHS survey data to examine how survey weights in the propensity score models should be evaluated. They considered the same three different formulations used by Lenis et al. (2017) and their results were inconclusive with respect to which methods of estimating the propensity score model was preferable. Additionally, they recommended that matched control units should retain their survey weights because these weights lead to a decreased bias of the estimated treatment effect. Paper 3 will also evaluate whether survey weights influence the estimation of outcome for the matched sample.

1.4 Overview of the three papers

With a continuous rise in survey nonresponse and the use of different modes of data collection, there is an increase in the number of studies focusing on ways of understanding and improving survey data quality. This is important because it helps to make a continuous improvement of survey processes which ensures timely collection of high quality data within the budgeted costs. However, a clear understanding and subsequent improvement of survey data quality is a complex undertaking since survey errors are interrelated and vary across surveys. In addition, the survey landscape is transforming quickly. Therefore, it is important to understand the impact of various errors on survey quality as this will help in the establishment of effective and efficient strategies for survey data collection.

Respondents influence survey quality by either choosing to participate in a survey or not (i.e. nonresponse error) (Groves, 1989; Groves et al., 2009). Therefore, it becomes important to understand response behaviour of survey participants and to target potential nonrespondents to improve response rates. One way of addressing this issue is by improving response propensity (RP) models in terms of their predictive power. This will in turn improve the accuracy of such models in their applications during survey management such as in adaptive and responsive survey designs. However, existing literature shows that RP models tend to have a relatively low predictive power of less than 8% in terms of Pseudo R^2 (Fricker & Tourangeau, 2010; Kreuter & Olson, 2011; Olson & Groves, 2012; Olson et al., 2012; Plewis et al., 2012). A number of ways may be explored to improve the predictive power of RP models such as the Bayesian approach that takes account of prior information (Duan, 2005; Fearn et al., 1996; Schouten et al., 2018; Viele et al., 2014). This approach is explored in the first empirical paper "An assessment of the utility of a Bayesian framework to improve response propensity modelling". This paper explores whether or not the use of a Bayesian approach, based on informative priors derived from previous waves, in a longitudinal context improves the predictive power of RP models. Classification tables, discrimination (sensitivity and specificity), prediction (positive and negative predictive values), and the area under the curve (AUC) of the receiver operating curves (ROC) are used to evaluate the predictive power, based on out of sample predictions. The data used in this paper are from Understanding Society, the UK Household Longitudinal Study (UKHLS). The findings indicate only a slight improvement in model fit when previous wave information is incorporated in response propensity models as informative priors. In addition, measures of classification, prediction, and discrimination only showed minimal gains in predictive and discriminative power of survey response predictions. These results contribute to a better understanding of the use of previous wave data as informative priors especially in adaptive and responsive designs.

Interviewers have long been known to affect data quality (Groves, 1989). This is because interviewers have varying skills in contacting respondents and eliciting their participation. Existing research has shown that interviewer characteristics such as age, gender, interviewer experience, and education are good at explaining some of the interviewer effects (Blom & Korbmacher, 2013; Groves & Couper, 1998; Purdon, Campanelli, & Sturgis, 1999). It is not only interviewers who have played a key part for maximising survey cooperation in interviewer-mediated surveys, but monetary incentives of various kinds have also played their part (Singer, 2002; Singer, Hoewyk, et al., 1999). Considering that interviewers play a key role in contacting respondents, it is likely that they also influence the effectiveness of incentives when offering them to sample members.

While the existing literature on the effects of incentives on response rates is substantial, little is currently known about the role of interviewers in determining whether incentives are deployed effectively. In Paper 2, the question "Do interviewers moderate the effect of monetary incentives on response rates in household interview surveys?" is explored. The paper uses data from three different UK face-to-face interviewer surveys. These are the 2015

National Survey for Wales Field Test (NSW 2015), the 2016 National Survey for Wales Incentive Experiment (NSW 2016), and Wave 1 of the UK Household Longitudinal Study Innovation Panel (UKHLS-IP). To account for the hierarchical structure between households and the interviewers, a multilevel modelling approach is adopted. The multilevel model also includes a random slope on incentive, to capture the variability of interviewers in the deployment of incentives. The results show significant and substantial variability between interviewers in the effectiveness of monetary incentives on the cooperation rates across all three surveys. However, none of the interviewer characteristics considered are significantly associated with more or less successful interviewers. These results are useful in identifying interviewers' performance in the deployment of incentives which may help in recruiting and training of interviewers especially on approaches of recognising and heightening the saliency of incentives in surveys.

Face-to-face interviewing has long been held as the "gold standard" mode of data collection that leads to the best data quality, in comparison to other modes. This is owing to its higher response rates compared to other modes, as well as to the interviewers who ensure that respondents remain motivated during the survey process (de Leeuw, 1992; Dillman et al., 2009). In spite of this many surveys are changing to alternative modes of data collection because of the substantial costs associated with conducting face-to-face interviews, the increasing nonresponse rates, and the increasing number of survey requests (Dillman et al., 2009; Williams & Brick, 2018). Also rapid technological advancements in recent years have led to changes in people's preference for data collection modes (de Leeuw & Hox, 2011; Peterson, Griffin, LaFrance, & Li, 2017).

The main alternate mode of data collection in face-to-face interviews is online surveys, which have been on the rise over the last 15 years (de Leeuw, 2018; Tourangeau et al., 2013). Online surveys are low cost, enable fast data processing, and are flexible in terms of providing more complex displays to respondents than face-to-face interviews (Beebe et al., 1997; Bethlehem & Biffignandi, 2011; de Leeuw, 2018; Dillman et al., 2009). The key concerns associated with online surveys are low response rates and susceptibility to satisficing, due to low motivation (de Leeuw, 2018; Kaminska & Foulsham, 2013). However, low response rate do not necessarily lead to nonresponse bias, and therefore is no longer an indicator of survey risk (Groves, 2006; Krosnick, 1999; Sturgis, Williams, Brunton-Smith, & Moore, 2017). Considering the rise of online surveys as an alternative to face-to-face interviews there is a need for a clear understanding of their similarities and differences in terms of data quality.

Direct evaluation of differences in data quality between face-to-face and online surveys is complicated because gold-standard criterion variables are not available (Dillman et al., 2009). Additionally, differences in survey estimates across face-to-face and online surveys consist of selection and measurement effects that are confounded (Vannieuwenhuyze & Loosveldt, 2013). On the strength of this, Kantar Public in the United Kingdom (UK) conducted a Community Life Survey (CLS) study, and assessed differences in data quality by applying nonresponse and attrition weighting to balance sample selection effects between general population samples interviewed online and face-to-face (Williams, 2017b). The study concluded that an online survey with low response rate probably produced data of a *higher* quality than a face-to-face survey with a considerably higher response rate. Given the longstanding consensus in survey research on the superiority of face-to-face interviewing, this must be considered a surprising conclusion. If this conclusion is robust, it is very important because it opens the possibility of conducting surveys considerably more cost-effectively, without incurring a decline in data quality.

The third paper entitled "Do low-response rate online surveys provide equal or better data quality than high response rate face-to-face designs? Separating sample selection from measurement effects", aims to assess whether this conclusion will be supported by applying propensity score matching approach, which is a robust method and well suited for estimating causal effects in mixed-mode designs. This paper uses the same CLS data that was used by Williams (2017b). This paper has two main objectives. The first objective adds to our understanding whether a low response rate online survey can produce data of equal or even better quality than face-to-face surveys. The second objective addresses how effective the propensity score matching approach is in removing selection effects and whether different formulations of survey weights in propensity score models and outcome analysis has an impact in the estimation of mode effect. The results show that the majority of total mode effects between the online and face-to-face surveys is due to measurement rather than selection effect. The results also that that propensity score matching cannot be assumed to be a completely effective method for removing selection effects in surveys with different modes of data collection. In addition, specification of different formulations of survey weights in propensity score models and outcome analysis are found to have no impact on the estimates of mode effects. These results indicate that survey designers need to be careful when switching from costly face-to-face interviews to more affordable online surveys. Additionally, results indicate that propensity score matching requires further optimisation and improvement to effectively remove selection effects in mixed-mode surveys.

Chapter 2 An assessment of the utility of a Bayesian framework to improve response propensity modelling (Paper 1)

2.1 Introduction

It has become more difficult in recent years to conduct high quality surveys because of declining response rates and increasing survey costs (Carlson & Williams, 2001; de Leeuw & de Heer, 2002). Declining response rates reduce stakeholder confidence in the ability of surveys to inform public policy due to concerns about the representativeness of samples and the generalisability of findings to wider populations. Therefore, survey researchers are keen to understand and address the factors which influence nonresponse. Such factors include socioeconomic and sociodemographic attributes of members of the public (Gjonça & Calderwood, 2004; Goldberg et al., 2001), salience of survey topics (Groves, Cialdini, & Couper, 1992), and survey design characteristics (Fan & Yan, 2010; Moss, 1981).

The increase in nonresponse rates over recent years has resulted in interest among survey practitioners in developing improved understanding of nonresponse behaviour. This has led to the development of response propensity (RP) models (Särndal & Swensson, 1987) to investigate the correlates of nonresponse (Durrant & Steele, 2009). Increases in nonresponse rates have also promoted research on the effect of strategies such as offering incentives (Singer, Hoewyk, et al., 1999), training of interviewers (Schnell & Trappmann, 2006) and implementation of mixed-mode designs (de Leeuw, 2005, 2018). However, for effective implementation of responsive design strategies it is necessary to know which sample units are more or less likely to respond and this is where RP models can be effective.

Olson & Groves (2012) employed RP models to predict changes of individual response propensities under responsive and adaptive strategies. Durrant et al. (2015) showed that the predictive power of RP models for final call outcome and length of call sequence improves when information from most recent calls is included as explanatory variables. However, often the proportion of the variance of the response outcome in the RP models that is explained by the explanatory variables in terms of pseudo R^2 is low and ranges between 2 and 8 percent (Fricker & Tourangeau, 2010; Kreuter & Olson, 2011; Olson & Groves, 2012). Therefore, ways of improving the predictions of RP models remain an active and important area of survey research.

Some of the steps taken for improving the predictive power of RP models in responsive and adaptive designs include collection and/or use of new auxiliary data such as paradata which are data about the survey process (Biemer et al., 2013; Durrant et al., 2015, 2017; West, 2011). Another way is to explore statistical methods for improving RP models. One possibility in the latter context is the use of a Bayesian approach, which is the focus of this paper. It investigates the utility of a Bayesian modelling approach for RP models. In particular, it evaluates whether or not specification of informative priors using existing knowledge about the response propensities of population sub-groups improves predictive and discrimination power. Model performance is assessed using a range of evaluation criteria such as sensitivity, specificity, area under the receiver operating characteristic (ROC) curve, positive and negative predicted values. Data from the UK Household Longitudinal Study (Understanding Society) are used.

The remainder of this chapter is structured as follows: Section 2.2 provides background on and motivation for the study. Section 2.3 describes the Understanding Society survey and explains how the analysis samples are constructed. The methodology for the analysis is then outlined in section 2.4, followed by results in section 2.5. Section 2.6 summarises the key findings, acknowledges limitations, and draws out implications for survey practice.

2.2 Background and Motivation

RP models as tools for evaluating nonresponse behaviour in surveys were introduced by David, Little, Samuhel, & Triest (1983) who extended the propensity score theory of Rosenbaum & Rubin (1983). RP models produce a single score as a function of variables that are observed for both respondents and non-respondents (Kalton & Flores-cervantes, 2003). Traditionally, the method for estimating response propensities is a logistic regression model where the outcome is a binary indicator of survey response versus nonresponse. Response propensities have been used for a variety of purposes, including obtaining a better understanding of nonresponse and associated mechanisms (Durrant & Steele, 2009), developing nonresponse weights (Little, 1986), providing guidance on interventions for adaptive and responsive survey designs (Groves & Heeringa, 2006), calculating representativeness indicators such as R-indicators and coefficients of variation (CVs) (Schouten & Cobben, 2007), and for predicting response outcomes either during or at the end of data collection (Durrant, D'Arrigo, & Müller, 2013; Durrant et al., 2011, 2015, 2017).

The effectiveness of RP models in helping survey researchers implement fieldwork decisions is hindered by their generally low predictive power. For example, a RP model developed by Olson & Groves (2012) to investigate within-person variation in response propensities over the data collection period had a pseudo R^2 of 2.2%. Olson et al. (2012) investigated the effect of respondents' choice on their preferred survey mode using RP models and obtained pseudo R^2 ranging between 3.2% and 7.7%. The low predictive strength is a result of the use of auxiliary variables which are not strongly correlated with response outcomes (Kreuter, Olson, et al., 2010). This implies that the choice of the auxiliary variables affects response propensities and tends to be specific to both the units sampled and the survey conditions in wider society (Brick & Montaquila, 2009).

One of the strategies adopted to improve the fit of RP models involves the collection of new kinds of auxiliary variables and paradata to be used as predictors (Biemer et al., 2013; Blom, 2009; Peytcheva & Groves, 2009; Sinibaldi & Eckman, 2015; Sinibaldi et al., 2014). For example, Durrant et al. (2015, 2017) found that the inclusion of call record variables, especially from the most recent calls, improves the predictive power of RP models from 9% to 26% in pseudo R^2. Sinibaldi & Eckman (2015) used interviewer observations at call level and observed an improvement of the RP model's predictions in terms of both pseudo R^2 and the AUC of the ROC curves. Likewise, Blom (2009) showed that explanatory power improves when demographic variables are combined with paradata using European Social Survey (ESS) data for nonresponse adjustment.

Historically, RP modelling has been implemented within a frequentist statistical framework, (Durrant et al., 2015, 2017; Olson & Groves, 2012; Olson et al., 2012; Sinibaldi & Eckman, 2015). However, the Bayesian framework for statistical modelling is becoming increasingly popular in the social sciences and holds promise for improvements to RP modelling. The main difference between frequentist and Bayesian frameworks lies in the treatment of the observed data and the interpretation of uncertainty. Statistical inferences based on the frequentist framework make probability statements about random events with known probabilities and to long run frequencies, while Bayesian statistics treats all unknown quantities as random variables and represents uncertainty over those quantities using probability distributions. (Fearn et al., 2004). In addition, Bayesian inferences are exact since they are conditioned on observed data satisfying the likelihood principle, unlike frequentist inference that relies on asymptotic approximations (Steel, 2007).

The starting point of Bayesian analysis is expressing prior knowledge about unknown parameters in the form of prior distributions. The observed data is then combined with the prior distribution using Bayes' theorem to obtain an updated prior in the form of posterior distributions (Fearn et al., 2004; Gill, 2014; Simon, 2009). In many practical situations, there is little or no previous knowledge on the phenomenon of interest. This leads to the specification of 'vague' prior distributions that have minimal influence on the analysis (Gill, 2014). However, when researchers have some existing knowledge about the parameters of interest it is possible to specify informative prior distributions (Gill, 2014). Information to specify informative priors can be derived from existing data, expert opinion, pilot studies, and scientific literature.

In the context of a longitudinal survey, posterior distributions that summarise knowledge on the parameters at the current wave may be used as prior distributions for subsequent waves. This may lead to better and more stable estimates of parameters and, therefore, improved predictions. This procedure is known as sequential Bayesian updating (SBU) (Lindley, 1972). SBU has been applied in fields such as traffic analysis (R. Yu & Abdel-Aty, 2013), big data applications using web sourced data (Oravecz, Huentelman, & Vandekerckhove, 2015), and in clinical trials (Viele et al., 2014). For example, Oravecz et al. (2015) found SBU to be computationally efficient in their analyses involving web sourced Alzheimer's Dementia data. In their study, model parameters were updated as new data became available without the need to repeatedly compute the likelihood. Schoot, Broere, Perryck, & Loey (2015) also found that Bayesian models with informative priors tended to have increased power and reduced bias when implemented for datasets with small sample size.

The use of a Bayesian approach using informative priors has attracted the attention of survey methodologist in recent years (Schouten et al., 2018; Wagner, 2016). For instance, Schouten et al. (2018) presented a Bayesian framework that included and updated prior knowledge for survey design parameters related to response and costs .They demonstrated the utility of informative priors derived from historic survey data and expert opinion when incorporated in the Bayesian model using the Dutch Health Survey. They found that a correctly specified Bayesian model leads to robust results compared to a "non-Bayesian model" especially when used for smaller sample sizes. Wagner (2016) showed that using informative priors of fixed coefficients in RP models derived from the data collected in the last 21 days of the previous quarter of a survey improved classification power from a low of 40% to a higher value of 64%. Both Schouten et al. (2018) and Wagner (2016) also noted that timelines of the historical survey data is of crucial importance with prior information derived from early stages of data collection more valuable compared to that obtained in later stages. However, these studies did not explore the utility of a Bayesian approach when informative priors are derived from auxiliary variables such as paradata in a longitudinal context which this study seeks to add to the literature.

A frequently voiced concern in the use of Bayesian analysis is the 'subjectivity' associated with the choice of informative priors (Bijak & Bryant, 2016). Therefore, when informative priors are used, it is imperative to quantify prior impact under different specifications which involves fitting models with vague priors and altering the variance component of informative priors (Evans, Jang, & Jan, 2011; Gill, 2014). This process is referred to as global sensitivity analysis, it quantitatively assesses the impact of priors on the likelihood function of the model

(Gill, 2014). The Bayesian approach also tends to be computationally demanding when implemented in models which are highly parameterised and have many cases (Lam, 2008; Rue et al., 2009). However, recent advances in hardware speed and the introduction of faster computation platforms such as integrated nested Laplace approximation (INLA) have effectively reduced the severity of this problem (Rue et al., 2009). Finally, more widespread implementation of Bayesian models among social scientists is currently achievable due to developments of user friendly Bayesian modelling platforms such as BUGS (Bayesian inference Using Gibbs Sampling), MLwiN, and STAN (Browne et al., 2016; Carpenter et al., 2016; Lunn, Spiegelhalter, Thomas, & Best, 2009).

2.3 Data

Understanding Society is a large-scale household longitudinal survey which collects information on health, work, education, income, family and social life and aims to explain their stability and changes among individuals and households living in the UK (Buck & McFall, 2012; Knies, 2014). The survey comprises three sample components: the general population sample (GP), the ethnic minority boost sample (EMB), and the British Household Panel Survey (BHPS). The survey uses a multi-stage sample design with clustering and stratification. Households are clustered within interviewers and within the primary sampling units (PSU). The details of sample selection are provided by Lynn (2009). The study also uses call record data and interviewer observation variables (Knies, 2014). The survey aims to achieve interviews with all individuals in sampled households who are aged 16 years and above and young people aged 10-15.

2.3.1 Analysis Sample

This study uses the GP sample covering Great Britain (GB) only for the analysis, since the Northern Ireland (NI) sample does not contain call record data, which are required in the analysis, since previous wave call record data is incorporated in this model. The BHPS sample is excluded from the analysis because it was not included in Wave 1 of Understanding Society which is needed for the analysis. The EMB sample is also excluded from the analysis as the rules for selection are different from the main GP sample and as this study is not interested in the specifics of this subsample. Here the focus is on the first five waves of data, collected between January 2009 and December 2014. The waves are linked pairwise (wave 1 and wave 2; wave 2 and wave 3 etc.) using unique personal identifiers. The auxiliary variables for the response outcome are obtained from the previous wave and therefore there are four datasets used for analysis. Details about the four pair-wise datasets across the five waves are presented in Table 2-1.

As the Bayesian analysis with informative priors adds more value when used for smaller sample sizes of observed data, this study also applied informative priors to subsamples. This is aimed at investigating whether data has dominating effect on the posterior results irrespective of the amount of previous wave data used to derive informative priors. Therefore, subsamples, which consist of 2%, 5% and 10% of the main sample, are randomly selected and the analysis repeated on each subsample.

Table 2-1: The number of households on each wave linked to previous wave auxiliary data, missing cases and wave final sample size

Waves	*Households linked to previous wave auxiliary data*	*Missing cases (survey and interview observations)*	*Final sample*
1 and 2	24,738	288 (1.2%)	24,450
2 and 3	19,791	618 (3.1%)	19,173
3 and 4	17,856	490 (2.7%)	17,366
4 and 5	16,705	578 (3.5%)	16,127

2.3.2 Dependent and Explanatory Variables

The response variable modelled in this study is the final call outcome. The final call outcome has a successful response if at least one interview is conducted in a household denoted by (1), otherwise unsuccessful (0).

The choice of household level response is motivated by the fact that in this study the interest is in including variables from the call record data (paradata) in nonresponse models, which here (as in most other surveys) are only recorded at the household level. The definition of the response outcome is informed by Durrant et al. (2015, 2017) and its distribution is presented in Table 2-2.

Table 2-2: Distribution of the final call outcome in the Final Analysis Sample

Waves	*At least one interview*	*No interview*	*Total*
2	18,928 (77.4%)	5,522 (22.6%)	24,450
3	15,741 (82.1%)	3,432 (17.9%)	19,173
4	15,016 (86.5%)	2,350 (13.5%)	17,366
5	14,271 (88.5%)	1,856 (11.5%)	16,127

The analysis also considers the length of call sequence as the response outcome since survey managers may want to know which households are more likely to respond in a shorter period. This knowledge potentially can help, saving survey efforts and costs.

The explanatory variables available for the analysis are split into four groups:

1. Geographical and design variables: (GORs, urban/rural indicator, and month and year of household issue).

2. Survey variables: (lone parents, pensioners in household, employment status, number of cars, highest education qualification in household, household income, tenure; household size).
3. Interviewer observations: (accommodation, relative condition of property, presence of unkempt garden in address, conditions of surrounding houses, presence of trash/litter/junk in street or road, heavy traffic on street or road, presence of car/van and children in household).
4. Call records data: (length of call sequence, proportion of noncontacts, proportion of appointments, proportion of contacts, proportion of other call outcomes and proportion of interviews). The denominator of all the proportions is the length of sequence.

2.4 Methodology

The final call outcome is modelled using binary logistic regression (Hosmer & Lemeshow 2000). Let the binary response of household i be denoted by $y_i, i = 1, \dots\ n$. The response variable for the final call outcome is given as:

$$y_i = \begin{cases} 1 & \text{successful final call outcome (at least one interview)} \\ 0 & \text{unsuccessful final call outcome (no interview).} \end{cases} \quad (2.1)$$

for each household i, response probabilities for y_i are denoted as $\pi_i = Pr(y_i = 1)$ and $(1 - \pi_i) = Pr(y_i = 0)$. Observed responses y_i are proportions with the standard assumption that they are binomially distributed

$$y_i \sim Bin(n_i, \pi_i) \quad (2.2)$$

where n_i is the number of trials. The logistic regression model is defined as

$$logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j = \mathbf{B}^T \boldsymbol{X}_j \quad (2.3)$$

where $\mathrm{B} = (\beta_0, \beta_1, \dots, \beta_j)$ is a vector of regression parameters and X_i is a vector of covariates at household level.

The Bayesian logistic RP models are fitted using the INLA package (Fong, Rue, and Wakefield 2010; Rue, Martino, and Chopin 2009) in the R statistical software. INLA produces fast and accurate approximations compared to Markov chain Monte Carlo (MCMC) alternatives for latent Gaussian models (Rue et al. 2016) . INLA's Bayesian inferences are approximated deterministically, making it practically feasible to fit models which contain many regression parameters and complex structures (Rue et al. 2009). A detailed description of the INLA methodology can be found in Rue et al. (2009).

To complete the model described in Equation 2.3, normal distributions denoted by $\beta_k \sim N(\mu_k, \sigma_k^2)$, $k = 1, \dots, j$ are specified as priors for regression parameters (Gelman, Jakulin, Pittau, & Su, 2008). The normally distributed priors and are not conjugate with the likelihood of the data and they are incorporated in the model by altering the weighted least squares step of the algorithm and augmenting the approximate likelihood with the prior distribution (Gelman et al., 2008). The basic idea of conjugacy implies that prior-to-posterior updating yields a posterior that is also in the same distribution family. The analysis starts by specifying vague normal prior distributions denoted by $\beta_k \sim N(0, 10000)$ for regression parameters in the model predicting the wave 2 final call outcome. Then posterior summaries are obtained from the INLA that summarise the knowledge on the parameters given the data. The posterior results are summarised in terms of the means that express the updated knowledge of the regression parameters and their variances. The estimated posterior means and variances are then used as informative priors for the subsequent wave analysis.

The global sensitivity analysis on specifications of different prior distributions will be assessed by altering the variance component of informative priors (Gill, 2014). Since the normal distribution is a location-scale family distribution, altering the variance parameter provides the best way of assessing the sensitivity of the informative priors because the variance influences the posterior results' dispersion. Therefore, posterior sensitivity is assessed by multiplying the informative prior variance parameters by a factor of 0.1, 2.0, 5.0, 10.0, and 100.0 and observing the effect on the resulting posterior distribution in terms of predictive and discriminative measures. This spectrum of mis-specified priors gives the relative weighting of the variance for the likelihood function from highly to less informative priors. As an uncertainty measure, variance works well for determining prior impact where higher variances "flatten" out the informative prior making it less informative (Gill, 2014). The different prior specifications used in this study are presented in Table 2-3.

Table 2-3: Different prior distributions used for Bayesian response propensity models in each wave

Prior type	*Specification of prior distribution for regression parameters in wave n*	*Model name*
Vague	$\beta_k^n \sim N(0{,}10000)$	M1
Informative	$\beta_k^n \sim \mathrm{N}\left(\beta_k^{n-1}, \left(\sigma_k^{n-1}\right)^2 \times 0.1\right)$	M2
	$\beta_k^n \sim \mathrm{N}\left(\beta_k^{n-1}, \left(\sigma_k^{n-1}\right)^2\right)$	M3
	$\beta_k^n \sim \mathrm{N}\left(\beta_k^{n-1}, \left(\sigma_k^{n-1}\right)^2 \times 2\right)$	M4
	$\beta_k^n \sim \mathrm{N}\left(\beta_k^{n-1}, \left(\sigma_k^{n-1}\right)^2 \times 5\right)$	M5
	$\beta_k^n \sim \mathrm{N}\left(\beta_k^{n-1}, \left(\sigma_k^{n-1}\right)^2 \times 10\right)$	M6
	$\beta_k^n \sim N\left(\beta_k^{n-1}, \left(\sigma_k^{n-1}\right)^2 \times 100\right)$	M7

The posterior results from the best fitting model in each wave are used as informative priors for the subsequent wave models. This analysis does not consider correlation structures among the regression parameters due to the large number of explanatory variables used in the models, which make it computationally demanding. The model parameters for frequentist models are fitted using Maximum Likelihood Estimation (MLE) in Stats Package in R statistical software for comparison purposes (R Core Team, 2015).

2.4.1 Model Selection

The variables included in the final RP models are selected in a two-step process for both the frequentist and Bayesian models. The first step uses univariate analysis to identify those variables that are unconditionally related to the final call outcome. The explanatory variables with p values < 0.05 for frequentist models and 95% credible intervals that do not cover zero for Bayesian models are selected for inclusion in the multivariable analysis. In addition, at this stage, contingency tables with zero or low cells that may cause numerical problems in models are grouped (i.e. categorical levels that have few cases are combined into one group). The correlations between each of the explanatory variables and the final call outcomes are assessed using Cramer's *V,* a measure of correlation for categorical variables (Liebetrau, 1983).

The next step involves fitting and refitting of frequentist and Bayesian multivariable models using a forward selection approach (Hosmer & Lemeshow, 2000). The explanatory variables that are significant in both the frequentist and Bayesian multivariable models are retained. This ensures that explanatory variables selected for the final frequentist and Bayesian models are the same. In the event that the frequentist and Bayesian approaches do not produce exactly the same models then an appropriate decision on which variables to include in the final analysis is necessary. For frequentist and Bayesian approaches, the Akaike information criterion (AIC) and Watanabe Akaike information criterion (WAIC) measures will be applied in selecting the final models respectively (Freese & Long, 2006; Gelman, Hwang, & Vehtari, 2013). Both AIC and WAIC are measures of predictive accuracy and are typically defined based on the deviance. AIC is calculated using the maximum likelihood estimate, while WAIC is computed using log pointwise predictive density and both adjust for the effective number of parameters. The process of variable selection is only applied on the wave 2 data with models for subsequent waves employing the same set of explanatory variables.

2.4.2 Model Evaluation

The fit of the frequentist and Bayesian models is evaluated using AIC and WAIC measures (Freese & Long, 2006; Gelman et al., 2013) with the lowest AIC and WAIC values indicating the best fit compared to alternative models. In addition, the proportion of variance in the final call outcome accounted for by the explanatory variables in the frequentist models is assessed using a Nagelkerke pseudo R^2 (Nagelkerke, 1991). The closer the values of the Nagelkerke pseudo-R^2 are to 1 the higher the proportion of the variability in the final call outcome is explained by the model.

Although the AIC, WAIC, and pseudo R^2 are useful for evaluating model adequacy, they cannot assess the accuracy of the model predictions of correctly classifying non-respondents and respondents (Plewis et al., 2012). In addition, using WAIC and AIC makes it difficult to compare the predictive performance of frequentist and Bayesian models directly. These challenges are addressed by adopting measures for classification, discrimination (sensitivity and specificity), prediction (positive and negative predicted values), and (AUC) of the ROC which addresses the issues of arbitrary cut-off values in discrimination and prediction (Durrant et al. 2015; Pepe 2003; Plewis, Ketende, and Calderwood 2012).

An overall summary of predictor power is the proportion of the correct classifications referred to as the classification rate, which measures the proportion of households that would be correctly classified by the model. Sensitivity is the proportion of households that experience no interview and are correctly predicted as such, while specificity is the proportion of households which are correctly predicted as providing at least one interview. The positive predictive value (PPV) is the probability that a household is indeed a nonresponse given that it is predicted as nonresponse, while the negative predicted value (NPV) is the probability that a household is indeed a response given that it is predicted as a response. The R package epiR is used to evaluate classification rate, sensitivity, specificity, positive and negative predictive values (Mark et al., 2016).

The AUC of the ROC curve measures the model's ability to discriminate between households which were not interviewed and those which had at least one interview (Plewis et al., 2012). The AUC represents an overall accuracy of model predictions and has a range of 0.5 to 1.0. A value of 0.5 means the model predictions are no better than random guessing, while a value of 1.0 represents perfect discrimination between households that experience at least one interview and those which do not. The ROC curves are implemented in the R pROC package, a tool for visualising , smoothing and comparing ROC curves (Robin et al., 2011).

These measures are evaluated using out-of-sample predictions of test data because this is less sensitive to outliers and overfitting (Hastie, Tibshirani, & Friedman, 2009). This is done by partitioning the analysis samples into training and testing subsets which are used for model fitting and evaluation respectively (Hastie et al., 2009). In this study, 50% of the sample is used for an out-of-sample prediction. The training and testing subsets are obtained by randomly splitting the given wave data using the R caret package (Kuhn et al., 2016). Cross-validation was done by splitting each dataset twice into a training dataset and a validation dataset.

2.5 Results

The results presented are for 23 models estimating response propensities of final call outcomes for Waves 2, 3, 4, and 5. For Wave 2, vague priors are specified for the Bayesian models because previous wave data is not available. A total of 9 models are fitted for the final call outcome at subsequent Waves (Waves 3, 4 and 5). The posterior summaries from the Bayesian model with the lowest WAIC among alternative models in the current wave are used as informative priors for the subsequent wave analyses. At each wave, a model with vague priors is used as the reference when comparing the predictive performance of informative prior models. Cramer's V values obtained for the final call outcome are less than 0.26 indicating weak bivariate relationships between response variable and explanatory variables used in all models.

2.5.1 Assessment of model fit using WAIC, AIC and Nagelkerke pseudo R^2

Table 2-4 presents pseudo-R^2 coefficients and the values for AIC and WAIC for the 23 models in Waves 2, 3, 4 and 5. Table 2-4 shows that in Wave 3, all models with different specifications of informative priors have lower WAIC values compared to model with vague priors except models (M2) and (M3) which have higher WAIC values indicating a poor model fit. In Wave 4, only model (M2) has a higher WAIC value in comparison to the vague prior model. In Wave 5, models with informative priors have higher WAIC values compared to a model with vague prior while models (M5), (M6), and (M7) have WAIC values similar to that obtained in vague model. This may be due to the introduction of mixed-mode data collection in the third and fourth quarters of Wave 3 in which interviews of unproductive households were attempted using computer assisted telephone interviewing (CATI) in place of computer assisted personal interviewing (CAPI) (Baghal, Jäckle, Burton, & Lynn, 2016). This potentially makes information from previous waves less relevant to the Wave 5 final call outcome. It is further observed that model (M2) which has tight variance and is considered highly informative has a poor fit in comparison to other models in all waves.

.

Table 2-4: Evaluation criteria for frequentist models using (Akaike information Criteria (AIC), Nagelkerke's pseudo R² and Watanabe Akaike Information Criteria (WAIC)) for Bayesian models

Wave	*Model*	*AIC*	*Nagelkerke R² (%)*	*WAIC*
1 and 2	Frequentist	12559.00	7.40	-
	M1	-	-	12561.34
2 and 3	frequentist	8700.50	4.91	-
	M1	-	-	8701.35
	M2	-	-	8704.27
	M3	-	-	8856.92
	M4	-	-	8692.62
	M5	-	-	8695.86
	M6	-	-	8699.08
	M7	-	-	8701.32
3 and 4	frequentist	6864.60	5.76	-
	M1	-	-	6865.25
	M2	-	-	6868.08
	M3	-	-	6997.24
	M4	-	-	6854.12
	M5	-	-	6858.06
	M6	-	-	6861.73
	M7	-	-	6870.71
4 and 5	frequentist	5592.5	6.43	-
	M1	-	-	5594.18
	M2	-	-	5810.01
	M3	-	-	5626.60
	M4	-	-	5603.74
	M5	-	-	5593.14
	M6	-	-	5592.24
	M7	-	-	5594.14

Table 2-4 also shows that the Nagelkerke pseudo R² for the frequentist models are between 4.9% and 7.4% for the final call outcomes, which are similar to pseudo R² values of nonresponse models reported in previous studies (Olson & Groves, 2012; Olson et al., 2012). To summarise, these results indicate that the use of informative priors leads to a slight improvement of model fit in the earlier waves of the survey compared to models with vague priors. However, the performance of the Bayesian models is poorer at later waves. This difference between earlier and later waves could be due to substantive changes introduced to the survey fieldwork in later waves.In addition, this may have been caused by the reduction in strength of borrowed information in later waves due to temporal effect (Schouten et al., 2018). That is, conditioning on most recent data is expected to be more informative about households in comparison to later data because households' characteristics are more likely to remain the same in short-term. The RP models with informative priors that have larger variances (standard deviation multiplied by a factor of 10 and 100) tend to have WAIC values similar to those of vague priors' models.

2.5.2 Classification table and AUC for ROC Curves, sensitivity, specificity and, positive and negative predictive value

Table 2-5 presents the classification tables and AUC values for ROC curves based on 50% out-of-sample predictions. For classification tables it is expected that 50% of the cases for the final call outcomes are classified correctly by chance, with higher values relative to 50% usually depicting higher predictive powers of models. The observed classification values for all models are 82%, 87% and 88% in Waves 3, 4 and 5 respectively. These values are similar to the proportion of households which had at least one interview since classification rates tend to be overly sensitive to the dominant categorical level of the response (Agresti, 2013). These classification values for the final call outcomes show that models are not performing better than the observed distribution.

The AUC values of ROC curve greater than 50% indicate that any discrimination for the outcome of interest is not due to random variation, with values above 70% considered to offer better discrimination (Hosmer & Lemeshow, 2000). For the final call outcomes, Table 2-5 shows the AUC values obtained in all waves range between 62% and 64%, indicating a minimal discrimination. In all waves, the differences in AUC values for models with informative and vague priors range between $\pm 0.01\%$ and $\pm 0.03\%$ which are negligible. Although there is an indication of slightly higher AUC values for RP models with informative priors they are not statistically significant. Overall, the results show that the use of informative priors does not lead to significant improvement in the predictive power of the models. It is evident that use of previous wave information does not lead to significant changes in either classification values or AUC values for RP models. This means using previous wave data adds no additional strength to the predictions of the final call outcome.

Table 2-5 also shows improvement of sensitivity values for the final call outcome model (M3) in waves 3 and 4 relative to model (M1). The sensitivity values for model (M2) give a non-numeric value (NaN)[4] in all waves indicating that a tight informative prior does not correctly predict any households that were not interviewed. This is because the informative prior specified is very strong since it puts most of its mass on parameter values that are large in absolute value and therefore strongly influences the posterior inference. Considering that only a few households are not interviewed compared to those interviewed, conditioning on a

[4] NaN results when the fraction's numerator is zero

tight informative prior predicts that none of the households will be correctly predicted as non-interviewed (i.e. nonresponse). The mis-specified informative prior models with larger variances (for global sensitivity analysis) have similar sensitivity values as vague priors except in Wave 5 that have slightly improved sensitivity values. In addition, the specificity values for models with informative and vague prior models are similar in each wave. Sensitivity and specificity results show that the use of previous wave information does not improve the discrimination power of the models. Table 2-5 also shows that, the positive and negative predictive values for final call outcome models with informative priors and vague priors in waves 3, 4, and 5 are similar with negligible differences of ±1% . It is important to note that sensitivity, specificity, PPV and NPV values in Table 2-5 are integers because of the nature of data used. Otherwise, they can take any numerical values.

Table 2-5: Results of classification table and AUC of ROC curves, sensitivity, specificity, positive predictive values (PPV) and negative predictive values (NPV) for the final call outcome

Wave	*Modelling approach*	*Classification (%)*	*AUC (%)*	*Sensitivity (%)*	*Specificity (%)*	*PPV (%)*	*NPV (%)*
1 and 2	frequentist	77.5	64.3	52.0	78.0	3.0	99.0
	M1	77.5	64.3	53.0	78.0	3.0	99.0
2 and 3	frequentist	81.7	62.4	25.0	82.0	0.0	100.0
	M1	81.7	62.4	27.0	82.0	0.0	100.0
	M2	81.7	56.5	Nan	82.0	0.0	100.0
	M3	81.7	62.0	50.0	82.0	0.0	100.0
	M4	81.7	62.4	40.0	82.0	0.0	100.0
	M5	81.7	62.4	30.0	82.0	0.0	100.0
	M6	81.7	62.4	27.0	82.0	0.0	100.0
	M7	81.7	62.3	27.0	82.0	0.0	100.0
3 and 4	frequentist	87.0	62.6	40.0	87.0	0.0	100.0
	M1	87.0	62.6	40.0	87.0	0.0	100.0
	M2	87.0	58.4	Nan	87.0	0.0	100.0
	M3	87.0	62.4	50.0	87.0	0.0	100.0
	M3	87.0	58.4	Nan	87.0	0.0	100.0
	M5	87.0	62.7	25.0	87.0	0.0	100.0
	M6	87.0	62.7	40.0	87.0	0.0	100.0
	M7	87.0	62.7	40.0	87.0	0.0	100.0
4 and 5	frequentist	88.5	63.6	45.0	89.0	1.0	100.0
	M1	88.5	63.6	45.0	89.0	1.0	100.0
	M2	88.5	52.4	Nan	89.0	0.0	100.0
	M3	88.5	63.3	43.0	89.0	0.0	100.0
	M4	88.6	63.7	38.0	89.0	0.0	100.0
	M5	88.6	63.7	50.0	89.0	1.0	100.0
	M6	88.6	63.6	56.0	89.0	1.0	100.0
	M7	88.6	63.7	45.0	89.0	1.0	100.0

The additional analysis involving subsamples selected randomly from main sample had similar results in terms of discrimination and prediction power as those obtained from the main sample analysis and these are presented in Appendix A. These results show that the sample size of the data considered in this study does not have an impact on the discrimination and prediction power in response propensity models. In addition, analyses involving length of call sequences as an outcome produced similar results as those obtained for the response outcome. This indicates that the use of informative priors based on previous waves does not lead to significant improvement of the predictive ability of the models. Additional analysis was also conducted with the aim of investigating whether strength of relationships between response and explanatory variables influences borrowed information from the previous wave using income and employment variables from this study's analysis sample. The results show that that these models' predictive and discrimination abilities are not different from those obtained in the main analysis. This indicates that the strength of correlation between variables in the data used for this analysis does not influence the effectiveness of the borrowed information.

2.6 Discussion

Household survey response rates have been steadily declining in developed countries over the last two decades. This has forced survey methodologists to introduce strategies such as mixed-mode designs (Couper, 2011), incentives (Mcgrath, 2005), improved training of interviewers (Schnell & Trappmann, 2006), and adoption of adaptive and responsive designs (Groves & Heeringa, 2006; Tourangeau, Brick, Lohr, & Li, 2016). However, these strategies are often not very successful and tend to increase survey costs. The overriding concern about nonresponse in household surveys is the weakening of the validity of inferences drawn from estimates due to unrepresentative samples. This in turn undermines the confidence of commissioners and key stakeholders in surveys for providing high quality evidence for understanding social and economic issues.

To better understand response behaviour and counter the negative effects of rising nonresponse rates there is a need to improve the predictive power of nonresponse models, which are generally low. In addition to contributing theoretical insights, RP models can be used to develop nonresponse weights and to underpin strategies for monitoring fieldwork in responsive and adaptive survey designs (Groves & Heeringa, 2006; Kim & Kim, 2007; Schouten et al., 2012). Recent studies have used paradata to improve the predictive power of nonresponse models (Kreuter et al. 2010; Kreuter & Olson 2013; Sinibaldi & Eckman 2015; Durrant et al. 2015, 2017). However, more work is required to make RP models more effective. This study has evaluated the potential utility of a Bayesian approach to fitting RP models. This is in principle attractive because it enables incorporation of previous evidence on response propensities through the specification of informative priors.

The findings indicate that RP models with informative priors are not significantly better in terms of WAIC when compared to vague prior models. Although models with informative priors have a slightly better predictive accuracy in terms of WAIC compared to models with vague priors, their specificity values are similar in each wave. Some small gains in sensitivity values for models with informative priors were observed in earlier waves of this study but this diminished and reversed in later waves. We speculate that this may have been due to the introduction of a mixed-mode design during the last two phases of Wave 3 which makes earlier information about the correlates of response from earlier waves less relevant. This implies that incorporating the most recent wave data as informative priors into RP models, improves their predictive power compared to using earlier wave data. This supports findings by Schouten et al. (2018).

It is also observed that altering the variance component of the informative prior did not produce notable changes in the range of the predictive and discrimination measures, an indication of robust results obtained in large samples irrespective of the specification of informative priors. The discrimination values indicate that models with better fit in terms of WAIC do not generally translate into having better discriminative power. Also, the AUC values as well as, positive and negative predicted values from models with informative priors showed no improvements in their predictions when compared to models with vague priors. In addition, discriminative and predictive results obtained from subsamples and subgroups are similar to those found in the main analysis.

An important assumption in this study involved specifying no correlations among regression parameters, which is informed by weak correlations between explanatory variables and also the complexity involved in trying to incorporate covariance structure with higher dimensionality (due to many explanatory variables) into the model. The length of call sequences as response outcome was also analysed. The discriminative and predictive results were similar to the results reported in this paper. It observed that using different samples with small sizes leads to similar conclusions as the main sample. However, it is important to note that subsamples were obtained randomly from the main data and it is probable that using a different survey with a small sample size might lead to different results.

The results show that, at least for this study, the use of informative priors derived from previous wave data in RP models leads to negligible improvement of the predictive and discriminative ability of the models. First, for effective borrowing of information from previous waves it is expected that explanatory variables in the models are able to optimally explain the variability of the outcome of interest. Therefore, if the explanatory variables in the RP model for a given wave are not explaining the variability of the response outcome well, it implies that such model's posterior estimates will also not provide additional information when used as informative priors for subsequent waves. That is, borrowing weak information from previous waves cannot improve predictive accuracy of subsequent waves since such informative priors do not bring any additional information. In addition, effectiveness of the informative priors derived from data of previous waves is dependent on how well auxiliary variables are correlated with final call outcome.

Previous studies suggest that available paradata and auxiliary data are not sufficiently correlated with the response outcomes for effective predictive accuracy in household survey responses (Olson & Groves 2012; Kreuter et al. 2010; Kreuter et al. 2010). However, in this study it was found that the strength of correlation between the outcome and explanatory variables in the data used for this analysis does not influence the effectiveness of the borrowed information. Furthermore, in longitudinal studies such as Understanding Society some new responsive and adaptive strategies are adopted as the survey progresses, which may also lead to changes in the auxiliary data compositions across waves for effective borrowing of previous wave's data via Bayesian sequential updating (Gill, 2014; Plewis et al., 2012; Schouten et al., 2018). According to Gill (2014), use of informative priors derived from previous data can be suspect if the data generating mechanism keeps changing over time relative to the data used for estimating the first posterior estimates. In Bayesian sequential updating it is not possible to include and control for any additional variables as the survey progresses since explanatory variables of the model are defined during the initial wave (Oravecz et al., 2015).

This study also noted that the data forming the likelihood component may also be having a dominating effect on the posterior results rendering information borrowed from previous waves less relevant. Usually the likelihood component depends on the sample size, which implies that the influence of an informative prior from previous waves decreases in longitudinal studies with large samples (Lynch, 2007; Schouten et al., 2018). However, the dominating effect of the likelihood component is not always dependent on the sample size but also how strongly the data contribute to the posterior. The results from the subsamples showed that previous wave data had a dominating effect on the posterior results irrespective of the specification of the priors. The results from mis-specified informative priors' shows robustness in the model specification since alterations in the variance component do not lead to large changes in the ranges of the predictive and discrimination measures. Although variance as an uncertainty measure works well for determining prior impact when altered, it is a poor detector of any prior and likelihood conflict which occurs when the prior puts all its mass in the tails of the likelihood. The prior and likelihood conflict may be detected using prior to posterior divergences measures: these measures were not considered in this study.

The results of this analysis contribute to a better understanding of the use of previous wave data as informative priors for response propensity models. Although the model results show no improvement in response predictions, these findings help to establish a new framework for the exploration of other sources of informative priors under different study settings. This author encourages researchers in this area to apply the method presented here to other applications to assess whether an improvement in performance could be achieved. Further work aims to explore other sources of informative priors such as elicitation from experts as described by O'Hagan et al. (2006) is recommended.

Chapter 3 Do interviewers moderate the effect of monetary incentives on response rates in household interview surveys? (Paper 2)[5]

3.1 Introduction

Declining response rates in developed countries have led survey researchers to focus on ways of improving survey cooperation (Brick & Williams, 2013; de Leeuw & de Heer, 2002). Amongst a wide range of measures, two key foci have been the role of interviewers and the use of incentives. In face-to-face surveys interviewers play an important role in gaining contact and cooperation from sample members (Campanelli & O'Muircheartaigh, 1999; Durrant et al., 2010; Hox & de Leeuw, 2002). The literature also provides abundant evidence regarding the effects of incentive on response rates (Laurie & Lynn, 2009; Pforr et al., 2015; Schröder, Saßenroth, Körtner, Kroh, & Schupp, 2013; Singer, Hoewyk, et al., 1999; Singer & Ye, 2013). However, existing research has not considered whether there is an interaction between interviewer behaviour and the effectiveness of incentives in gaining response. The aim of this paper is to investigate the role that interviewers play with respect to the effectiveness of incentives on survey response and cooperation. Findings will help improve our understanding of using incentives in interviewer-mediated household surveys.

Interviewers play a critical role as a link between the survey organisation and sample members. They are responsible for making contact and achieving cooperation from sample members and in doing so they communicate many aspects of the survey and its design to the sample members such as the survey topic, the importance of the study, the sponsor and the availability of incentives (West and Olson 2010). This may in turn motivate participation by sample members, as set out in the Leverage Saliency Theory (LST) (Groves et al., 2000). The LST is a conceptual framework that describes how multiple factors may influence a sample unit's decision to participate in a survey and that depends on how salient these factors are when an interviewer introduces the survey and makes a request for participation. Existing studies have documented various interviewer characteristics associated with their ability to stimulate survey participation. These include gender (Hansen, 2006; Hox & de Leeuw, 2002), years of experience (Durrant, Groves, Steele, et al., 2010; Hansen, 2006; Hox & de Leeuw,

[5] Kibuchi, E., Sturgis, P., Durrant, G. B., & Maslovskaya, O. (2018). Do interviewers moderate the effect of monetary incentives on response rates in household interview survey. J*ournal of Survey Statistics and Methodology*[5]

2002) and age (Blom, de Leeuw, & Hox, 2011; Durrant et al., 2010). Interviewers' characteristics influence the doorstep interaction between interviewer and sample member. For example, experienced interviewers are usually found to be more successful at obtaining cooperation because of their ability to tailor the survey request to the respondent's motivation and concerns (Campanelli & O'Muircheartaigh, 1999).

In addition, existing studies have covered extensively the role of incentives in motivating survey response (Singer, 2002; Singer, Groves, et al., 1999; Singer, Hoewyk, et al., 1999). Incentives are often used to facilitate survey recruitment and to stimulate participation among sample members (Church 1993; Singer et al. 1999). Usually incentives are effective in surveys that are expected to experience low response rates: they are offered as inducements to either compensate for the absence of interest in the survey topic or the lack of a sense of civic obligation (Groves et al., 2000; Singer, Hoewyk, et al., 1999). Studies find that incentives have a positive effect on response rates and that larger incentives induce greater survey participation but at a decreasing rate (Cantor et al., 2008; Singer, Groves, et al., 1999).

Given both the influence of interviewer and the use of incentives, it seems natural to consider the influence that interviewers have on the effectiveness of incentives on survey response and cooperation. It is possible that the effect of incentives will vary between interviewers, if some interviewers are more effective at leveraging incentives than others. For example, interviewers can tailor their introductions in a particular way such that they highlight the availability of a monetary or non-monetary reward at households that are most likely to be sensitive to it (Campanelli et al., 1997; Groves & Couper, 1998). Similarly, interviewers may feel more confident in their doorstep approach when they know an incentive is available which may positively affect their persuasive efforts (Singer & Ye, 2013). This joint influence is the focus in this paper. To analyse this interaction effect multilevel models are fitted to three different surveys which include a randomized incentive. The datasets used in this study are National Survey for Wales-Field Test 2015, National Survey for Wales-Incentive Experiment 2016, and the UK Household Longitudinal Study Innovation Panel Wave 1 (2008).

The structure of the paper is as follows. Sections 3.2 provides a literature review of how interviewers influence survey response and section 3.3 reviews the influence of incentives on response. Section 3.4 describes how interviewers might influence the effectiveness of the incentives. Section 3.5 describes the data, followed by the methodology employed for the analysis in section 3.6. The results are presented in section 3.7, and section 3.8 discusses some implications of the results for survey practice.

3.2 Background and Motivation

3.2.1 How interviewers influence response rates

It has long been recognised that interviewers play an important role in gaining response and cooperation (Campanelli et al., 1997). Normally, face-to-face surveys consistently achieve higher response rates than those undertaken by self-administration or by telephone, a difference that is largely attributable to the role of interviewers. The mechanism by which interviewers affect response rates varies , and depends on their diverse characteristics, attitudes and personalities (Blom et al., 2011; Campanelli & O'Muircheartaigh, 1999; Durrant et al., 2010; Hox & de Leeuw, 2002). This is the reason behind significant interviewer effects on survey contact and response that have been found across a range of sample designs and international contexts (Campanelli et al., 1997; Durrant et al., 2010; Durrant & Steele, 2009; Hox & de Leeuw, 2002). For example, Blom, Leeuw, and Hox (2011) found interviewer intra-class correlation coefficients of 0.27 for non-contact and 0.08 for cooperation across ten countries in the 2008 European Social Survey.

Interviewers brief respondents on key survey features such as incentives, topic and sponsor during their initial interaction and this may motivate survey participation (Couper & Schlegel, 1998). It is common for studies to send advance letters that contain the most important features of a survey (Groves & Couper, 1998; Groves et al., 2000; Singer et al., 2000). However, not all households read advance letters. Furthermore, individuals in households may read advance letters but may fail to pass that information onto other members of the household or be away at the time when an interviewer calls. For example, Singer, Hoewyk, and Maher (2000) and Brick et al (1997) found that advance letters overall increase response and cooperation rates by nonsignificant percentage points of 0 to 3. This indicates that interviewers will always play a crucial role in promoting response rates irrespective of the survey design adopted.

Interviewers who have the ability to adapt their approach to specific characteristics of sample units maximise response rates by identifying and presenting positively valued aspects of the survey to respondents using a technique labelled "tailoring". By tailoring, interviewers adjust what they say in an introduction based on factors they judge will be favourably received by the sample unit outside the constraints of the standardized interview (Groves & Couper, 1998). For example, an interviewer may make a mention of incentives that are on offer to those respondents who may have a high positive leverage on incentives. This may motivate their participation in a survey simply because incentives are offered even if they are not interested in other aspects of the survey.

The general conclusion of the conceptual mechanisms that make some interviewers more successful in tailoring their introductions are still not well understood (Blom & Korbmacher, 2013; Groves & Couper, 1996; West & Blom, 2017). Some studies suggest that experienced interviewers are better at tailoring their approaches to the range of household types and concerns (Groves & Couper, 1998; Lemay & Durand, 2002). This is because experienced interviewers are good at recruiting and maintaining interactions with potential respondents (Lemay & Durand, 2002), and also have lower appointment and interviewer postponement times (Durrant & D'Arrigo, 2014), even though they are often allocated to more difficult areas (Purdon et al., 1999; West & Blom, 2017).

Although Durrant and D'Arrigo (2014) did not find evidence that interviewers who are good at tailoring their approaches tend to be more effective, other studies have found a positive relationship between interviewers' self-confidence and the likelihood of achieving higher response and cooperation rates (Durrant et al., 2010; Groves & Couper, 1998; Hox & de Leeuw, 2002). This is thought to arise from the positive effect of confidence on the quality of doorstep interactions (Groves & Heeringa, 2006; Hox & de Leeuw, 2002). Singer and Kohnke-Aguirre (1977) also found that interviewer expectations and experience influence the overall behaviour of potential respondents towards survey participation. The conclusions that can be drawn from these studies may, therefore, imply that interviewer skills and experience in recognising, interpreting, and addressing visual cues and the confidence with which they approach the task of obtaining cooperation on the doorstep are likely to influence response and cooperation rates and increase the effectiveness of incentives.

3.2.2 How Incentives Influence Response Rates

To counter the low response rates due to the absence of other non-financial motivating factors such as engagement to survey topic, sense of civic or moral obligation and enjoyment of social interaction, many surveys offer incentives (Groves et al., 2000; Singer & Maher, 2000). Incentives improve response rates by either facilitating contact with potential respondents, or by stimulating their cooperation. Based on Leverage-Saliency theory, incentives may motivate survey participation of sample units who might otherwise not have participated. This is especially common among some respondents who have higher saliency on incentives, and therefore serves as an important leveraging factor in determining a respondent's decision to participate in surveys (Groves et al., 2000). Some respondents may perceive incentive payments as an act of compensation for the time and effort they have put into the survey process, as posited by Economic Exchange theory (Biner & Kidd, 1994). Incentives, especially those prepaid, are also effective at establishing the social exchange of trust making sample units more willing to reciprocate by participating in surveys (Blau, 1964). Lastly, incentives may invoke norms of reciprocity in a way that respondents feel a sense of obligation to provide an interview because of the incentive offered before the survey request (Biner & Kidd, 1994; Blau, 1964).

Incentives may be administered in several forms: monetary, non-monetary, pre-paid (i.e. unconditional) and promised (i.e. conditional) (Cantor et al., 2008; Church, 1993; Singer, Groves, et al., 1999). Monetary incentives are in the form of cash rewards while nonmonetary incentives are comprised of gifts such as pens, calendars, diaries, as well as summaries of the survey results (Lavrakas, 2008). Monetary incentives are more effective in motivating participation than nonmonetary incentives (Cantor et al., 2008; Church, 1993; Singer, Groves, et al., 1999). Prepaid incentives are offered prior to survey participation and tend to be effective in reducing refusals in comparison to promised incentives that are only provided upon completion of the survey (Cantor et al., 2008; Church, 1993; Lavrakas, 2008; Singer, Hoewyk, et al., 1999). However the effectiveness of the prepaid incentives does not necessarily imply that they are more cost effective (Brick, Montaquila, Hagedorn, Roth, & Chapman, 2005). Prepaid incentives used to secure refusal conversion are also as effective at improving response rate as those sent prior to the initial contact with the household (Cantor et al., 2008). Incentives are also more effective in self-administered surveys and surveys that are expected to have low response rates (Mercer et al., 2015; Singer, Hoewyk, et al., 1999). This is likely to be because there is more scope for the incentive to act as a replacement for non-monetary motivations among a larger pool of potential respondents.

The existing literature shows that use of the incentives has a positive effect on response rates in interviewer-mediated surveys (Cantor et al., 2008; Singer, Hoewyk, et al., 1999). A meta-analysis of 39 experiments in interviewer-mediated surveys by Singer, Hoewyk, et al. (1999) found that monetary incentives have a positive and significant effect on survey response. In the Singer, Groves, et al. (1999) analysis, they found that, on average, each dollar of incentive paid per interview results in about a third of a percentage point increase in response rates when compared with the zero-incentive condition. Lynn (2001) using a face to face 2000 UK Time Use Survey pilot study found that offering a £10 incentive led to a household response rate of 65%, higher than the no incentive group rate of 56%, indicating that incentives had a significantly positive effect on response rates. However, studies by Singer, Groves, et al. (1999) and Cantor et al. (2008) found that the size of the increase in the response rate declines with additional increases in the value of the monetary incentive (i.e. the 'dose'-response relationship is a curvilinear relationship).

Stratford, Simmonds, & Nicolaas (2002) used data from a National Travel Survey and found that a £10 conditional incentive significantly improved the response rate by 5-percentage points compared to no incentive. In the same study, Stratford, Simmonds, & Nicolaas (2002) found only 1 percentage point difference in response rates between £5 and £10 conditional incentives, indicating that response rates do not increase at the same rate as the amount of incentive offered. Boreham & Constantine (2008) using Understanding Society Innovation Panel data (Wave 1) found that response rates among the respondents offered a £10 incentive and those offered £5 rising to £10 per adult if all adults in the household completed their CAPI interviews in person were the same, at 61%. Therefore, simply offering higher valued incentives does not necessarily lead to a linear increase in response rates but a levelling off effect usually occurs (i.e. curvilinear relationship between increases in incentives and response rates) (Cantor et al., 2008; Gelman, Stevens, & Chan, 2002; Mercer et al., 2015). Incentives may also improve time-efficiency and cost-effectiveness in surveys because they promote early responses (Lavrakas et al., 2012).

The amount of incentive required for both recruitment and retention of respondents may vary depending on the sensitivity of the study. In general, amounts offered in longitudinal studies are greater than in cross-sectional studies because of the need to retain panel members over time (Singer & Ye, 2013). In the United Kingdom, several studies have shown that giving a £10 incentive improves response rates by a significant percentage of 4 to 9% (Hanson, Sullivan, & Mcgowan, 2015; Lynn, 2001; Stratford et al., 2002). Although offering £5 incentive may improve response rates in comparison to no incentive group the increment is not always significant (Aumeyr et al., 2017; Boreham & Constantine, 2008; Stratford et al., 2002). In general, the existing evidence demonstrates that monetary incentives have a robust, positive effect on the probability of survey cooperation.

3.2.3 How interviewers may influence the effectiveness of incentives

Existing studies attribute the positive effects of incentives on response rates primarily to respondents' behaviour and perception (Currivan, 2005; Patrick et al., 2013; Singer, 2002; Singer, Hoewyk, et al., 1999). Although these studies show that incentives have an independent positive effect on response rates, they do not rule out the possibility that the interviewer might moderate the incentive effect. There are good reasons to assume that they might. First, it is likely that interviewers expect those sample members receiving an incentive to be more cooperative and therefore may be more confident when approaching them, leading to an outcome that is expected (Hox & de Leeuw, 2002). This may result in an increased survey response because confident interviewers are known to have good powers of persuasion which greatly enhance the chances of gaining cooperation among respondents (Groves & Couper, 1996; Singer, Hoewyk, et al., 1999; Singer & Ye, 2013). Second, interviewers may also attempt to tailor their interviews by heightening the salience of the incentives at addresses where they believe these are likely to be effective (Groves et al., 2000). Third, interviewers are the primary conduit of the information between survey organisations and sample members, so are essential for ensuring that potential respondents are aware that an incentive is available. This is because while most surveys highlight incentives in advance letters, many respondents do not open, let alone read their mail (Stoop, 2005). It is likely that those people who do not read advance letters are busy and uninterested in the survey topic and therefore more susceptible to monetary incentives.

Given the substantial attention paid to both interviewers and incentives in boosting response rates, surprisingly few studies have considered both incentives and interviewers in the same study. Lynn (2001) investigated interviewer expectations and attitudes towards incentives using the 2000 UK Time Use Survey pilot that included an incentive experiment. Lynn also investigated interviewer perceptions regarding the incentives offered to respondents, using a focus group. He found that approximately half the interviewers believed that incentives had little or no effect on response and cooperation rates, while the other half felt that incentives had a negative effect on response rates. While Lynn's study did not aim to assess whether interviewers varied in how successful they were at using incentives to increase response rates, he was able to demonstrate that interviewers vary in their beliefs about the effectiveness of incentives and that these beliefs may not always be accurate.

Singer et al. (2000), investigated effects of incentives on interviewers, using data from the Survey of Consumer Attitudes, a telephone survey of the American public. Singer and colleagues randomly assigned interviewers and respondents to three groups: in groups 1 and 2 respondents received an advance letter and $5, while respondents in group 3 received only the advance letter. Interviewers' in-group 1 were unaware of the incentive, while interviewers in groups 2 and 3 were made aware of the incentive level via messages on their computers. Interviewers in groups 1 and 2 achieved response rates of 76% and 75%, respectively, compared to 62% for interviewers in group 3. The difference of 1 percentage point between interviewers for group 1 and interviewers for group 2 who were aware of the incentive was not significant. Singer et al. (2000) concluded that, although the unconditional incentive increased the response rate, interviewer expectations about the likely cooperativeness of sample members had no additional effect. That is, incentives had a direct influence on respondents but not through their effects on interviewer expectations. They further suggested that interviewers' expectations concerning the ease or difficulty of interviewing respondents, might affect response rates. This study did not examine the interaction of interviewer characteristics with incentives to see whether they are particularly effective among certain group of interviewers.

Stratford et al. (2002) designed an experiment on the National Travel Survey aimed at testing the effect of offering monetary incentives to every household member conditional on full cooperation from the whole household. In addition, they investigated interviewers' attitudes towards incentives. The sample members were assigned to two experimental groups that were promised £5 and £10 respectively, conditional on full cooperation from the whole household, and a control group that received no incentive. To reduce any interviewers' confounding effects each interviewer was assigned addresses in the control and both the experimental groups. The final response rate for the control group was 62%, and the experimental groups for £5 and £10 each had 66% and 67% respectively. They found that interviewers might have put less effort into persuading reluctant respondents in the control group resulting in the significant differences in final response rates. Some interviewers also felt that offering £10 was more successful in encouraging a full response than £5. However, this positive expectation by interviewers was not supported by the response rates obtained because the difference was only one percentage point between the £5 and £10 incentive groups. In conclusion, interviewers may have certain expectations among those households offered incentives that may in turn influence their efforts and behaviour towards such households in trying to motivate them to participate in surveys.

This study focuses on testing whether interviewers differentially affect the effectiveness of incentives using a multilevel modelling approach. In addition, the study tests whether interviewer observable characteristics such as age, sex, and experience are associated with this effect. While existing studies have focused only on interviewer effects on incentives that are brought about through their influence on overall response rates, here the interactions between interviewer characteristics and incentives are also investigated. This study uses rich data from three datasets that enables the comparison of results across different survey settings.

3.3 Data

To ensure that conclusions are robust, three different face-to-face surveys with similar but somewhat different designs are used. The three studies considered are: the National Survey for Wales – Field Test 2015 (NSW 2015), the National Survey for Wales-Incentive Experiment 2016 (NSW 2016), and the UK Household Longitudinal Study Innovation Panel (Wave 1) (UKHLS-IP). Incentives in all three surveys were offered conditional upon the completion of the questionnaire and random allocation of addresses in these experimental conditions, was implemented within interviewer workloads. Only response outcomes, before any re-issuing of the questionnaires are used, in order to ensure that the random assignment of incentives within interviewers is maintained.

3.3.1 National Survey for Wales (NSW) Field Test 2015

The National Survey for Wales involved interviewing a randomly selected sample of people aged 16 and over across Wales. The Welsh government commissioned Kantar Public (previously TNS-BMRB) and Beaufort Research to carry out the NSW 2015, a large-scale field test between May and September 2015 (Hanson, Sullivan, & Mcgowan, 2016). The sample design of the NSW 2015 was based on a stratified, single-stage random selection of addresses across Wales drawn from the small user Postcode Address File (PAF), belonging to the Royal Mail. Further details on the NSW 2015 sample design are included in the technical report by Hanson, Sullivan, and Mcgowan (2015). The survey questionnaire and all supporting materials were available as standard in both English and Welsh. Adults aged 16 or over within each sampled household were interviewed face-to-face and each interview lasted for around 25 minutes. Where a household contained more than one adult, a single adult was randomly selected to represent others in the household.

The aim of the incentive experiment was to assess the extent to which response rates improved by offering respondents a £10 gift-card upon completing an interview. The experimental group (N=2,965) received a £10 conditional incentive and the second group received no incentive (N=2,830). The households which were randomly selected to be offered a conditional £10 received advance letters mentioning the incentive, while the other half of households received advance letters that contained no information about incentive. To ensure that any differences in response rates between respondents who were offered £10 and those offered no incentive are not attributed to any interviewer abilities, addresses that were offered incentives were randomly allocated within each assignment. Interviewers were required to mention the incentive on doorsteps to those households that had been offered with the aim of encouraging participation. The household level variables available in this dataset include incentive and urban/rural indicator. The socio-demographic characteristics available for interviewers include age, experience and gender. To protect the identity of interviewers, the National Survey for Wales did not provide either primary sampling units (PSU's) or middle layer super output area (MSOA) identifiers. The survey was implemented by a team of 86 interviewers with the number of households interviewed by each interviewer ranging from14 to134. Further details on the NSW2015 sample design can be found in Hanson, Sullivan, & Mcgowan (2015).

3.3.2 The National Survey for Wales Incentive Experiment 2016 (NSW 2016)

The Welsh Government commissioned the Office of National Statistics (ONS) to conduct the National Survey for Wales - incentive experiment 2016 between July and October 2016 (Aumeyr et al., 2017). The aim of NSW 2016 experiment was to find whether incentives should be introduced as the standard on its survey, and if so the size of such an incentive. The annual sample for NSW follows a design that is a stratified, single-stage random selection of addresses across Wales and is representative of all adults aged 16 or over living in private households in Wales. The sample was drawn from the Royal Mail Small Users Postcode Address File (PAF). The stratification was by Local Authority (LA) using an allocation designed to ensure that a minimum effective sample size was achieved in each LA based on estimated response rate. Further details on the sample design may be obtained from NSW 2016-17 Technical Report (Aumeyr et al., 2017).

The experiment design followed a standard ONS design principle whereby half of the addresses in each odd numbered quota[6] were offered a £5 incentive conditional on participation (N=3,604) and addresses with an even quota number were offered no incentive (N=3,467). This was to ensure that the experimental conditions were not confounded by interviewers' characteristics and geographical areas. The incentive experiment ran from July to October 2016. Originally, it was intended to run the experiment until December 2016, but it was terminated at the end of October 2016 as both experimental and control groups experienced lower response rates at 55% and 54% respectively which were lower than expected. With the aim of boosting response rates, a new £10 incentive conditional on participation was introduced to the full sample from November 2016. This current study will only consider the experiment sample size from July to October 2016 that consist of 7,071 households across the two conditions. It is crucial to note that the expected response rate for NSW 2016 was based on the NSW 2015 field test that found statistically significant increase on response rate by over 4 percentage points by offering a conditional £10 (Aumeyr et al., 2017). Therefore, the aim NSW 2016 incentive experiment was to investigate the effect of a conditional £5 incentive on response rates compared to a conditional £10 incentive given in NSW 2015.

The household characteristics variables provided in this dataset include incentive, and population density of the area as well as interviewer characteristics that include age, gender and experience. To make sure that interviewers are not identifiable, ONS provided only rural/urban identifiers and regional indicators for the purpose of analysis. During this study, 85 interviewers from ONS were involved. Socio-demographic characteristics of 10 (12%) interviewers who conducted interviews on 249 (3.5%) households were missing because they had not given consent for the use of their personal data and had already left the organisation by the time the data were released for this study. The final analysis sample had 6,122 households after excluding 742 (10.5%) ineligible households and those interviewed

[6] Each quota contained between 20 and 30 addresses on average.

by interviewers with missing socio-demographic characteristics. A sensitivity analysis of NSW 2016 including the 12% of interviewers with missing data, showed no substantive differences from the main results of the complete data.

3.3.3 UK Household Longitudinal Survey Innovation Panel Wave 1 (UKHLS-IP)

The UKHLS-IP is part of the Understanding Society survey. The main purpose of the Innovation Panel is to conduct methodological experiments, and testing aimed at of advancing knowledge in the methodology of designing longitudinal surveys (Baghal et al., 2016; Boreham & Constantine, 2008). The sample for the IP wave 1 consists of 2,786 addresses from 120 primary sampling units (PSUs) across Great Britain (Boreham and Constantine 2008). The sample design is based on equal probability and was drawn from the small user Postcode Address File (PAF), a list of addresses that receive fewer than 25 items of mail per day (Boreham & Constantine, 2008).

The experimentation in the IP Wave 1 contained four randomised split-ballot experiments designed to evaluate the use of incentives and variation in question design protocols (Boreham & Constantine, 2008). The value of incentive offered in order to achieve the required response rates was determined using a randomised three-way split sample design. The gross sample of the IP data was randomly allocated to three experimental groups, with each group receiving a different incentive condition: Group 1 £5 per adult, Group 2 £10 per adult, and Group 3 £5 per adult, rising to £10 per adult if all adults in the household completed their CAPI interviews in person. For the purposes of this current analysis, Groups 2 and 3 were combined because each household received a total of £10 and the response rates achieved in each of the two groups was the same at 61% (Boreham & Constantine, 2008). Single person households, randomly assigned to the third group, received £5 initially which then increased to £10 if they participated. Each household received an unconditional cash voucher of the appropriate amount (£5 for groups 1 and 3 and £10 for group 2) in advance, along with a letter explaining that all household members would be sampled to participate in the Understanding Society survey. The advance letter also explained the amount of incentive that households would receive after participating in the survey. In each group, a £3 incentive was also offered for each young person (ages of 10 to 15 years) who filled in a self-completion questionnaire. The vouchers with the exception of the voucher they had already received with advance letter, were sent to respondents after the interview, together with a thank you letter (Boreham & Constantine, 2008).

For UKHLS-IP, the total number of issued households were 2,786, of which 263 (9.4%) were not eligible. An additional 26 households were also issued making the final analysis sample upto 2,523 households. For the issued household sample, variables selected for the final analysis sample were restricted to those containing information from responding and nonresponding households. There were 27 (1.0%) households in the UKHLS-IP that did not successfully merge with interviewer data due to lack of common unique identifiers. The Innovation Panel data with interviewer characteristics was then linked to aggregated census variables (i.e. factor scores) from the 2011 census. A total of 21 census count variables were combined using a factorial ecology model (Rees, 1971), with a total of five neighbourhood indices extracted. Factorial ecology model uses factor analysis to analyse social aspects by treating an outcome as an interaction of many factors arising at individual, community and societal levels (Janson, 2003; Rees, 1971). This means that a factorial ecology model can describe a set of socioecological macro-units by means of a set of variables which are analysed using factor analysis and the resulting factor scores are clustered into homogenous categories (Janson, 2003). The measures considered in this study cover the extent of *concentrated disadvantage* (areas with a higher number of single parent families, those on income support and unemployed, fewer people in managerial and professional occupations, and fewer owner occupiers), *urbanicity* (high population density and domestic properties, and relatively little green space) and *population mobility* (higher levels of in- and out-migration and more single person households). The other variables account for differences in the neighborhood *age structure* (with higher scores for areas with a younger population), and *housing structure* (higher scores for areas with more terraced and vacant properties), and the police recorded crime rate.

The aggregated census variables file obtained from data provider was restricted to MSOA for England only. This led to the exclusion of 342 (12.3%) households contained in 57 MSOA's from Wales and Scotland. In addition, 31 (1.1%) of households in 5 MSOA's from England did not successfully merge with Innovation Panel data due to lack of common unique identification codes. Therefore, the final analysis sample contained 2,123 households after excluding 263 (9.4%) ineligible households. The number of interviewers involved in this study was 107 and the households interviewed by each interviewer ranged from 2 to50. However, detailed first issue outcomes were not available for the UKHLS-IP so in this study it is only possible to model response/nonresponse for this survey, rather than cooperation conditional on contact. Analysis of the two Welsh surveys show that the results are substantively the same for both response and cooperation, so this is not considered to be an important limitation. Further details about the UKHLS-IP can be found in Boreham and Constantine (2008).

3.3.4 Analytical Approach

To test for the effect of interviewers on the effectiveness of incentives the experimental variations in the incentives offered are examined by considering response and cooperation rates. Response rate is a function of all various nonresponse sources such as non-contacts, refusals, and other unproductive responses while cooperation rate is a function of refusals only. Different dynamics lead to noncontacts and refusals in face-to-face surveys (Groves & Couper, 1998). Noncontact is related to accessibility impediments such as locked gates, no-trespassing signs, and intercoms. Refusals occur only after contact is made and the decision to participate or not is influenced by the respondent's openness to a survey and also by other factors such as incentives offered, interviewer behaviour, sponsor and topic of survey (Durrant et al., 2010; Groves & Couper, 1998). Therefore, by considering both response and cooperation rate this study considers the possible counteracting biases of different types of nonresponse. The first definition of survey response is based on AAPOR RR2[7]((APPOR), 2016; Lavrakas, 2008):

$$RR = \frac{(I + P)}{(I + P) + (R + NC + O) + (UE(NC) + UE)} \qquad (3.1)$$

where RR denotes Response Rate, I Interview, P Partial Interviews, R Refusals, NC Non-Contacts, O Other Unproductive, UE(NC) Unknown Eligibility (non-contacted), and UE Unknown Eligibility. Generally, the response rate is the ratio of all households interviewed out of all eligible samples units in the study. The cooperation rate (CR) depends on those contacted and is defined as:

$$CR = \frac{(I + P)}{(I + P + R)} \qquad (3.2)$$

The distributions of the response outcomes for the three datasets are presented in Table 3-1. A 2 by 2 chi-square test was used to test the association between response and incentive condition in each of the three surveys. The null hypothesis H_0 assumes that there is no association between response and incentive condition, while the alternative hypothesis H_a assumes some association does exist. The response rates for those offered an incentive for NSW 2015 and NSW 2016 are 53% and 55% compared to only 50%, and 54% for those not offered an incentive respectively. For the Innovation Panel Wave 1, the response rates for those offered £10 are 61% compared to only 56% for those offered £5. The $P - values$ are statistically

[7] The disposition code in response rate described above is:
AAPOR = American Association for Public Opinion Research

significant at the 95% level of confidence for both NSW 2015 and UKHLS-IP indicating that the higher response rates on incentive groups are not due to random variation. However, chi-square test for NSW 2016 is not significant.

Table 3-1: Incentives and fieldwork outcomes for the three surveys

	NSW2015		**NSW2016**		**UKHLS-IP**	
	£10	**£0**	**£5**	**£0**	**£10**	**£5**
Interviews	1,387	1,228	1,772	1,664	1,020	469
Refusals	640	670	954	961	-	-
Non-contact	285	289	265	250	-	-
Other nonresponse	285	273	230	233	-	-
Total nonresponse	1210	1232	1,449	1,444	660	374
Ineligible	368	370	383	359	175	88
Cooperation Rate	68%	65%	65%	63%	-	-
Response Rate	53%	50%	55%	54%	61%	56%
Total issued sample	2,965	2,830	3,604	3,467	1,855	931

Note: Only total nonresponse is available for UKHLS-IP at first issue

3.4 Methodology

The influence of interviewers on the effectiveness of incentives on survey response is assessed using multilevel logit models. In survey nonresponse, multilevel models have been widely applied to examine interviewer effects (Durrant & Steele, 2009; Hox & de Leeuw, 2002; Vassallo et al., 2015). The model applied has the following form. Let y_{ij} denote the binary response for household i $(i = 1, \dots, i)$, interviewed by interviewer j $(j = 1, \dots, j)$ where

$$y_{ij} = \begin{cases} 1 & \text{cooperation /response} \\ 0 & \text{refusal} \end{cases} \qquad (3.3)$$

where y_{ij} is assumed to follow a Bernoulli distribution, , with conditional response probabilities defined as $\pi_{ij}\ \Pr(y_{ij} = 1)$ and $1 - \pi_{ij} = \Pr(y_{ij} = 0)$. The multilevel logistic regression model accounting for interviewer effects takes the form

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_{1j}x_{1ij} + \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{j}\boldsymbol{\alpha} + \mu_{1j}x_{1ij} + \mu_{oj} \qquad (3.4)$$

where β_0 is the intercept , $\beta_1$is the coefficient for the incentive condition, x_{1ij} is a dummy indicator of the incentive group for household *i* within the assignment of interviewer *j*, $\mathbf{x}'_{ij}$ is a vector of household-level characteristics with coefficient vector $\boldsymbol{\beta}$, $\mathbf{z}'_j$ is a vector of interviewer-level covariates with coefficient vector $\boldsymbol{\alpha}$, μ_{0j} is a random intercept and μ_{1j} is a random coefficient for the incentive variable. The random intercept and slope, μ_{0j} and μ_{1j}, are assumed to follow a normal distribution with zero mean and variance matrix Ω_μ defined as

$$\begin{bmatrix} \mu_{0j} \\ \mu_{1j} \end{bmatrix} \sim N\left(0, \Omega_\mu\right) \text{ where } \Omega_\mu = \begin{bmatrix} \sigma_{\mu 0}^2 & \\ \sigma_{\mu 01} & \sigma_{u1}^2 \end{bmatrix} \tag{3.5}$$

where $\sigma_{\mu 0}^2$ is the intercept variance, $\sigma_{\mu 1}^2$ is the variance in slope and $\sigma_{\mu 01}$ is the covariance between random intercept and coefficients residuals. Positive values of $\sigma_{\mu 01}$ indicate that interviewers who achieve high response rates on average (i.e. intercept) are the interviewers where the effect of the incentive is greater, and negative values indicate the opposite. Cross-level interactions between interviewer characteristic variables and the incentive variable are included in Equation (3.4) to test whether observable characteristics of interviewers are associated with variability in the effectiveness of deploying incentives. Quantification of the random slope variance (i.e. incentive effect variance) will be evaluated by providing a range around the fixed effect (i.e. incentive effect) within a 95% confidence interval (CI) based on the random effect variance (i.e. variance in slope) (Lorah, 2018; Snijders & Bosker, 2012). This is defined as:

$$\text{Random Effect 95\% CI} = \text{fixed effect} \pm \left(1.96 \times \sqrt{\text{random variance}}\right) \tag{3.6}$$

The analysis of interviewer effects can be complicated by the confounding of interviewer assignments and area. This happens because interviewer assignments may be clustered within particular geographic areas making it difficult to distinguish the effects of interviewers on survey outcomes from area compositional effects (Campanelli & O'Muircheartaigh, 1999; Durrant et al., 2010). Failure to account for differences in the area-level composition of interviewer assignments can result in over-estimation of the magnitude of interviewer effects (O'Muircheartaigh & Campanelli, 1998). In face-to-face surveys, it is difficult and costly to have a fully interpenetrated survey design that randomly assigns interviewers to households. Where there is an overlap between interviewer assignments and areas, this can be mitigated using a cross-classified multi-level model (Durrant & Steele, 2009). However, this could not be done for the three datasets analysed here, because it was not possible to obtain geographic identifiers for the two Welsh surveys and the UKHLS-IP did not contain sufficient crossing of interviewers and areas to implement a cross-classified model. Therefore, any potential area effects on survey response were controlled for in models by using area level characteristics as fixed effects to assess their impact on the interviewer random effects. This is because survey nonresponse is influenced by area effects such as similarities in socio-economic characteristics, accessibility, and urbanicity across geographic regions of the sampled units

(Haunberger, 2010). Therefore, controlling for area variables can explain some variability in household responses although it is difficult to quantify to which extent. It is also important to note that area variables that are not significant were excluded from the final model on the assumption that area effects are absent in household response.

3.4.1 Modelling Estimation and modelling strategy

Models are estimated using the Markov Chain Monte Carlo (MCMC) methods implemented in MLwiN software (Browne et al., 2016; Fearn et al., 2004). MCMC estimation allows fitting of Bayesian models, by specifying the prior distributions for the model parameters. The decision to use Bayesian approach is informed by the fact that it allows estimation of robust variance estimates when the number of higher-level units are small and the data are imbalanced (i.e. number of units per interviewer is not equally distributed) (Gelman & Hill, 2007). The starting values for the fixed effects are the second-order penalised quasi-likelihood (PQL) estimates. Priors for the variance matrix are assumed to follow an inverse Wishart distribution $p(\Omega_{\mu}^{-1}) \sim Wishart_n(n)$, where n is the number of rows in the variance matrix and is an estimate for the true value of the variance matrix Ω_{μ} (Browne et al., 2016). The starting values for variance parameters (i.e. $\sigma_{\mu 0}^2$ and $\sigma_{\mu 1}^2$) are 0.1 and 0 for covariance (i.e. $\sigma_{\mu 01}$). A forward selection strategy is used for selecting the variables to include in the final model (Hosmer & Lemeshow, 2000). The first step is the specification of the base model that includes only incentives as the fixed effect.

The second step of modelling involves the inclusion of random intercept and random slope across incentive variable one at a time. Then, explanatory variables are added to obtain the final model. In addition, changes in the parameter estimates on the random part will be tracked as the model becomes more complicated. A Deviance information Criterion (DIC) is used to evaluate whether the added random effects are leading to a better model fit (Spiegelhalter, Best, Carlin, & van der Linde, 2002). The DIC is a Bayesian measure of model fit that penalises for model complexity which enables nested model comparisons, with smaller DIC values indicating a better fit. That is, DIC is the sum of the posterior expectation (mean) of the deviance function ($\overline{D}$) and the effective number of parameters (pD). The term ($\overline{D}$) measures the goodness-of-fit of the model and the term (pD) measures the model complexity. When comparing DIC values, a model with a DIC value of at least 3 points lower than the previous model is considered to have a significantly better fit (Rasbash, Steele, Browne, Goldstein, & Charlton, 2012; Spiegelhalter et al., 2002). For discrete response models, the Wald test is usually also an alternative to test significance for the variance parameters. However, the Wald test has an approximate chi-squared distribution and therefore is not appropriate for testing significance of variance parameters because variance parameters are not normally distributed (Welham, Gezan, Clark, & Mead, 2014). The Wald test tends to have a large positive value because the ratio obtained after dividing the variance estimate by its standard error estimate tends to have a large positive value with respect to the variance and covariance matrix. Naturally, variances can only take positive values and therefore Wald's test for variance parameters tends to be a one-sided test. However, Wald test will be used evaluate the significance of the covariance between the random intercept and random coefficient. The covariance value is significant if the ratio obtained after dividing the covariance estimate by its standard error is greater than 2.

The models fitted in this study had a burn-in length of 10,000 and then 200,000 iterations. In order to avoid undue influence of starting values, different burn-in lengths were tried as recommended by Fearn et al. (2004). The Brooks-Draper and Raftery-Lewis diagnostic were checked to determine how long the chain must be run for, to obtain accurate posterior estimates (Browne et al., 2016). Table 3-2 presents the different specifications of multilevel models fitted for each data set.

Table 3-2: Specifications of the models fitted for each survey

Model	**Fixed and random components specified**
1: model 1(Base)	Incentive
2: model 1 + area level variables	Model 1 + area level variables
3: model 2 + random intercept (interviewers)	Model 2 + significant area level variables from model 2 + random intercept across interviewers
4: model 3 + random coefficient (interviewers)	Model 3 + random coefficient for incentives across interviewers
5: model 4 + interviewer characteristics	Model 4 + interviewer characteristics
6: model 5 + cross-level interactions	Model 5 + cross-level interactions for incentive and interviewer characteristics

3.5 Results

3.5.1 National Survey for Wales Field Test 2015

Table 3-3 presents variance estimates and DIC values for various specifications of the NSW2015 models. The inclusion of population density of area variable in models 2 improves the model fit significantly since the DIC values for response and cooperation reduce by 9 and 3 respectively compared to incentive only model 1. The random intercept model in Table 3-3 with interviewer effects (model 3) serves as a benchmark with which to compare other models controlling for interviewer effects. The inclusion of a random intercept across interviewers in models 3 for both response and cooperation improves the model fit significantly in terms of DIC changes in comparison with models 2 which is consistent with findings of Durrant, Groves, and Steele (2010) and Blom, Leeuw, and Hox (2011). The test of the hypothesis that the relationship between incentives and household response and cooperation varies as a function of interviewers is investigated in model 4 by including a random slope for incentive. After controlling for interviewer variation for the regression coefficients of the household incentive in model 4, the DIC values for response and cooperation reduce by 13 and 15 respectively in comparison with random-intercept only model. This indicates that introducing a random coefficient leads to an improvement in model fit.

The inclusion of interviewer characteristics in model 5 leads to a non-significant reduction in DIC by 0.2 for response model. This implies that interviewer characteristics do not have a significant effect on model fit for household response. However, the DIC value increases by -0.01 in the cooperation model 5 indicating that interviewer characteristics do not improve model fit significantly. The DIC change in models 6 with cross-level interactions effects for both response and cooperation are not significant. This shows that interviewer characteristics do not moderate the relationship between incentives and household response and cooperation in the NSW Field Test 2015 survey.

The variance for the random intercept decreases slightly after controlling for interviewer characteristics for both response and cooperation in model 5. The variance of the random coefficient on incentive for cooperation is slightly higher than response rate indicating that interviewer effectiveness in deployment of incentives is pronounced on survey cooperation when compared with response. On the other hand, the variance of the incentive random coefficient increases slightly for both response and cooperation in model 6 after controlling for the cross-level interactions between incentive and interviewer characteristics. The covariance between the random intercept and random coefficient for both response and cooperation models are positive but non-significant. The results for quantification of the incentive effect indicates that deployment of incentives by interviewers improve both response and cooperation on average, although some interviewers may actually obtain lower response and cooperation rates in incentive group. This is because the range of slopes for models 4, 5 and 6 for both response and cooperation are predicted to range from negative to positive values. For example, model 4 for response has a range from -0.22 to 0.69 indicating that deployment of incentives may actually lead to a decline in response rates.

Table 3-3: Variance estimates and DIC values for various specifications of the models in NSW 2015 based on response (RR) and cooperation (CR)

Model	**Response Propensity models based on RR**					**Cooperation Propensity models based on CR**				
	Intercept Variance (SD)	*Coefficient Variance (SD)* *Range of slope (,)*	Covariance (SD)	DIC	*DIC Change*	*Intercept Variance (SD)*	*Coefficient Variance (SD)* *Range of slope (,)*	*Covariance (SD)*	DIC	DIC Change
	Interviewer	*Interviewer*				*Interviewer*	*Interviewer*			
1: model 1(Base)	-	-	-	7000.388	-	-	-	-	4994.822	-
2: model 1 + area level variables	-	-	-	6992.588	7.800	-	-	-	4990.156	4.666
3: model 2 + random intercept (interviewers)	0.213 (0.049)	-	-	6759.186	233.402	0.284(0.066)	-	-	4771.789	218.368
4: model 3 + random coefficient (interviewers)	0.133 (0.052)	0.054 (0.027) (-0.221, 0.689)	0.011 (0.031)	6751.928	7.258	0.255(0.071)	0.070 (0.038) (-0.314, 0.722)	0.013 (0.044)	4763.824	7.965
5: model 4 + interviewer characteristics	0.139 (0.047)	0.063 (0.031) (-0.329, 0.655)	0.027(0.028)	6751.556	0.240	0.168(0.056)	0.078 (0.042) (-0.339, 0.755)	0.038 (0.034)	4763.395	-0.079
6: model 5 + cross-level interactions	0.138 (0.045)	0.065 (0.032) (-0.151, 0.849)	0.029 (0.027)	6750.275	1.097	0.163 (0.057)	0.094 (0.052) (-0.313, 0.889)	0.034 (0.038)	4765.957	-2.562

Range of slope=Quantification of the incentive effect variance

Table 3-4 presents the estimated coefficients, their standard deviations and the corresponding 95% credible intervals for the NSW Field Test 2015 models 5 and 6. Table 3-4 shows that incentive has a positive and not significant effect on survey cooperation. The interviewer characteristics (i.e. experience, age and gender) controlled for in this model have a non-significant effect on cooperation. The random slope variance values of 0.08 and 0.09 for models 5 and 6 respectively are significant indicating that interviewers vary in the effectiveness with which they deploy incentives. The non-significant cross-level interactions in model 6 show that interviewer characteristics do not significantly moderate the relationship between incentives and survey cooperation.

The cross-level interactions between the three interviewer characteristic variables – age, sex, and experience and the incentive dummy are all non-significant, indicating that these interviewer characteristics do not explain interviewer variability in the effectiveness of incentives on cooperation. The covariance between the random intercept and random coefficient, $\sigma_{\mu01}$, is non-significant, with a posterior estimate of 0.03 indicating that the effectiveness of incentive deployment among interviewers is not related to the overall cooperation rate an interviewer achieves on their assignments. The results for the response model are substantively the same as those of the cooperation model and are presented in Appendix B.2.

Figure 3-1 plots the difference in the mean predicted rates of response (left panel) and cooperation (right panel) for each interviewer derived as fitted values from the models 6 for response and cooperation. Each green and blue dot in Figure 3-1 represents an interviewer, with the left Y axis being the difference in the response and cooperation rates for households in the incentive and non-incentive conditions respectively. The grey and brown triangles show the mean overall response and cooperation rates respectively (plotted against the right Y axis) for each interviewer across all eligible households in their assignment. It can be observed that the differences in interviewers' response and cooperation rates among those households offered an incentive and those not offered range from -13% to +14% and from -10% to +12% respectively. This indicates that interviewers' performance varies substantially for both survey cooperation and response. The interviewers who have a negative percentage difference in response and cooperation rates tend to perform worse among households offered incentives in comparison to those not offered an incentive.

Table 3-4: Estimated coefficients for models 5 and 6 for NSW Field Test 2015 Cooperation

		Model 5				*Model 6*			
Variable *{Reference Category}*	*Category*	*β*	*SD*	*95% Credible Interval*		*β*	*SD*	*95% Credible Interval*	
Intercept		0.409	0.282	-0.167	0.942	0.239	0.320	-0.409	0.857
Incentive *{no incentive}*	£10 Incentive	0.208	0.079	0.055	0.365	0.288	0.363	-0.433	0.991
Interviewer age *{young}*	Lower middle	-0.037	0.208	-0.444	0.368	0.048	0.237	-0.406	0.526
	Upper middle	0.170	0.212	-0.246	0.585	0.345	0.233	-0.103	0.808
	old	0.226	0.245	-0.253	0.706	0.428	0.278	-0.102	0.982
Interviewer Experience *{less}*	Lower middle	-0.026	0.221	-0.450	0.410	-0.027	0.255	-0.535	0.471
	Upper middle	0.364	0.247	-0.114	0.858	0.403	0.285	-0.170	0.959
	Highest	0.430	0.235	-0.021	0.898	0.386	0.274	-0.159	0.929
Interviewer Sex *{Female}*	Male	-0.094	0.136	-0.362	0.174	0.048	0.237	-0.406	0.526
Incentive *{£10 per adult}**Gender *{Female}*	£10 per adult *Male					0.035	0.180	-0.316	0.385
Incentive *{£10 per adult}* * Age *{young}*	£10* Lower middle					-0.021	0.269	-0.540	0.506
	£10* Upper Middle					-0.124	0.264	-0.638	0.394
	£10* Old					-0.412	0.317	-1.035	0.216
Incentive *{£5}* * Experience *{less}*	£10*Lower Middle					-0.027	0.281	-0.587	0.516
	£10*Upper Middle					-0.176	0.312	-0.791	0.434
	£10*Highest					0.109	0.297	-0.480	0.688
$\sigma_{\mu 0}^2 = var(\mu_{oj})$		0.168	0.056	0.076	0.294	0.199	0.065	0.097	0.349
$\sigma_{\mu 1}^2 = var(\mu_{1j})$		0.078	0.042	0.026	0.196	0.085	0.047	0.025	0.205
$\sigma_{\mu 01} = cov(\mu_{0j}, \mu_{1j})$		0.032	0.036	-0.044	0.100	0.020	0.042	-0.073	0.096
DIC		4763.395				4765.957			

It can also be observed that interviewers who are good at achieving higher responses and cooperation among households not offered incentives also tend to have slightly higher response and cooperation rates among incentivised households. However, this pattern is moderate as indicated by plots in Figure 3-1 and is consistent with positive covariance in model 6 for both response and cooperation that are not significant in Table 3-3.This indicates that interviewer's response and cooperation rates do not influence their effectiveness in the deployment of incentives. Not all of this variability is attributable to how skilful interviewers are in deploying incentives and simply reflects random variability in response propensities across interviewer assignments. A better sense of the effect of interviewers on incentive effectiveness can be achieved by taking the expected cooperation rate for an incentivised household using interviewers from the top and bottom deciles of the random coefficient variance, while holding all other variables constant. For response, this shows that interviewers in the top performing decile achieve an expected response rate of 54% for incentivised households compared to 48% for those in the bottom decile and compared to 53% for the median interviewer for non-incentivised households, a quite substantial difference. The corresponding figures for cooperation are 67% and 64% for the top and bottom deciles, respectively, and 68% for the median interviewer for non-incentivised households. There is no obvious relationship between the overall response rate and the effectiveness of the incentive within interviewers, so we find no evidence that interviewers who are, on average, better at obtaining cooperation are also more effective in deploying the incentive.

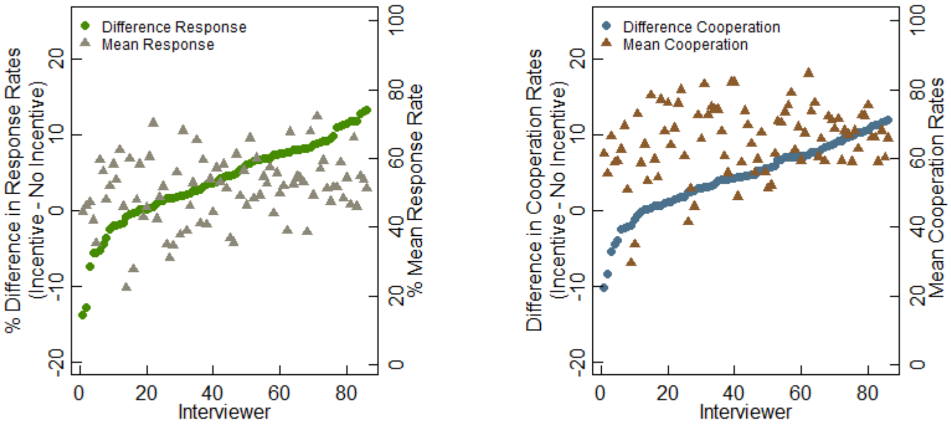


Figure 3-1: Difference in predicted rates of response (left panel) and cooperation (right panel) for incentive and non-incentive households by interviewer for NSW Field Test 2015 in model 6

3.5.2 National Survey for Wales Incentive Experiment 2016

Table 3-5 presents the results for various specifications of the NSW incentive experiment 2016. The inclusion of population density of area variables in models 2 significantly reduces the DIC values by 66 and 36 for response and cooperation respectively when compared with to models 1. This indicates that population density of area improves the model fit for survey response and cooperation. The random intercept models 3 for both response and cooperation are highly significant in terms of DIC changes when compared with models 2.

After controlling for the interviewer variation for the regression coefficients of the household incentive in models 4, the DIC values reduce by 17 and 15 for response and cooperation respectively when compared with the random-intercept only models 3. This implies that the variance of the incentive coefficient for both response and cooperation are significant and that interviewers vary in the deployment of incentives. Controlling for interviewer characteristics does not lead to significant reduction of DIC values in model 5 for either response or cooperation. The DIC values reduce by 3 and 4 after controlling for cross-level interactions effects in model 6 for both response and cooperation compared to the previous model 5. This shows that interviewer characteristics significantly moderate the relationship between incentives and household survey response and cooperation although this is at borderline.

The variances for the interviewer random intercept increase slightly for response after controlling for area level and interviewer characteristics variables. However, the variance for the random intercept for cooperation reduces after controlling for area characteristics. There is a change of variance after controlling for interviewer characteristics on survey cooperation. There is a slight reduction in variance for the incentive coefficient after controlling for cross-level interactions between incentives and interviewer characteristics for both survey response and cooperation. It is important to note that these changes in variances are quite small indicating that inclusion of interviewer characteristics and corresponding cross-level interactions do not explain interviewer variability in the effectiveness of incentives. The variance for the random coefficient on incentive for cooperation is slightly higher than the response rate, indicating that interviewer effectiveness in the deployment of incentives is more pronounced on survey cooperation than response. The negative covariance values for response and cooperation are not significant. The results for quantification of the incentive effect indicates are also similar to those obtained for NSW 2015 and show that deployment of incentives by interviewers improve both response and cooperation on average, although some interviewers may actually obtain lower response and cooperation rates in incentive group.

Table 3-5: Variance estimates and DIC values for various specifications of the models in NSW 2016 based on response (RR) and cooperation (CR)

Model	**Response Propensity models based on RR**					**Cooperation Propensity models based on CR**				
	Intercept Variance (SD)	*Coefficient Variance (SD)* *Range of slope (,)*	*Covariance (SD)*	DIC	*DIC Change*	*Intercept Variance (SD)*	*Coefficient Variance (SD)* *Range of slope (,)*	*Covariance (SD)*	DIC	DIC Change
	Interviewer	*Interviewer*				*Interviewer*	*Interviewer*			
1: model 1(Base)	-	-	-	8443.080	-	-	-	-	6753.065	-
2: model 1 + area level variables	-	-	-	8377.019	66.061	-	-	-	6716.787	36.278
3: model 2 + random intercept (interviewers)	0.148(0.036)	-	-	8178.454	198.595	0.136(0.036)	-	-	6571.087	145.700
4: model 3 + random coefficient (interviewers)	0.130 (0.038)	0.069 (0.033) (-0.437, 0.592)	0.003 (0.027)	8160.613	17.841	0.128 (0.042)	0.074 (0.038) (-0.457,0.609)	-0.008 (0.031)	6556.021	15.066
5: model 4 + interviewer characteristics	0.128 (0.042)	0.068 (0.032) (-0.432, 0.590)	-0.005 (0.030)	8161.125	0.281	0.122 (0.044)	0.075 (0.039) (-0.454,0.619)	-0.021 (0.034)	6555.917	0.104
6: model 5 + cross-level interactions	0.132 (0.041)	0.062 (0.030) (-0.322, 0.654)	-0.007 (0.028)	8156.666	3.340	0.122 (0.043)	0.067 (0.035) (-0.385,0.629)	-0.021 (0.032)	6551.744	4.062

Range of slope=Quantification of the incentive effect variance

Table 3-6 presents the estimated coefficients, their standard deviations and the corresponding 95% credible intervals for models 5 and 6 obtained using NSW Incentive Experiment 2016 cooperation. It can be observed that incentive has a positive and non-significant effect on cooperation. The population density variable indicates that households living in towns and urban areas have a significant negative effect on cooperation consistent with findings by Groves and Couper (1998). None of the interviewer characterstics is significant. Also, the cross-level interactions show that interviewer characteristics do not significantly moderate the relationship between incentives and survey cooperation. This finding is consistent with NSW Field Test 2015 and shows that interviewer characteristics do not significantly moderate the relationship between incentives and survey cooperation. The covariance value for cooperation is negative and not significant. The fact that inclusion of cross-level interactions in model 6 leads to a significant improvement of model fit in Table 3-5 although they are non-significant indicates that statistical non-significance does not imply an effect is improbable (Wasserstein, Schirm, & Lazar, 2019).

From Figure 3-2, it can be observed that effectiveness in deployment of incentives varies across interviewers with the range of percentage differences in response and cooperation probabilities lying between -8% and 17%. However, it is also the case that the relationship between mean differences of both survey response and cooperation and mean response and cooperation rates for interviewer is not well evident. This finding explains the negative covariance in the random coefficient model. In conclusion, interviewers vary in their effectiveness of deploying incentives with some even performing worse among incentivised households.

The effect of interviewers on incentive effectiveness for response shows that interviewers in the top performing decile achieve an expected response rate of 59% for incentivised households compared to 49% for those in the bottom decile and compared to 55% for the median interviewer for non-incentivised households. The corresponding figures for cooperation are 65% and 45% for the top and bottom deciles, respectively, and 58% for the median interviewer for non-incentivised households. This indicates a substantial difference in the effect of interviewers on incentive effectiveness for both response and cooperation. However, there is no obvious relationship between the overall response and cooperation rates and the effectiveness of the incentive within interviewers, so we find no evidence that interviewers who are, on average, better at obtaining cooperation are also more effective in deploying the incentive. These findings are consistent with NSW Field Test 2015.

Table 3-6: Estimated coefficients for the models 5 and 6 for NSW Incentive Experiment 2015 Cooperation

Variable *{Reference Category}*	*Category*	*Model 5*				*Model 6*			
		β	*SD*	*95 % Credible Intervals*		*β*	*SD*	*95 % Credible Intervals*	
Intercept		0.928	0.123	0.684	1.165	0.927	0.132	0.674	1.193
Incentive *{no incentive}*	£5 Incentive	0.083	0.069	-0.053	0.219	0.122	0.111	-0.099	0.341
Population density of area {*Village*}	Hamlet and isolated dwellings	0.143	0.142	-0.134	0.421	-0.144	0.143	-0.134	0.424
	Town and Fringe	-0.359	0.114	-0.581	-0.134	-0.353	0.116	-0.582	-0.126
	Urban	-0.328	0.106	-0.534	-0.121	-0.326	0.106	-0.532	-0.121
Interviewer age {*young*}	Upper Middle	-0.260	0.108	-0.472	-0.049	0.136	0.152	-0.163	0.435
Interviewer Experience {*less*}	Upper middle	-0.066	0.171	-0.403	0.272	-0.275	0.195	-0.657	0.105
	Highest	-0.302	0.186	-0.672	0.064	-0.303	0.221	-0.735	0.127
Interviewer Sex *{Female}*	Male	0.147	0.130	-0.107	0.402	-0.159	0.131	-0.414	0.106
Incentive *{£10 per adult}**Gender *{Female}*	£10 per adult *Male					-0.270	0.148	-0.560	0.025
Incentive *{£10 per adult}* * Age *{young}*	£10* Upper Middle					0.020	0.246	-0.285	0.389
Incentive *{£5}* * Experience *{less}*	£10*Upper Middle					0.052	0.221	-0.553	0.987
	£10*Highest					0.020	0.246	-0.464	0.499
$\sigma_{\mu 0}^2 = var(\mu_{oj})$		0.122	0.044	0.055	0.224	0.120	0.043	0.055	0.220
$\sigma_{\mu 1}^2 = var(\mu_{1j})$		0.075	0.039	0.023	0.172	0.067	0.036	0.023	0.154
$\sigma_{\mu 01} = cov(\mu_{0j}, \mu_{1j})$		-0.021	0.034	-0.103	0.032	-0.021	0.032	-0.096	0.031
DIC		6555.917				6551.774			

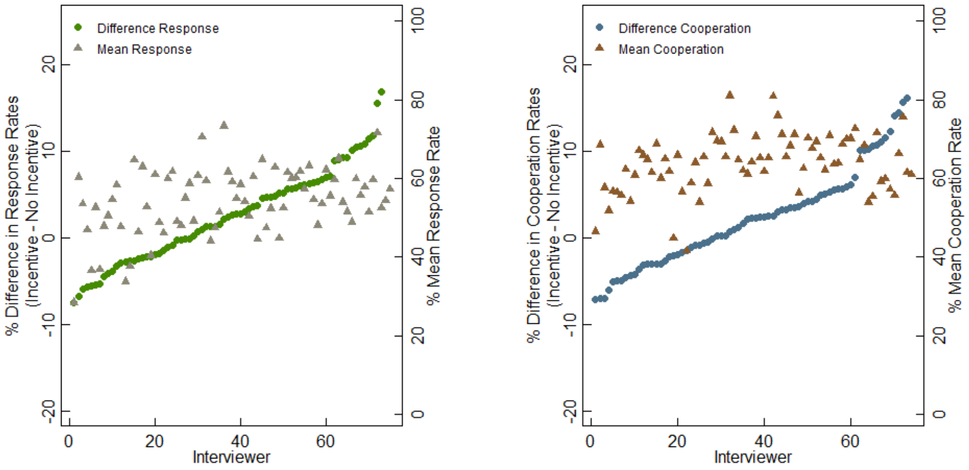


Figure 3-2: Difference in predicted rates of response (left panel) and cooperation (right panel) for incentive and non-incentive households by interviewer for NSW Incentive Experiment 2016 in final model 6

3.5.3 Innovation Panel Wave 1

Table 3-7 shows the variance estimates and DIC values for multilevel models fitted for the Innovation Panel data, which as a household longitudinal survey, has a different design from the Welsh cross-sectional study. Here, in this study the focus is only on wave 1 and the response outcome, because the original outcomes before re-issuing were not available. The DIC values for model 2 reduce by 25 after controlling for neighbourhood characteristics when compared with model 1. After controlling for the interviewer variation for the regression coefficients of the household incentive in model 4, the DIC value reduces by 9 when compared with the random-intercept only, model 3. This provides evidence of a between interviewer difference in the effectiveness of the incentive. These findings indicate that interviewers vary in how effective they are at deploying incentives and this is consistent with both the NSW field test 2015 and the NSW incentive experiment 2016. The DIC value in model 5 after controlling for interviewer characteristics has a slight increase indicating that controlling for interviewer characteristics variables does not improve the model fit. The DIC value for model 6 for cooperation increases by 0.5 after controlling for cross-level interactions indicating that interviewer characteristics do not moderate the relationship between incentives and household response. This finding is consistent with the NSW 2015 field test.

The variance for the random intercept reduces after controlling for interviewer characteristics for response in model 5. The variance for the incentive random coefficient reduces for response in model 6 after controlling for the cross-level interactions between incentive and interviewer characteristics. The covariance value for cooperation is positive and not significant which is consistent with the results obtained for the NSW 2015 field test. This indicates that there is no support from UKHLS-IP for the idea that interviewers who, on average, obtain higher response rates might also be more effective in their deployment of incentives. The results for quantification of the incentive effect indicates are similar to those obtained for NSW 2015 and NSW 2016 which show that deployment of incentives by interviewers improve both response and cooperation on average, although some interviewers may actually obtain lower response and cooperation rates in incentive group.

Table 3-7: Variance estimates and DIC values for various specifications of the models in Innovation Panel based on both response (RR)

	Intercept Variance (SD)	Coefficient Variance (SD) *Range of slope (,)*	Covariance (SD)	DIC	DIC Change
	Interviewer	*Interviewer*			
1: Base	-	-	-	2869.213	-
2: model 1 + neighbourhood characteristics	-	-	-	2844.657	24.556
3: model 2 + random intercept (interviewers)	0.759 (0.177)	-	-	2591.580	253.077
4: model 3 + random coefficient (interviewers)	0.492 (0.182)	0.178 (0.104) (-0.444, 1.210)	0.134 (0.098)	2582.256	9.324
5: model 4 + interviewer characteristics	0.442 (0.191)	0.131 (0.089) (-0.460, 0.958)	0.131 (0.089)	2582.650	-0.394
6: model 5 + cross-level interactions	0.438 (0.186)	0.130 (0.090) (-0.489,0.924)	0.001 (0.111)	2583.195	-0.545

Range of slope=Quantification of the incentive effect variance

Table 3-8 presents the estimated coefficients and corresponding credible intervals for the standard multilevel models 5 and 6 for the Innovation Panel Wave 1 data. This is despite model 4 being the most parsimonious though not significantly different. This was necessitated by the need to provide comparisons with the results obtained for NSW 2015 and NSW 2016. The results are consistent with those for the NSW 2015 and the NSW 2016; the fixed effect for the incentive predicting response is positive but non-significant and the interviewer characteristics - age, gender, and experience - are all non-significant, as are the interactions between these variables and the incentive fixed effect.

Table 3-8: Estimated coefficients for models 5 and 6 for Innovation Panel Response

Variable *{reference category}*	*Category*	*β*	*SD*	*95% Credible Interval*		*β*	*SD*	*95% Credible Interval*	
Intercept		0.344	0.413	-0.445	1.218	0.383	0.647	-0.893	1.679
Incentive *{£5 per adult}*	£10 per adult	0.249	0.116	0.019	0.482	0.217	0.549	-0.471	1.767
Urbanicity		-0.175	0.072	-0.318	-0.033	-0.213	0.078	-0.368	-0.061
Housing structure		0.232	0.070	0.096	0.372	0.302	0.077	0.153	0.455
Population Mobility		-0.190	0.082	-0.351	-0.031	-0.269	0.094	-0.455	-0.089
Gender *{Female}*	Male	-0.229	0.203	-0.627	0.165	-0.477	0.304	-1.089	0.117
Age *{less than 40 years}*	41 to 50 years	-0.379	0.477	-1.342	0.546	0.550	0.654	-0.764	1.852
	50 to 60 years	0.473	0.427	-0.377	1.297	0.730	0.512	-0.246	1.754
	> 60 years	0.359	0.444	-0.510	1.215	0.251	0.675	-1.103	1.567
Experience *{less than 2 yrs.)*	3 to 6 years	-0.203	0.230	-0.658	0.260	-0.355	0.326	-1.010	0.258
	7 to 9 years	-0.172	0.293	-0.751	0.400	-0.658	0.413	-1.479	0.151
	>10 years	-0.368	0.394	-1.143	0.409	-0.869	0.570	-1.998	0.220
Incentive *{£5 per adult}**Gender *{Female}*	£10 per adult *Male					0.140	0.143	-0.345	0.624
Incentive *{£5 per adult}* * Age *{less than 40 years}*	£10 per adult *41 to 50 years					-0.024	0.611	-1.250	1.147
	£10 per adult *50 to 60 years					-0.110	0.551	-1.194	0.951
	£10 per adult *> 60 years					0.077	0.567	-1.040	1.165
Incentive *{£5 per adult}* * Experience *{less than 2 yrs.)*	£10 per adult *3 to 6 years					0.035	0.266	-0.496	0.552
	£10 per adult *7 to 9 years					0.366	0.327	-0.273	1.010
	£10 per adult *>10 years					0.198	0.428	-0.642	1.045
$\sigma_{\mu0}^2 = var(\mu_{oj})$		0.442	0.191	0.206	0.908	1.143	0.359	0.582	1.984
$\sigma_{\mu1}^2 = var(\mu_{1j})$		0.131	0.089	0.047	0.423	0.126	0.085	0.028	0.349
$\sigma_{\mu01} = cov(\mu_{oj}, \mu_{1j})$		0.131	0.089	-0.068	0.315	0.130	0.180	-0.274	0.349
DIC		2582.650				2583.195			

Three of the area level variables are significantly associated with response; the higher the urbanicity and population mobility, the lower the level of survey response, while areas with a housing structure comprising more terraced housing and vacant properties have higher levels of response. Even after controlling for these differences in area composition, the random coefficient for the incentive is significant, with a variance of 0.13. This suggests that the between interviewer variability in the effectiveness of the incentive is caused by interviewer behaviour, rather than by the differences of the people allocated to interviews supporting earlier findings of significant change of DIC with inclusion of the random coefficient. The cross-level interactions show that interviewer characteristics do not significantly moderate the relationship between incentives and survey cooperation. These findings are consistent with those obtained for the NSW 2015 and the NSW 2016.

Figure 3-3 plots the difference in the mean predicted rates of response for each interviewer 5derived as fitted values from the model in Table 3-8. It shows a very similar pattern to what was seen in Figures 3-1 and 3-2 for the Welsh surveys, with substantial between-interviewer variation in response probabilities between high and low incentive groups with a range of -21% to +18%. Visually, there is no evidence of a positive correlation between percentage difference in response rates and the overall response rate for each interviewer, although this difference is not statistically significant.

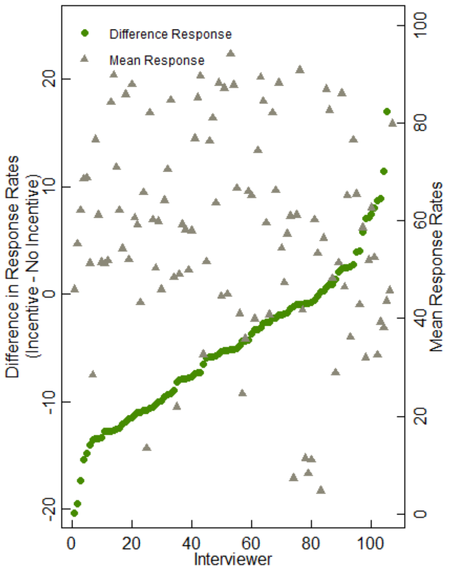


Figure 3-3: Difference in predicted rates of response for incentive and non-incentive households by interviewer for UKHLS-IP

3.6 Discussion

Survey methodologists have demonstrated that monetary incentives play a crucial role in motivating survey response and cooperation (Cantor et al., 2008; Church, 1993; Singer & Kulka, 2002; Singer & Ye, 2013). Incentives only increase headline response and cooperation rates by small percentages, which implies that most of the money spent on incentives is wasted. This is because the majority of respondents in any survey using a monetary incentive would have agreed to provide an interview anyway. However, based on Leverage Saliency Theory (LST) some respondents are susceptible to being converted from refusal to interview with the provision of an incentive (Groves et al., 2000). In turn, it is possible that interviewers might play an important role in determining the rate of such 'conversions' in face-to-face interviews because interviewers' characteristics influence survey response and cooperation (Blom et al., 2011; Durrant, Groves, & Steele, 2010; Hansen, 2006; Hox & de Leeuw, 2002). There is substantial research on the effects of incentives and interviewers on survey response and cooperation. However, it is surprising that little attention has been paid to identifying whether interviewers differentially influence the effectiveness of incentives on survey response and cooperation. Therefore, the motivation for this study has been to explore the influence of interviewers in survey response and cooperation in the deployment of incentives.

In this study interviewers influence on the effectiveness of incentives for survey response and cooperation has been explored. The findings indicate that interviewers vary in the deployment of incentives and the range of percentage difference in response and cooperation rates lies between -13% and 14% for NSW 2015, -8 % and 17% for NSW 2016, and -21% and 18% for Innovation Panel. This might be explained by the fact that under the norm of reciprocity interviewers may be more confident in approaching those households that have received an advance letter by restating the value of the incentives being offered (Singer et al., 1983; Singer & Maher, 2000). This implies that interviewers have the ability to make incentives more effective in promoting survey participation. This may be achieved by tailoring their interactions with potential respondents and reminding respondents about the incentive being offered which might lead to the outcome they expect (Blau, 1964; Singer & Ye, 2013).

The findings also show that interviewers who perform better in gaining good response and cooperation rates fare no better in the deployment of incentives, at least insofar as response rate on a single survey is a good measure of response rate attainment for interviewers. However, the effects of interviewers on incentive effectiveness are substantively as well as statistically significant; exchanging interviewers from the top to the bottom decile of interviewer performance would yield an expected 14 to 15 percent increase in the effect of the incentive relative to the control condition. This indicates that the effectiveness of incentives in motivating survey participation may be enhanced by recruiting appropriate interviewers and offering them better training to improve their efficiency of deploying incentives. The results do not provide evidence that survey response and cooperation is associated with the interviewer characteristics (i.e. age, gender, and experience) controlled for in this study. In addition, the cross-level interactions of interviewer characteristics and incentive were not significant. This implies that interviewers' variability in their effectiveness of deploying incentives is not moderated by interviewer characteristics. Therefore, variability of interviewers in deployment of incentives may be explained by other factors such as interviewers' attitudes and personalities towards households offered incentives, and those not offered incentives. This study did not control for interviewer attitudes and personalities due to data unavailability. Therefore, future research that aims to provide a clear understanding of the interactions between incentives and interviewers' attributes and personalities on survey response is required.

In summary, the results have three important implications for survey practice. First, the approach implemented here to identify interviewer effectiveness in deploying incentives could be used as a way of identifying underperforming interviewers. This kind of monitoring is now routinely implemented in adaptive and responsive surveys as a way of identifying interviewers who miss their fieldwork targets (Edwards, Maitland, & O'Connor, 2017; Kreuter, 2013). Therefore, "incentive performance" can be used alongside other forms of paradata to raise flags against interviewers on this performance dimension. Although further consideration is required to understand how this would be adapted in surveys in which all households are offered the same incentive.

Second, the approach used here can also provide guidance to survey organisations on the appropriate recruitment and training of interviewers on the deployment of incentives. Most survey organisations are now offering incentives aimed at improving response rates because these have persistently declined over the last 20 years. However, such incentives may be counterproductive if interviewers put too great a reliance on them and end up reducing their effort of convincing reluctant respondents to participate in surveys. Survey organisations may gain more benefit by pointing out to interviewers the interdependence that exists between them and effectiveness of incentives. Potentially, this will improve the way interviewers tailor their interactions with respondents rather than relying on incentives alone to improve response rates. In addition, the training may involve imparting skills on how to recognise and heighten the saliency of incentives in households where they are more likely to be effective.

Third, the ability to identify interviewers at the top end of the performance distribution offers opportunities for a better understanding of the sorts of strategies employed by more successful interviewers. Such interviewers may be encouraged to share their ideas and best practice with poorly performing ones. That is, good interviewers will be in position to steer the poorly performing interviewers in the right direction in terms of mediating the effects of incentives on survey response. This approach of good interviewers mentoring poor ones will in long run be cost effective for survey organisations. Additionally, information on successful approaches to incentive use that are identified in this way could be integrated into sections of interviewer briefings which address doorstep approaches, both for general and survey-specific training. In summary, highlighting to interviewers that the way they administer incentives can have substantial effects on their response outcomes can positively influence their subsequent behaviour.

This study has notable advantages when compared to other studies that have tried to investigate interviewer effects on incentives. First, data obtained from three different face-to-face surveys were analysed. This makes the findings and conclusions drawn from this study robust because of the comparisons made across all three surveys. Second, the application of multilevel models leads to an estimation of interviewer effects on incentives simultaneously with the effects of cross-level interactions between incentives and interviewer level characteristics. In conclusion, the results show that interviewers moderate the effects of incentives on both survey response and cooperation. The interviewer effects on incentives for survey cooperation are moderately smaller in comparison to survey response by DIC changes. The findings further show that cross-level interactions between incentives and socio-demographic characteristics for interviewers are not significant. The nonsignificant relationship between interviewer socio-demographic characteristics and survey response makes it hard to identify interviewers who are effective in deploying incentives. Therefore, with this data, it has been possible to show that interviewers differentially influence the effectiveness of incentives. However, it has not been possible to confirm the interviewer characteristics that influence their differences.

Nonetheless, this study has some limitations. First, the surveys considered all use a relatively narrow range of incentive values which are administered to all households in the incentive condition. Caution should therefore be exercised in generalising to contexts where larger incentives are used, or where incentives of varying values are targeted at different sub-groups of the sample based on response propensities (Lavrakas, McPhee, & Jackson, 2016). The results in this study also have little relevance to the use of incentives in online surveys, which comprise a large and growing proportion of total survey volume, both in the UK and internationally.

Second, the data used for this study did not allow controlling for survey variables because these variables are not available for both respondents and non-respondents. Singer, Hoewyk, et al., (1999) found that the effects of incentives tend to be relatively modest after controlling for survey variables. Probably, controlling for survey variables might have reduced the magnitude of the effects observed in this study. Survey process data (paradata) may be an alternative to survey variables since paradata contains rich information for both respondent and non-respondents. However, controlling for paradata in this study may be inappropriate because paradata may contain some interviewer bias. This may in turn introduce bias on estimates of interviewer effects on incentives. Third, this study did not control for variables measuring interviewer attitudes, beliefs, behaviours, and personalities that might explain why some interviewers are more effective in deploying monetary incentives compared to others. Future studies should address these issues in detail

Chapter 4 Do low-response rate online surveys provide equal or better data quality than high response rate face-to-face designs? Separating sample selection from measurement effects (Paper 3)

4.1 Introduction

For many years, face-to-face interviews have been the treated as the 'gold standard' method of data collection in survey research (de Leeuw, 1992; de Leeuw, 1992; Dillman et al., 2009). This is mainly due to higher contact and cooperation rates obtained in face-to-face interviews compared to other modes. The positive features of face-to-face interviews can mostly be attributed to interviewers who are tasked with locating and persuading sample members to participate in surveys. Additionally, interviewers highlight key survey design features such as incentives, survey topic, and the sponsor of the study in persuading potentially reluctant respondents to take part. Interviewers are also able to motivate respondents to complete questions, to provide explanations and clarifications for complex or ambiguous questions, and use show cards and other supporting materials all of which should, in principle at least, improve measurement quality. However, the substantial costs of conducting face-to-face interviews, and increasing nonreponse rates have necessisated the use of alternative modes of data collection (Dillman et al., 2009; Williams & Brick, 2018). The rapid pace of technological advancement in recent years has also transformed people's daily communication habits, leading to changes in data collection mode preferences (de Leeuw & Hox, 2011; Peterson et al., 2017; Tourangeau et al., 2013).

The main alternative mode of data collection to face-to-face interviews are online surveys, which have been substantially increasing in number and volume over the last 15 years (Callegaro, Lozar Manfreda, & Vehovar, 2015). Online surveys are considerably cheaper than face-to-face interviews although they tend also to have considerably lower response rates (de Leeuw, 2018; Dillman et al., 2009; Vannieuwenhuyze, Loosveldt, & Molenberghs, 2017) and higher rates of missing data (Heerwegh & Loosveldt, 2008; Jäckle et al., 2015; Lesser et al., 2012). One of the key concerns regarding online surveys is that the low response rates they tend to achieve may result in potentially large nonresponse biases. However, studies have shown that low response rates do not necessarily lead to nonresponse bias (Fricker & Tourangeau, 2010; Groves, 2006). In a meta-analysis of studies that produced estimates of nonresponse bias, Groves & Peytcheva (2008) found that high response rates reduce the risk of bias, although some surveys with low nonresponse rates had estimates with high relative nonresponse bias. Other studies have come to similar conclusions (Krosnick, 1999; Meterko et al., 2015; Rindfuss, Choe, Tsuya, Bumpass, & Tamaki, 2015; Sturgis et al., 2017; Wright, 2015).

This raises the possibility that low response rates may not be as strong an indicator of data quality as has traditionally been assumed. As a consenquence this raises the question of whether the longstanding presumed benefits of face-to-face interviews compared to alternate modes are as great as has conventionally been thought, since there is a lack of empirical evidence to support the idea (Burkill et al., 2016; de Leeuw, 1992; Tourangeau & Yan, 2007; Villar & Fitzgerald, 2017). Moreover, the presence of an interviewer may have a detrimental effect on response quality for surveys with sensitive questions (Burkill et al., 2016; Heerwegh, 2009; Kreuter, Presser, & Tourangeau, 2008). Respondents interviewed face-to-face tend to take social norms into account when providing answers to sensitive behavioural and attitudinal questions which leads to social desirability bias – the tendency to understate socially undesirable attitudes and behaviours and to overstate those that conform to social norms (Kaminska & Foulsham, 2013; Tourangeau & Yan, 2007). Online surveys are less prone to social desirability bias because respondents answer survey questions without being so influenced by the social presence of interviewers leading to more candid and accuarate responses to sensitive questions (Kreuter et al., 2008).

Thus, despite the longstanding assumption that face-to-face surveys provide the highest quality data, there are good grounds for questioning the extent to which this will always be the case. However, evaluation of differences in data quality between surveys conducted in different modes is complicated because gold-standard criterion variables are rarely available, so it is generally not possible to estimate bias (Dillman et al., 2009; Hox, de Leeuw, & Klausch, 2017; Klausch & Schouten, 2015; Vannieuwenhuyze & Loosveldt, 2013). Additionally differences in survey estimates across modes comprise a mix of sampling, selection, and measurement errors (Klausch & Schouten, 2015; Vannieuwenhuyze & Loosveldt, 2013). Selection effects are non-observational errors caused by differential coverage and nonresponse, while measurement effects are observational errors that arise during the process of reporting and recording an answer (Voogt & Saris, 2005; Weisberg, 2005). For proper evaluation of data quality between face-to-face interviews and online surveys it is crucial to differentiate between selection and measurement effects. However, this is not straightforward because selection and measurement effects are confounded (Vannieuwenhuyze & Loosveldt, 2013).

One of the strategies that can be used to separate the two sources of mode differences is to render the different modes comparable with regard to sample composition by using weighting or propensity matching (Lee, 2006; Lugtig et al., 2011; Vannieuwenhuyze & Loosveldt, 2013; Vannieuwenhuyze, Loosveldt, & Molenberghs, 2010). For example, a recent study conducted by Kantar Public in the United Kingdom assessed differences in data quality by applying nonresponse and attrition weighting to balance sample selection effects between general population samples interviewed online and face-to-face (Williams, 2017b). The study concluded that an online sample with a low response rate probably produced data of a *higher* quality than a contemporaneous face-to-face survey with a considerably higher response rate. If this conclusion is robust, it is very important because it opens the possibility of conducting surveys considerably more cost-effectively, without incurring a decline in data quality.

In this paper the data in Williams (2017b) is reanalysed using a different approach: propensity score matching (PSM). Lugtig et al. (2011) concluded that PSM is an effective approach for separating selection and measurement effects between modes. The approach of Lugtig et al.(2011) is extended here by using three different ways of estimating propensity scores (PS) based on how survey weights are included in the models : (1) unweighted, (2) weighted, and (3) unweighted with weights as covariate. Using each of the three different propensity score models, propensity score matching is used to create matched samples. The mode effects are then estimated using three different approaches within each matched sample depending on how survey weights are handled in outcome analysis: (1) no survey weights; (2) matched respondents in each mode retain their natural survey weights; and (3) matched respondents in one mode inherit the weights of the respondents in the reference mode. Generally, ignoring survey weights in complex design surveys may lead to bias and inaccurate variance estimates (Andrews & Oster, 2017).

This paper has two complementary objectives. First, it adds to understanding of how effective PSM is in removing selection differences between samples interviewed in different modes. Second, it uses the outcome of this assessment to evaluate whether it is reasonable to conclude that a low response rate online survey can produce data of equivalent to or even better quality than face-to-face surveys as suggested by Williams (2017b). The remainder of the paper is structured as follows. Sections 4.2 provides a literature review on the effect of mode of interview on survey data quality and section 4.3 reviews the application of PSM for removing differences in selection effects between samples. The next section 4.4 describes the data and analysis strategy, with the key findings from the analyses presented after that in section 4.5. The final section 4.6 of the paper summarises the key findings, considers the limitations of the methodological approach, and discusses the implications of the results for survey practice.

4.2 Survey Mode and Data Quality

Data quality in surveys has no universally accepted definition because researchers and experts tend to have different understandings depending on their discipline and methodological traditions. Broadly, however, survey quality can be defined in terms of two main perspectives: Total Survey Error (TSE) (Biemer & Lyberg, 2003) and quality management sciences (L. E. Lyberg, 2012). The TSE paradigm is the most widely used framework for defining survey data quality in the context of mean squared error (MSE) which is the sum of the random errors (i.e. variance) and squared systematic errors (i.e. bias) (Biemer, 2016; Cochran, 1977; Groves, 1989; Groves & Lyberg, 2010). Survey data with minimal MSE is deemed *appropriate for intended use* and *meets end user needs* (Alizamini, Pedram, Alishahi, & Badie, 2010; Biemer & Lyberg, 2003). However, it is difficult to accurately know the level of minimum MSE which the data may be deemed appropriate (Ellen Hansen et al., 2016; Vehovar et al., 2012). First, MSE is calculated differently for different survey parameters and the true scores used in bias estimation are often unknown since they are obtained from benchmark surveys which their accuracy is not guaranteed. Finally, it is often difficult to distinguish and separate the combination of different error sources which constitutes MSE.

The choice of data collection mode affects both who responds to a survey and how they answer which, in turn affects survey data quality (de Leeuw, 2018; Dillman, 2002; Jäckle, Roberts, et al., 2010). There is a wealth of evidence on how different methods of data collection influence survey data quality in the context of selection and measurement effects (de Leeuw, 2005, 2018; Dillman, 2002; Jäckle, Roberts, et al., 2010) which is too substantial to review in its entirety here. Instead, the focus of this study is limited to face-to-face interviews and online probability surveys. For online probability surveys, the sample units are sampled randomly from a list of addresses or pre-recruited from a panel of randomly recruited volunteers (Toepoel, 2012).

In principle, face-to-face interviews have several strengths compared to online probability surveys. First, interviewers can locate and persuade sample members to participate in surveys leading to higher response and cooperation rates (de Leeuw, 1992). In contrast , internet coverage is not universally available, and this can lead to noncoverage error (Blasius & Brandt, 2010; Tourangeau et al., 2013). Interviewers can verify the identity of the surveyed person to ensure that they are interviewing the sampled respondent (Couper, 2000), which is not possible in online surveys. Interviewers can also probe for explanations of responses allowing in-depth data collection and a better understanding of complex questions and responses (de Leeuw, 2005; Szolnoki & Hoffmann, 2013). Finally, interviewers can observe a respondent's body language and facial expressions, allowing them to make adjustments as needed if respondents are distracted or feel uncomfortable (Groves, 1989; Holbrook et al., 2003; Schober, 2018). This reduces rates of item missing data and breakoffs in face-to-face interviews compared to online surveys which are completed in a less controlled environment (de Leeuw, 1992; de Leeuw, Hox, & Huisman, 2003; Krosnick, 1991). This is despite questionnaires in online surveys being considerably shorter (Allen, 2016).

The positive features of face-to-face interviews come with significantly higher costs (de Leeuw, 2005, 2018). First, the process of interviewer recruitment and training is resource intensive. Second, locating respondents and conducting interviews is time consuming and resource intensive especially for hard to reach respondents. Respondents may also be unwilling to admit socially undesirable behaviours or opinions in person on sensitive questions, leading to biased responses (Burkill et al., 2016; de Leeuw, Hox, & Kef, 2003). The empirical evidence comparing the benefits and drawbacks of face-to-face interviews and online surveys in the current technological era is critically lacking. This is despite the fact that there has been a substantial shift from face-to-face interviews to online self-administration in the survey industry over the past fifteen years (de Leeuw, 2018).

The appeal of online surveys is largely driven by lower costs, technological advancement and societal change (de Leeuw, 2018). Online surveys are less expensive, enable fast data processing, and are flexible in terms of providing more complex displays to respondents (Beebe et al., 1997; Bethlehem & Biffignandi, 2011; Tourangeau et al., 2013). Despite these strengths, online surveys are potentially more susceptible to satisficing behaviour due to lower motivation compared to face-to-face interviews where interviewers motivate respondents (Kaminska & Foulsham, 2013; Krosnick, 1991; Krosnick, Narayan, & Smith, 1996).

Several studies have investigated mode effects across face-to-face and online surveys (Burkill et al., 2016; Heerwegh, 2009; Klausch & Schouten, 2015; Kreuter et al., 2010; Revilla & Saris, 2013; Williams, 2017b). Most studies have focused on differences in terms of response rates, item-nonresponse, satisficing and social desirability. For example, Burkill et al. (2016) using the National Survey of Sexual Attitudes and Lifestyle (Natsal-3) compared responses to 7 demographic and 31 behavioural and opinion questions provided by respondents between face-to-face interviews and online surveys. They found significantly higher response rates to sensitive questions in the online survey compared to the face-to-face interviews.

Villar & Fitzgerald (2017) investigated measurement differences between face-to-face and online respondents in the UK European Social Survey (ESS) Round 5 survey. They found that face-to-face interviews provided lower item nonresponse than online surveys. Heerwegh, (2009) found that an online survey generated higher item nonresponse but lower socially desirable responses compared to face-to-face interviews. Schouten et al. (2013) found large mode effects between face-to-face interviews and online surveys in a large-scale mixed-mode experiment linked to the Dutch Crime Victimisation Survey conducted on 2011. They concluded that biases in interviewer-mediated and self-administered surveys when benchmarked with respect to face-to-face interviews are not equivalent and should not be treated as one sample. Revilla & Saris (2013) used a split ballot multitrait-multimethod (SB-MTMM) approach to evaluate differences in data quality between online and face-to-face modes in terms of the strength of the relationship between latent variables and observed responses. They found that data quality does not vary between face-to-face and online surveys. However, they noted instances where there was a variation in data quality and it was usually higher in the online survey compared to face-to-face.

Of particular relevance to this paper is a study by Williams (2017b) which attempted to separate measurement from selection effects in parallel face-to-face and online surveys as part of the 2015 UK Community Life Survey (CLS). This survey asked questions on volunteering, donating, community engagement, civil duty and well-being. The samples considered were: initial face-to-face, online (follow up), and address based online surveying (ABOS). A second face-to-face sample was collected at the same time as online surveys to correct for change over time since interviews for the initial face-to-face sample were conducted at an earlier date than online samples. Williams concluded from this study that the majority of the total mode effect was caused by measurement rather than selection effects. This conclusion was based on weighting the online (follow up) survey and correcting for the change over time to make the sample composition similar across modes. The resulting differences between face-to-face and the weighted online (follow up) surveys were assumed, on this basis to comprise of only measurement differences.

Furthermore, Williams (2017b) concluded cautiously that the online samples provide better quality data than face-to-face survey. This was because the majority of questions in CLS were more susceptible to social desirability bias in face-to-face surveys compared to online surveys. Usually, online surveys as a self-administered mode tend to have superior measurement properties for sensitive attitudinal and behavioural questions compared to face-to-face interviews (Kaminska & Foulsham, 2013; Kreuter et al., 2008; Roberts, 2007). Finally, Williams (2017b) found that the differences between the online (follow up) and ABOS samples were small in magnitude, and attributed these to selection effects because both samples are in the same measurement mode.

The conclusion that the online samples provide better quality data rests on four key assumptions. First, there should be no selection effect differences between the initial face-to-face survey which was used as the basis for online (follow up) and the later face-to-face survey used to correct for change over time. Second, attrition weighting should remove all selection differences between the online (follow up) sample and the initial face-to-face survey. However, this is may not have been the case because a review by Tourangeau et al. (2013) found that weighting schemes only remove 30-60% of selection effects between online surveys and face-to-face surveys. Third, in estimating mode effects using a pre-recruited online sample, the comparison assumes that having previously completed the same questionnaire face-to-face had no effect on the online answers. Yet there is good evidence that repeated interviewing of this nature can result in panel conditioning effects (Sturgis, Allum, & Brunton-Smith, 2009). Last, the differences between the two online surveys should be purely attributed to differential selection effects. However, this may not be the case because of the potential for different mixes of device types being used across the two samples (Lugtig & Toepoel, 2016).

4.3 Separating Selection and Measurement effects

One approach that can be used to separate selection and measurement effects involves using common variables in each sample to make the different modes equivalent with regard to sample composition (Vannieuwenhuyze & Loosveldt, 2013). The common variables are used as predictors of the respondent’s propensity to be in a specific mode. Conditional on this propensity, remaining differences between modes are assumed to be due to measurement effects. Propensity score matching (PSM) tends to be more successful in removing bias under correct model specification, compared to weighting (Ertefaie & Stephens, 2010; Hahn, 1998; Hirano et al., 2003). This is because PSM is capable of providing good covariate balance between matched groups which ensures that any differences are not as a result of differences on the matching variables.

Lugtig et al. (2011) applied PSM to a survey carried out in different modes and found it to be effective at removing selection effects between online and face-to-face samples. They found large differences between telephone and online surveys, even after matching which they took as indicating the presence of measurement differences across the two modes. On the other hand, they found that PSM removed the differences caused between two online samples. This led them to conclude that PSM is an effective way of separating measurement differences from sample selection effects in surveys using different modes. However, the Lugtig et al. (2011) study had three limitations. First, they considered only 7 questions for evaluation of mode effects which is a small number of variables for generalising to all survey purposes. Second, the matched samples had fewer than 250 respondents, due to a high number of respondents discarded during the matching process. This means their tests of difference were low powered (Caliendo & Kopeinig, 2008). Lastly, the study failed to include the influence of survey weights in the estimation of propensity scores for matching purposes

and in the outcome analysis of mode effects for matched samples. Considering that most mixed-mode surveys use complex survey designs it is important to consider the influence of survey weights in the estimation of propensity scores and outcome analysis for matched samples.
Several studies have addressed the use of survey weights in propensity scores (Austin et al., 2018; DuGoff et al., 2014; Lenis et al., 2017; Ridgeway et al., 2015; Zanutto, 2006). For example, Zanutto (2006) concluded that it is important to incorporate survey weights from complex surveys in outcome analysis, but not when estimating propensity score models. According to Zanutto (2006), ignoring survey weights in the outcome analysis may substantially affect the estimates of population level effects. Dugoff et al. (2008) recommended that survey weights should be incorporated as a covariate in the propensity score model. Dugoff et al. (2008) also gave the same recommendation as Zanutto (2006) that survey weights should be incorporated in the outcome analysis when making inferences about the population level estimates. However, Ridgeway et al. (2015) recommended that survey weights should be included as weights in propensity score model since they lead to treatment effects with the lowest MSE.

Recent publications consider the use of sampling weights in the context of PSM in complex designs Austin et al. (2018) and Lenis et al. (2017). They consider three ways of incorporating survey weights in propensity score models: (1) unweighted model, (2) weighted model and (3) unweighted model with survey weights included as a covariate in the model. Lenis et al. (2017) found that survey weights incorporated in propensity score models do not influence the estimation of the population treatment estimates. On the other hand, Austin et al. (2018) produced inconclusive findings on which of the three different formulations of the survey weights on propensity score models was preferable.

Austin et al. (2018) and Lenis et al. (2017) also investigated which survey weights to assign to matched samples in outcome analyses, by considering three possible specifications: (1) no survey weights; (2) matched units in each sample retain their natural survey weights; and (3) matched control unit inherits survey weight of the treatment unit it is matched to. Austin et al. (2018) recommend that matched control units should retain their survey weights because they lead to decreased bias. On the other hand, Lenis et al. (2017) suggested that matched control units should use inherited weights of the treated units they are matched to as survey weights, because this specification can be beneficial when the missing data mechanism is missing at random. To be specific, where the nonresponse depends on the baseline covariates and the treatment assignment. Therefore, based on this lack of clear consensus of which survey weights to use for matched control units for outcome analysis, it becomes necessary to consider all three different specifications in the context of mixed-mode designs.

4.4 Data

Data for this study comes from the Community Life Survey (CLS) study, which was carried out between July and September 2014 (Williams, 2017b). This involved administration of the CLS questionnaire in three independent samples: a face-to-face; an online (follow up) survey drawn from an existing face-to-face survey; and an Address Based Online Survey (ABOS) for adults aged 16 years and above. The study design and corresponding response rate (RR) for each survey are presented in Figure 4-1 and they are described in more detail below. Since both online (follow up) and ABOS samples had both paper and online completions, Figure 4-1 presents only those respondents who responded using online, the focus of this study.

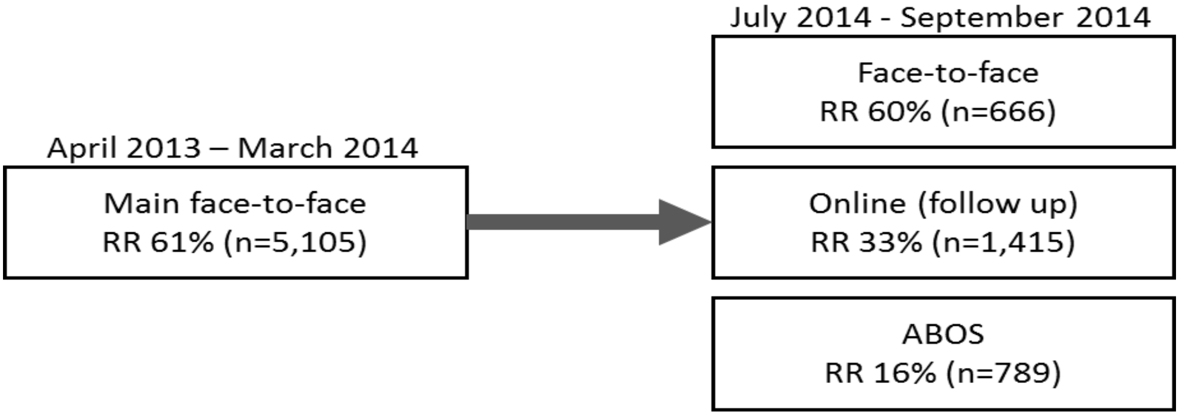


Figure 4-1: Representation of the CLS study

4.4.1 Face-to-face Survey

A multi-stage random sample design was employed for the face-to-face CLS. A stratified random sample of postal sectors was drawn in England with probability proportional to size. Addresses within each selected postcode had an equal probability of selection at the second stage of sampling. Where the number of dwelling units was greater than one, the interviewer used a random number generator to sample one household. The same random number generator was used to sample one adult for interview at sampled addresses containing more than one adult. The data collection was between July and September 2014 and six in-person interviewer visits were conducted before a case was considered non-contact. The issued sample size was 1,110 and 666 respondents were successfully interviewed representing a 60% response rate.

4.4.2 Online (Follow up) Survey

The online (follow up) survey was drawn from respondents who had participated in the main face-to-face Community Life Survey of 2013-14 who had given consent to be re-contacted. The sample design of the 2013-14 CLS was the same as the face-to-face survey described above. The fieldwork was undertaken from July to September 2014. The number of respondents for the main CLS 2013-14 was 5,105 and 4,219 (83%) gave consent to be re-contacted to participate in an online survey. Of those re-contacted, 1,576 (37%) responded with 1,415 (89.8%) using online completion and 161 (10.2%) completing paper questionnaires returned by post. The postal sub-sample was excluded in this analysis because the focus is on face-to-face interviews and online surveys.

4.4.3 Address based online Surveying (ABOS)

The Address Based Online Surveying (ABOS) design involves drawing a stratified random sample of addresses from the Royal Mail's Residential Postcode Address File (PAF) with addresses sampled with equal probability (Williams, 2017a). After drawing the sampled addresses, invitation letters containing username(s), password(s) and the survey website url are sent to occupant(s) inviting the resident adult(s) to complete the survey online. Where there is more than one eligible adult at an address, all adults are asked to complete the survey, up to a maximum of four adults. The intention of allowing more than one individual from the same household to participate in a survey was introduced to minimise issues that may arise within household sampling stage when respondents ignore sampling instructions in self-completed surveys. However, this may lead to multiple completions by one respondent in the same household. Therefore, to ensure that the data quality is achieved from sampled individuals an algorithm is used to verify that the data obtained meet the set standards. However, the ABOS design does not control which household is selected in multi-household addresses. The ABOS design also has a paper option for individuals who do not have access to the internet or who prefer to complete a paper questionnaire. Fieldwork for this survey was undertaken in July to September 2014. The number of respondents completing an interview was 834, representing a response rate of 17% with 789 (94.6%) using online completions and 48 (5.4%) using postal completions which were excluded for the final analysis.

4.5 Methodology

Propensity Score Matching (PSM) is used to remove sample selection differences between the three independent samples by matching respondents between samples on a set of observed covariates (Imbens, 2004; Rosenbaum & Rubin, 1983). The aim is to generate a matched sample such that for every respondent in one survey mode there is at least one respondent from the other sample with similar characteristics on the vector of matching variables. The propensity scores are estimated using logistic regression (Agresti, 2013). Let y_i denote the binary outcome (i.e. survey modes assigned to survey participants) for respondent i ($i = 1, \dots, n$) where y_i is assumed to be conditionally distributed as Bernoulli, with conditional response probabilities defined as $\pi_i = Pr(y_i = 1)$ and $1 - \pi_i = Pr(y_i = 0)$. The logistic regression model takes the form

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_j x_i = \mathrm{B}^T X_i \quad (4.1)$$

where $\mathrm{B} = (\beta_0, \beta_1, \dots, \beta_j)$ is a vector of regression weights and X_i is a vector of covariates at the respondent level. The choice of variables to include as covariates in the propensity score model is important because it affects bias, variance and MSE of the estimated treatment effects (Austin et al., 2007; Brookhart et al., 2007; Smith & Todd, 2005). First, only those variables that have a direct effect on the probability of treatment assignment (i.e. mode of data collection) and are related to the outcome of interest should be included in a propensity score model (Brookhart et al., 2007; Guo & Fraser, 2014). Additionally, Brookhart et al. (2007) recommend inclusion of variables that are not related to treatment assignment but are related to the outcome of interest (i.e. potential outcomes). These variables lead to a reduction in the variance of estimated treatment effects without increasing bias. Lastly, only those variables that are measured at baseline should be included in the propensity score model to avoid using variables that might themselves be subject to mode effects (Austin, 2011a).

In this analysis, a set of 12 socio-demographic and area-level variables deemed appropriate covariates for inclusion in the propensity score models are considered. The literature shows that sociodemographic questions are less prone to measurement effects compared to behavioural and attitudinal questions which are majority of questions in CLS (Brookhart et al., 2007; Burkill et al., 2016; de Leeuw & Hox, 2011). This is a small number of variables compared to the 32 variables that Williams (2017b) considered in the computation of attrition weights. Williams considered both socio-demographic and attitudinal variables. However, this is problematic because inclusion of attitudinal variables in computation of attrition weights may remove some mode effects resulting in biased estimates because these variables are themselves subject to measurement effects across modes (Brookhart et al., 2007; Cuong, 2013). The final propensity score model consists of the variables with significant univariate relationship with the binary outcome (i.e. choice of mode) based on the 95% significance level (Hirano & Imbens, 2001). The adequacy of the propensity scores estimated using the propensity score model is determined by evaluating the area of common support (Austin, 2011a; Leite, 2017). This is the extent of the overlap in the distribution of propensity scores of respondents in different modes and is evaluated using histograms and boxplots (Austin, 2011a; Leite, 2017). Once the adequacy of common support is attained, the next step involves matching respondents between different modes.

Respondents for two different modes are matched on the logit of the propensity score using so-called 'greedy' nearest neighbour and calliper matching (G-NNCM) without replacement (Rosenbaum, 2002; Stuart, 2010). This is implemented using one to one matching where each respondent in a given mode is matched to one respondent from the other mode. One to one matching allows the mode with a smaller number of respondents to drive the power of matching which leads to an increased homogeneity of the matched sample, resulting to a reduction in bias of treatment effect (Cohen, 1988).

G-NNCM begins with randomly ordering respondents of two different modes based on their propensity scores. Then the first respondent from one mode is selected followed by finding the corresponding respondent with the closest propensity score within a specified calliper from the other mode. The two matched respondents are then removed from the matching sample and the next respondent is selected for matching purposes. The G-NNCM has a superior performance in terms of reduced bias for estimated treatment effects in matched sample compared to other matching algorithms (Austin, 2012). Specification of calliper during matching also improves the quality of matched samples by avoiding bad matches which are normally discarded if they are not within the defined width (Austin, 2008b; Smith & Todd, 2005). A matching calliper of width equal to 0.2 of the standard deviation of the logit of the propensity score is used because it leads to a better reduction in bias of estimated effects compared to other alternatives (Austin, 2009b; Caliendo & Kopeinig, 2008; Guo, Barth, & Gibbons, 2006). G-NNCM with calliper matching is implemented using the MatchIt package in R (Austin, 2011a; Ho, Imai, King, & Stuart, 2009).

The quality of matched samples is assessed in terms of covariate balance. Covariate balance is defined as the similarity of the empirical distributions of the full set of covariates included in the propensity score model and is evaluated using histograms, absolute standardised mean differences (SMD) and chi-square tests (Leite, 2017; Linden, 2015; Stuart, 2010). For histograms, categories of each covariate for matched samples are overlapped and any nonoverlapping areas indicate a lack of covariate balance. The SMDs are used to quantify the difference in means of the pooled standard deviation between matched samples (Austin, 2011a; Stuart, 2010). SMD is a robust approach for evaluating covariate balance before and after matching because it is not affected by the sample size. Adequate covariate balance for matched samples is achieved if the values of SMD are below 0.1 standard deviation for all the covariates used in the propensity score model (Austin, 2011a). According to Nguyen et al. (2017), the SMD threshold of 0.1 leads to unbiased estimates of treatment effects. A Chi-square test for independence is also used to test whether the frequencies of categorical covariates used in propensity score models are statistically equivalent across the matched samples. Covariate balance for a matched sample is achieved if the chi-square test for independence is not significant for the covariates considered.

The outcome analysis for the mode effects in the matched sample are estimated based on three specifications of outcome analysis: (1) no survey weights on the outcome analysis, (2) matched respondents from either mode retain their natural weights, and (3) matched control respondents inherit the weights of the treated respondents to which they are matched. Therefore, nine different methods for estimating the measurement effects were applied: three different methods for estimating the propensity score combined with three different analytical strategies within each matched sample.

4.5.1 Estimation of selection and measurement effects

Mode effects in this study are evaluated using the Absolute Percentage Differences (APD) between the same variables measured in different modes. The APD estimates are used because they are more intuitively interpretable compared to other measures such as standardized scores or relative absolute differences (Schouten et al., 2013). The APD is calculated by taking the un-signed difference in the proportion for each survey outcome across independent samples. It is important to note that APD estimates are computed for behavioural and attitudinal questions only. For categorical variables, APDs are calculated for each category with one category omitted for the combined analysis. That is, for a categorical variable with K response levels, $(K-1)$ APD estimates are derived, where the omitted categorical level is the one with the lowest frequency. Therefore, APD is the proportion in each category at each survey question and the proportion in the final achieved sample. For the computations of the proportions, the frequency of each category was treated as numerator while the sum of the frequencies of the given category and the omitted category level was the denominator.

To reduce undue influence of differences between sparse cells on the estimation of mode effects, only categories with proportions ranging between 5% and 95% are considered. Categorical levels with proportions that are not within this range of 5% and 95% are dropped from the analysis. The APD estimates are compared before and after matching and presented graphically based on different formulations of the propensity scores. The median is preferred as a measure of central tendency due to outliers and skewness in the distribution of APD estimates.

4.6 Results

Figure 4-2 presents histograms before and after matching for the three different analysis samples (face-to-face vs online (follow up), face-to-face vs ABOS, and online (follow up) vs ABOS). The propensity scores obtained based on three different specifications of survey weights in the propensity score models were similar. Therefore, only the results for the weighted model are presented here while the analyses for other model specifications are presented in Appendix C. The X-axis represents respondent propensities of using a given mode of data collection and the Y-axis represents the number of respondents. The face-to-face sample is represented by the solid grey bars and the online (follow up) sample is represented by shaded black bars. The shaded red bars represent the ABOS sample.

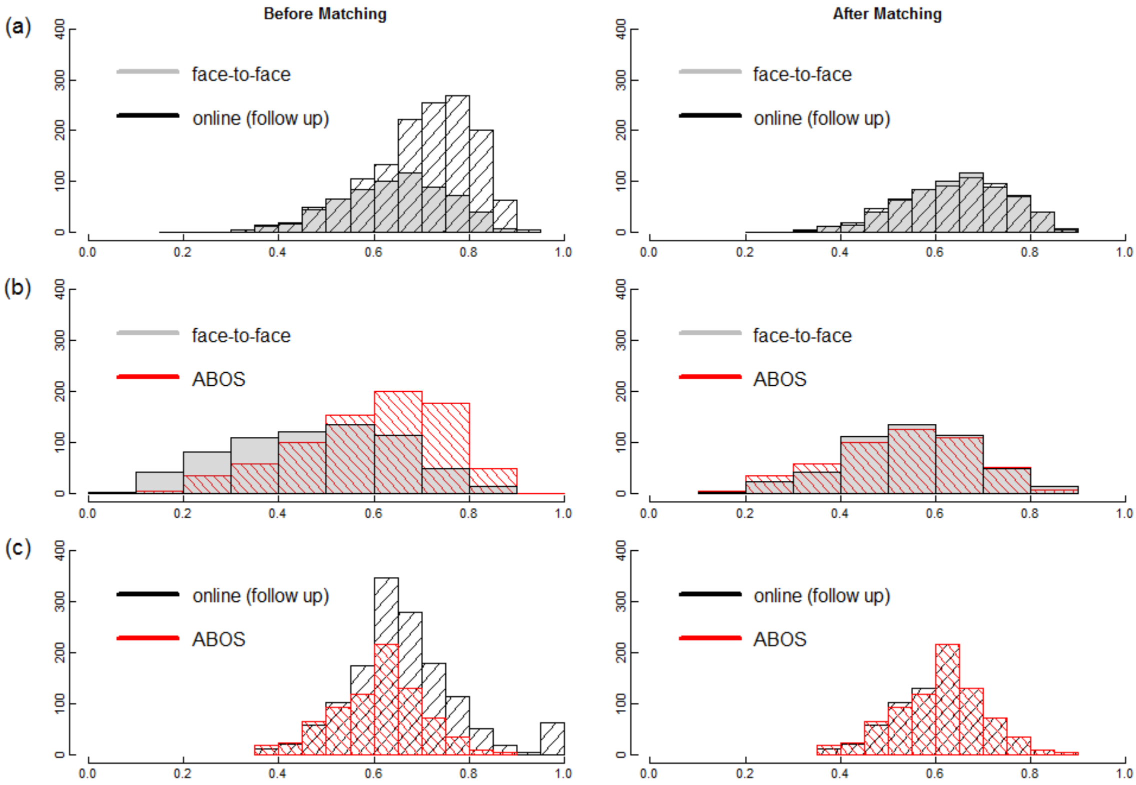


Figure 4-2: Histograms of propensity scores distributions before and after matching face to-face and online (follow up) (top panel), face-to-face and ABOS (middle panel) and ABOS and online (follow up)

Histograms before matching show that some respondents have overlapping propensity scores implying that they have a positive probability of being assigned to each mode when matched. This indicates that common support is potentially adequate to estimate measurement effects with propensity matching approach, because the distribution of the respondent in one mode is contained within the distribution of the other mode, and therefore adequate matches of respondents can be found between modes. The histogram for the face-to-face and ABOS before matching indicate that estimating the measurement effects using propensity score matching may be difficult because there are areas of the distribution of the face-to-face respondents without any ABOS respondents and vice versa, which can result in poor matching. However, the use of caliper matching will improve the quality of matched sample by avoiding bad matches, although this may lead to a higher number of unmatched respondents resulting to an increase in the variance of the estimated measurement effects (Caliendo & Kopeinig, 2008). This adequacy of common support is further supported by extent of the overlap in the distribution of propensity scores for matched samples presented by histograms after matching across the three different samples.

Table 4-1 shows the sample sizes before and after matching for the three samples. Of interest is the number of respondents discarded in the mode with smaller sample size before matching. This is because the number of respondents discarded influences the size of the variance of the estimated mode effects in the matched sample (Cohen, 1988). For the face-to-face and online (follow up) only 3% face-to-face respondents were discarded. A similar percentage of 3% for ABOS respondents were discarded when ABOS matched with online (follow up). This indicates that face-to-face and online (follow up), and ABOS and online (follow up) samples had only a few unacceptable matches with the defined calliper. The low percentage of discarded respondents indicates a higher homogeneity of matched samples resulting in a reduction in bias of the estimated mode effects (Caliendo & Kopeinig, 2008). However, the percentage of face-to-face respondents discarded after matching face-to-face and ABOS samples is 26%. This is because many face-to-face and ABOS respondents' propensity scores are not within the defined calliper of 0.2 standard deviations and are therefore discarded to avoid poor matching. The higher number of unmatched respondents in face-to-face and ABOS samples results in an increase in the variance of the estimated mode effects (Caliendo & Kopeinig, 2008). Therefore, based on these results, matching has been successful in all the three matched samples. However, it should be noted that estimated mode effects from matched face-to-face and ABOS sample may be susceptible to a higher variance. This is because they are many types of people in face-to-face survey who are not found in ABOS sample, which results in bigger differences between unmatched and matched samples for this comparison. It is crucial to note that the higher number of unmatched respondents between face-to-face and ABOS samples can be reduced by using matching with replacement approach.

Table 4-1: The sample sizes before and after matching (weighted model)

	Face-to-face and online (follow up)		**Face-to-face and ABOS**		**ABOS and online (follow up)**	
	Face-to-face	*Online (follow up)*	*Face-to-face*	*ABOS*	*ABOS*	*Online (follow up)*
Before matching	666	1,410	666	781	781	1,410
After matching	649	649	492	492	760	760
Discarded	17(2.5%)	761(54.0%)	174(26.1%)	289(37.0%)	21(2.7%)	650(46.1%)

Table 4-2 presents the standardised mean differences (SMD) for the three samples based on three different formulations of the propensity score models (unweighted, weighted and weight as a covariate). The first column of Table 4.2 represents the variables that were included in the propensity score models. The results show good covariate balance for the three different formulations, since they all produced SMD lower than 0.10. If the variables used in the matching are sufficient to account for sample selection differences, residual differences in APD estimates after matching can be interpreted as being due to measurement effects.

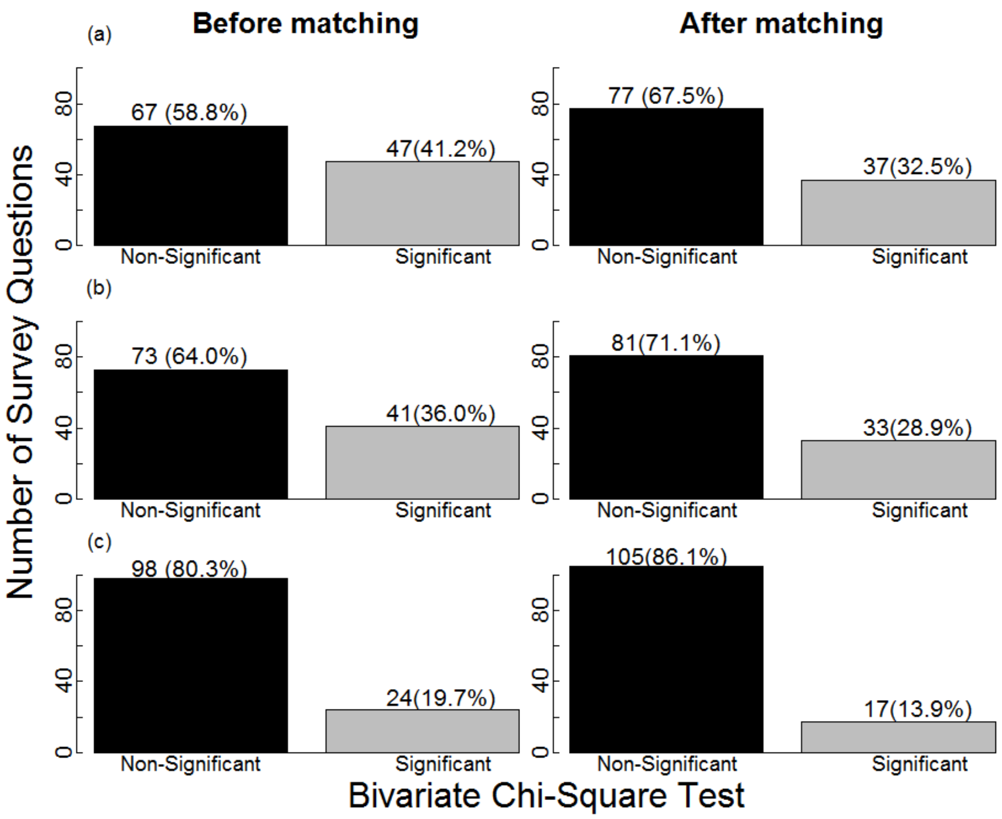
The PSM has effectively balanced three samples based on the observed covariates. Formulation of the propensity score model based on different survey weight specification had negligible impact on the baseline balance. This is consistent with the findings of Austin et al. (2018) that none of the different propensity score models resulted in a better balance of baseline covariates than other specifications.

Figure 4.3 shows bivariate chi square tests of the survey questions before and after matching in three samples. The Y-axis contains the number of questions while X-axis the bivariate chi-square test (i.e. significant or non-significant). Matching will be deemed effective between two different modes if the percentage of survey questions with significant bivariate chi-square test is no greater than 5% after matching. If this assumption is met, then any selection and measurement differences in matched sample are deemed to be caused by chance.

It is clear from Figure 4.3 that selection differences exists across three samples after matching. This is because the percentage of survey questions with significant chi-square tests is greater than 5% in all matched samples. The fact that 14% of survey questions were significantly different after matching the two-online sample suggests that matching was not completely effective in removing selection differences. Considering a third of survey questions for matched face-to-face and online samples are significantly different suggests that the APD estimates obtained will comprise not just measurement differences but also some selection differences. However, a reduction of survey questions with significant bivariate chi-square tests is observed across three samples at 9% ,7% and 6% for face-to-face and online (follow up), face-to-face and ABOS, and ABOS and Online (follow up) samples respectively. This suggests that matching removed some selection differences across the three matched samples, as would be expected. The higher percentage of survey questions with significant bivariate chi-square tests for face-to-face and online (follow up) at 33%, and face-to-face and ABOS at 29% compared to the two online samples at 14% is evidence of mode effects between face-to-face and online samples. This corresponds to the analysis of Williams (2017b) that 42% of survey questions had significant t-scores of the estimated mode effects compared to only 11% of survey questions that had sample effects.

Table 4-2: Standardised Mean Differences (SMD) for baseline covariates used in propensity score models based on three different formulations

		Face to face and ABOS			**Face to face and online (follow up)**			**ABOS and online (follow up)**		
Variable {Ref}	*categories*	*Unweighted model*	*weight as covariate model*	*Weighted Model*	*Unweighted model*	*weight as covariate model*	*Weighted Model*	*Unweighted model*	*weight as covariate model*	*Weighted Model*
Propensity scores		0.03	0.04	0.03	0.08	0.07	0.06	0.02	0.03	0.02
Age	16 to 34 year*s*	0.01	0.01	0.07	0.04	0.01	0.00	0.00	0.00	0.02
	35 to 49 years	0.00	0.05	0.02	0.06	0.03	0.01	0.02	0.00	0.04
	50 to 64 years	0.00	0.00	0.01	0.08	0.03	0.05	0.01	0.03	0.02
	65 to 74 years	0.02	0.05	0.00	0.02	0.04	0.02	0.04	0.03	0.01
	Over 75 years	0.03	0.02	0.08	0.04	0.05	0.09	0.02	0.01	0.02
Race {Others}	White	0.05	0.01	0.06	0.04	0.02	0.01			
Number of adults in household {1}	2	0.01	0.03	0.07	0.02	0.01	0.07	0.02	0.08	0.05
	3	0.01	0.02	0.01	0.05	0.06	0.06	0.02	0.02	0.05
	4 or more	0.01	0.06	0.04	0.01	0.01	0.05	0.04	0.03	0.04
Income	0 to < £15K	0.03	0.01	0.05	0.01	0.03	0.01	-	-	-
	£15K to <£40K	0.01	0.01	0.00	0.02	0.04	0.03	-	-	-
	>£40K	0.04	0.07	0.05	0.01	0.00	0.03	-	-	-
Tenure {Private rent}	Mortgaged	0.01	0.05	0.05	0.08	0.07	0.06	-	-	-
	Outright ownership	0.01	0.03	0.04	-	-	-	-	-	-
	Social rent	0.02	0.02	0.01	-	-	-			
Education {No qualification}	Other qualification	0.06	0.03	0.06	0.03	0.00	0.01	-	-	-
	Degree or above	0.04	0.04	0.00	0.01	0.04	0.04	-	-	-
GOR {London)	East Midlands	0.08	0.07	0.02	0.02	0.03	0.08	0.01	0.06	0.01
	East of England	0.02	0.05	0.03	0.02	0.02	0.03	0.02	0.02	0.01
	North East	0.02	0.01	0.02	0.03	0.02	0.06	0.03	0.01	0.01
	North West	0.01	0.00	0.01	0.01	0.04	0.02	0.00	0.00	0.02
	South East	0.04	0.05	0.02	0.08	0.01	0.02	0.01	0.01	0.00
	South West	0.02	0.01	0.03	0.02	0.02	0.07	0.06	0.04	0.00
	West Midlands	0.01	0.01	0.03	0.05	0.04	0.02	0.03	0.03	0.01
	Yorkshire and Humberside	0.02	0.01	0.00	0.02	0.01	0.04	0.01	0.02	0.04
Number of children {0}	1	-	-	-	0.01	0.03	0.01	0.05	0.02	0.02
	2	-	-	-	0.10	0.07	0.03	0.00	0.00	0.05
	3 or more	-	-	-	0.02	0.02	0.03	0.01	0.03	0.04
Sampling weights		-	0.02	-	-	0.05	-	-	0.01	-



*

Figure 4-3: Barplots of bivariate chi-square tests of the survey questions before and after matching (weighted model) for face-to-face and online (follow up) (a), face-to-face and ABOS (b), and online (follow up) and ABOS (c)

Figure 4-4 shows the p-values obtained from bivariate chi-square tests of the survey questions before and after matching in the three samples. The survey questions on the X-axis are ranked based on the p-values before matching, while the Y-axis represents the values of p-values. P-values before matching are represented by the black filled circles and the green stars represent p-values after matching. The red dotted line represents a p-value of 0.05 since a bivariate chi square test is deemed significant if $p - value \leq 0.05$. It can be observed that majority of survey questions with significant p-values before matching also tend to have significant p-values after matching across the three samples. This indicates that matching has less influence on selection and measurement differences across the modes considered. It is also important to note that only a few survey questions changed from being significant before matching to nonsignificant after matching.

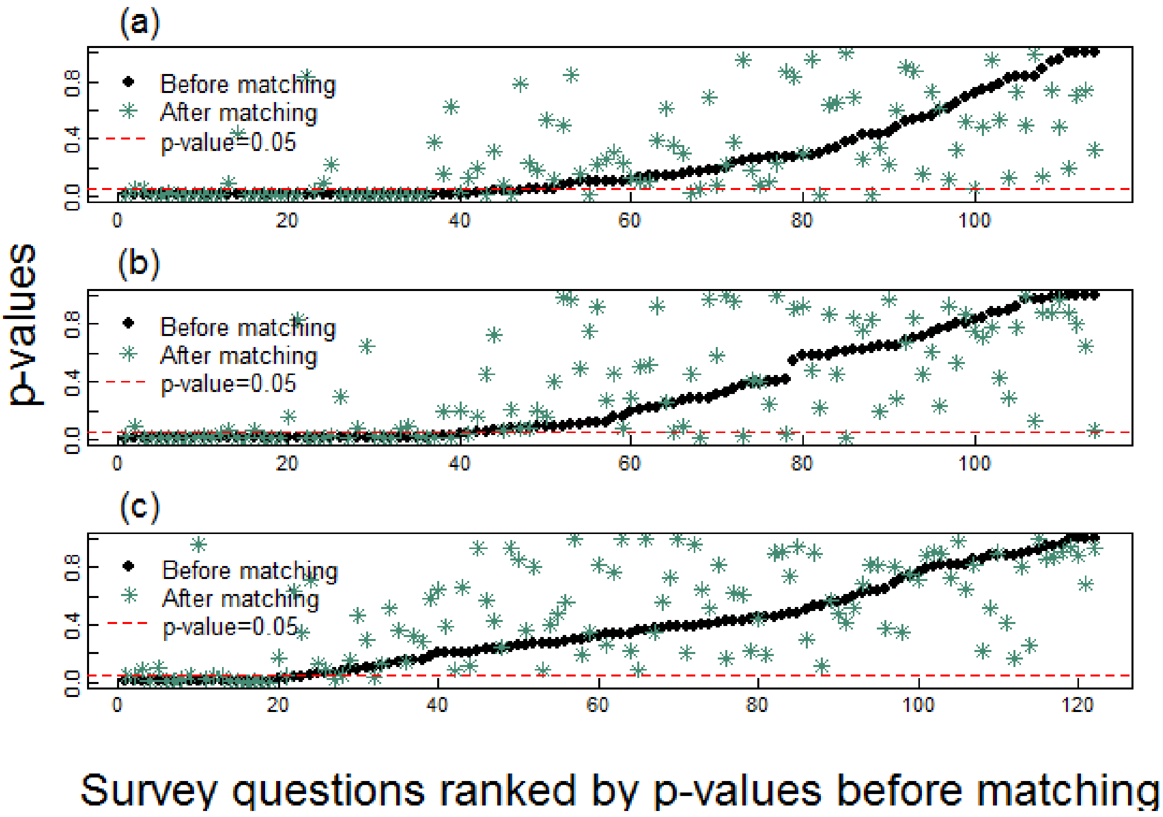


Figure 4-4: P-values by survey questions before and after matching (weighted model) for face-to-face and online (follow up) (a), face-to-face and ABOS (b), and online (follow up) and ABOS (c)

Figure 4-5 summarises the results of APD estimates obtained for the three samples (i.e. face-to-face vs ABOS, face-to-face vs online (follow up), and ABOS vs online (follow up) before and after matching. Different specifications of survey weights in propensity score models and outcome analysis (i.e. application of different specifications of survey weights when estimating APD after matching) did not have an impact on the estimated APDs. Therefore, APD estimates will be presented for matched samples obtained using propensity score estimated in a weighted model and with survey weights not controlled for in outcome analysis. Each dot represents an APD estimate for each survey question before (black) and after (green) matching. The survey questions on the X axis are ranked based on absolute percentage differences (APD) before matching. The pattern of the plots in Figure 4-3 is similar across the three analysis samples before matching. The APD estimates vary across the survey questions in the three analysis samples after matching.

The mode effect is considerably larger comparing face-to-face to online surveys before matching compared to the difference between the two online samples. The median APD for the face-to-face and online (follow up) is 5 percentage points before matching, increasing to 5.5 percentage points after matching. In general, the expectation is that the average mode effect should decrease after controlling for selection effects. However, the counter-intuitive pattern where the APD increases here may be attributed to selection and measurement effects having different signs, which is to say that they counteract each other (Schouten et al., 2013). That is, because the APD combines selection and measurement differences which can be in opposite directions, the asymptotic expectation of the APD after matching is not zero.

The median APD for the face-to-face and ABOS surveys reduces from 4.2 before matching to 4.0 after matching. This suggests that almost all of the mode difference between face-to-face and online surveys is due to measurement effects, if we assume that the matching successfully removes the selection effect component of the difference (Lugtig et al., 2011). However, this is not the case because APD estimates for the face-to-face and ABOS surveys contain both selection and measurement differences since a third of survey questions had significant bivariate chi-square tests after matching. Despite this drawback, these results are consistent with the findings of Williams (2017b) who estimated average measurement effects of 3.8% using nonresponse and attrition weighting to remove selection differences.

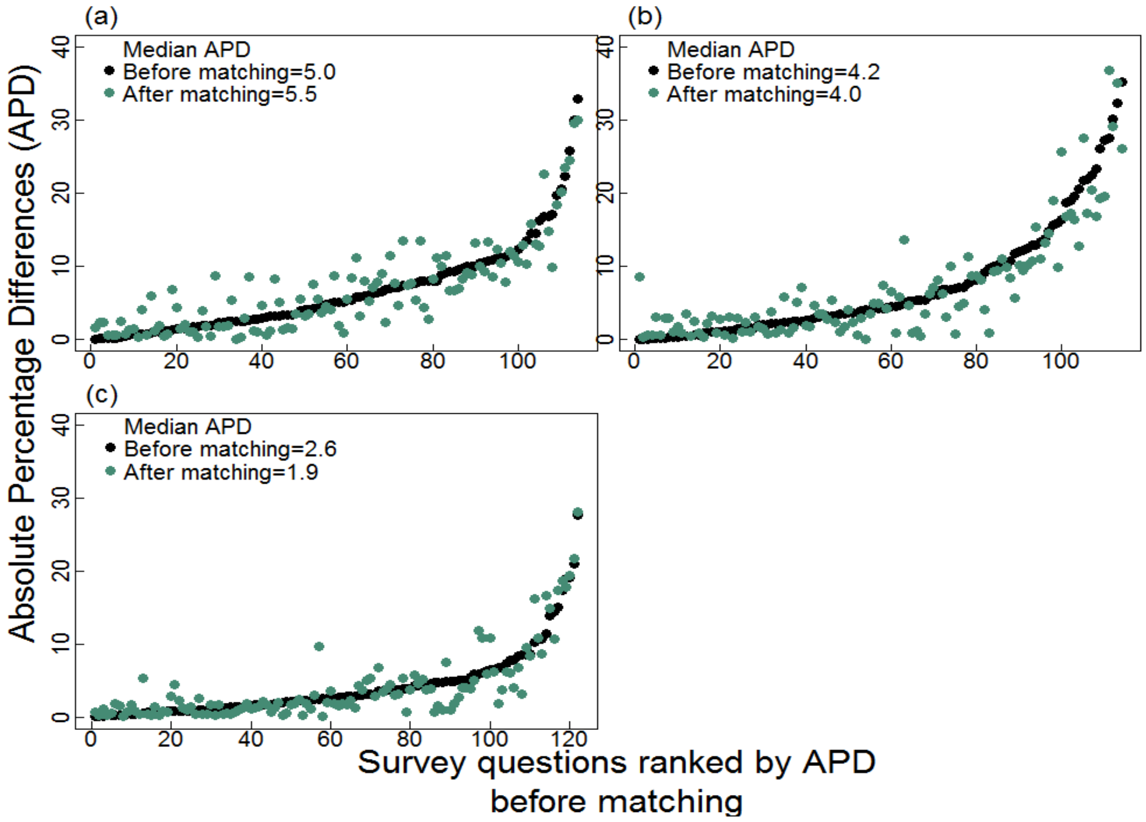


Figure 4-5: Estimated mode effects by Question before and after matching (weighted model) for face-to-face and online (follow up) (a), face-to-face and ABOS (b), and online(follow up) and ABOS (c)

The median APD for the two online surveys before matching is 2.6 percentage points which reduces to 1.9 percentage points after matching. As the two surveys are in the same mode, this suggests that the matching has not been successful in removing all the selection differences. As noted earlier, however, it is unclear how much of the 1.9% APD is due to selection differences. This supports a finding by Tourangeau et al. (2013) who found that weighting methods remove only 30-60% of selection effects in online surveys. This is probably because important variables are omitted in the propensity score models, leaving selection bias after matching. The APDs between the two online samples could also be due to a different mix of device types used to complete the survey (de Bruijne & Wijnant, 2013). This could not be controlled for in this analysis because it was not possible to obtain an indicator of device type used to complete the survey.

Figure 4-6 summarises the APD estimates based on the size of the APD before and after matching: 0-2.5%, 2.6-5.0%, 6.0-10.0%, 11.0-15.0%, 16.0-20.0%, and >20.0%. The X-axis represents the number of survey questions while the Y-axis represents APD estimates categorised into six levels namely: 0-2.5%, 2.6-5.0%, 6.0-10.0%, 11.0-15.0%, 16.0-20.0%, and >20.0%). Figure 4-6 aims to show whether matching exerts different influences across the distribution of mode of effects in the different APD categories. For example, it can be observed that 30% of survey questions in face-to-face and online (follow up) are classified in category 0-2.5% before matching which increases to 33% after matching indicating a 3-percentage point increase. On the hand, the percentage of survey questions in category 2.6-5.0% reduced by 7 percentage points from 20% before matching to 13% after matching. This representation allows an investigation of APD in other parts of the distribution which may be missed in APD medians. To obtain the percentage number of survey questions with median APD greater than 5% in each sample, the percentage points for categories 6.0-10.0%, 11.0-15.0%, 16.0-20.0%, and >20.0% are summed. The summed percentages of survey questions with APD greater than 5% in face-to-face and online (follow up) increased by 3 percentage points from 51% before matching to 54% after matching. On the other hand, for face-to-face and ABOS, the percentage of survey questions with median APD greater than 5% reduced by 2 percentage points from 44% before matching to 42% after matching; while for the two online surveys the reduction was by 1 percentage point after matching. This indicates that the matching does not make much difference across the different classifications of the distribution of the mode effects. This is because the effects are evenly distributed over the range of magnitudes.

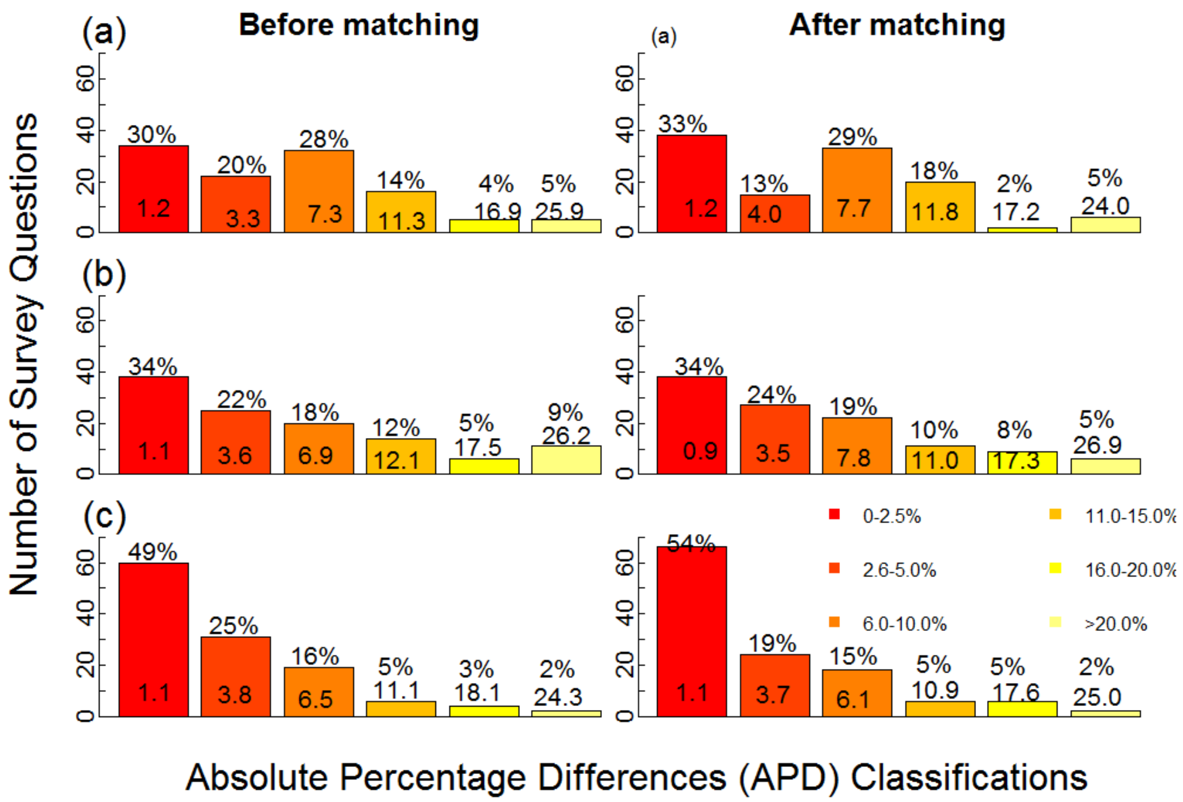


Figure 4-6: Barplots of Absolute Percentage Differences (APD) classifications with corresponding medians and percentages before and after matching (weighted model) for face-to-face and online (follow up) (a), face-to-face and ABOS (b), and online (follow up) and ABOS (c)

It is important to note that substantial differences in the APD could have remained after matching due to missing important confounders from the vector of matching variables. Socio-demographic variables are generally not strong predictors of respondent selection into different modes, so relying on these characteristics alone may lead to underestimation of the magnitude of selection effects (Vannieuwenhuyze et al., 2017). For this, reason a sensitivity analysis was performed to determine whether incorporating attitudinal variables in the propensity score model reduced the size of the mode effects. The online (follow up) and ABOS samples were used because their attitudinal and behavioural variables were measured in the same mode and the results are presented in Figure 4-7. This is important because controlling for attitudinal and behavioural variables obtained using different modes in propensity score models may remove some mode effects resulting in biased estimates.

Figure 4-7 shows that no improvement was found in the size of the mode effect after incorporating attitudinal and behavioural variables such as respondents' wellbeing, satisfaction with local area, attachment to neighbourhood and loneliness into the matching vector. In fact, the effect size of the median APD differences increased by 0.4 percentage points from 1.9% to 2.3% after inclusion of the attitudinal variables. In addition, the number of respondents that were discarded after controlling for attitudinal variables in propensity score model was higher than the number of unmatched respondents in matched samples that were based only on socio-demographic variables (Appendix C). This is because attitudinal variables may not be good predictors of the selection process in a given mode, because they introduce more variability between groups. In turn, this leads to a higher number of unmatched respondents.

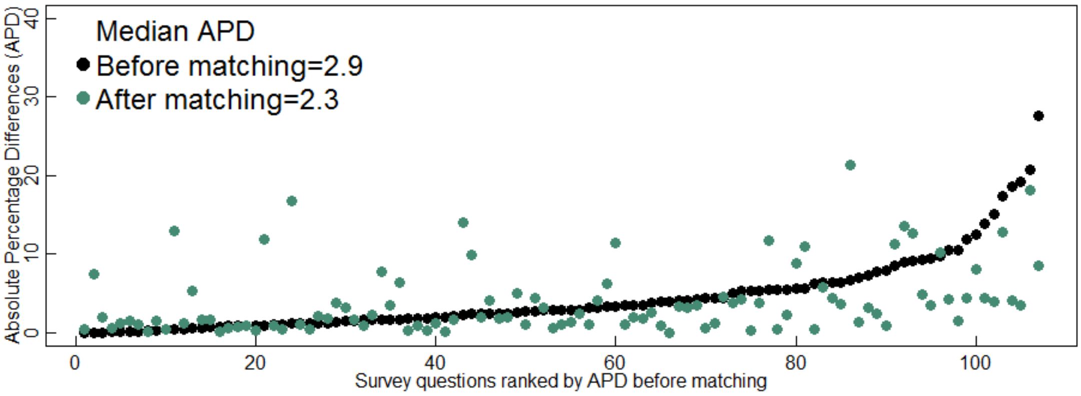


Figure 4-7: Estimated mode effects based by Question before and after matching (weighted model with attitudinal variables) for ABOS and online (follow up).

4.7 Discussion

Face-to-face interviewing has been the backbone of social survey research for many decades, achieving higher response rates, lower item missing data, and longer interviews compared to other modes of data collection (Heerwegh, 2009; Roberts, 2007). However, in recent years many surveys have changed from face-to-face to online administration mostly driven by substantially lower costs associated with online surveys (de Leeuw, 2005; Dillman et al., 2009). However, existing empirical evidence is not clear as to which mode leads to better data quality due to lack of gold standard criterion variables and the confounding of measurement and selection effect (Burkill et al., 2016; de Leeuw, 1992; Tourangeau & Yan, 2007; Vannieuwenhuyze & Loosveldt, 2013; Villar & Fitzgerald, 2017).

It is against this backdrop of uncertainty about data quality between face-to-face interviews and online surveys that the Community Life Survey (CLS) carried out a mode comparison study in 2014 (Williams, 2017b). Williams (2017b) concluded that the address based online probability sample with a low response rate produced data with lower net error compared to a high response rate face-to-face interview survey. However, given the longstanding consensus in survey research on the superiority of face-to-face interviewing, this must be considered a surprising conclusion. For this reason, the motivation for this paper has been to assess whether Williams' conclusion is reasonable by reanalysing the Community Life Survey mixed mode study using a different methodological approach. To be clear, the contention in this paper is not that Williams' (2017b) analysis is flawed. Rather, the aim is to assess whether Williams' (2017b) key conclusion is robust to comparing the mixed-mode samples using propensity score matching, an efficient estimator of mode effects with lower MSE compared to weighting methods (Ertefaie & Stephens, 2010; Hahn, 1998; Hirano et al., 2003). Moreover, PSM is an appropriate choice for making this assessment as it has previously been shown to be an effective method for removing selection effects in mixed modes (Lugtig et al., 2011)

The findings lead to the conclusion that the majority of the total mode effect between the online and face-to-face surveys is due to measurement rather than selection effects, on the assumption that matching successfully removes selection effects. The results show large differences between face-to-face and online (follow up) samples. A direct comparison between face-to-face and ABOS samples also found large mode differences. Smaller differences were found between the two online samples, an analysis which was not conducted by Williams (2017b). The large differences between face-to-face and the two online surveys, and the fact that matching makes little difference to the size of the mode effects suggests that the differences are primarily due to measurement rather than selection effects. This is consistent with the conclusions of Williams (2017b). However, this conclusion is subject to the caveat that the matching did not remove all selection effects, because of the differences in the two online samples after matching. The results clearly showed that neither of the two online surveys comes close to the face-to-face interview after matching. Therefore, it is not possible to conclude from this evidence that the online surveys produced a higher quality data compared to the face-to-face mode. Taking a closer look at face-to-face and online samples, it can be seen that in some instances the total mode effects increased after matching, an indication that selection and measurement effects counteract each other for some variables (Schouten et al., 2013; Tourangeau, 2017).

The second conclusion relates to the utility of PSM for removing selection effects from surveys administered in different modes (Lugtig et al., 2011). The findings demonstrate that PSM cannot be assumed to remove selection effects in all contexts. This is because differences are observed between the two online surveys after matching. This means that matching and by implication the weighting approach used by Williams (2017b) does not completely remove selection differences. Unfortunately, it is not possible to ascertain the degree of the selection effects that remain after matching using a statistical test. This is because the remaining selection effects are confounded with measurement effects. However, the fact that the matching makes little difference between face-to-face and online samples does suggest that a larger part of the differences in APD is due to measurement effects. It was also found that it does not make any difference to the estimated mode effect, whether or not survey weights were incorporated in the estimation of propensity score models and outcome analysis. This is likely because survey weights and are computed using socio-demographic variables and therefore provide similar information as other covariates when controlled for in propensity score models. In addition, selection probabilities are usually computed using sampling design variables such as Government Office Region (GOR) which is controlled for in the propensity score model.

The results in this study lead to three implications for survey practice. First, the approach that has been implemented here indicates that there are substantial mode differences between online probability and random face-to-face interviews after matching. Therefore, these data provide more evidence that survey designers have to be cautious in switching a survey from one mode to another. Second, PSM needs further optimisation for effective removal of selection effects in mixed-mode designs. This might be achieved by incorporating variables from the sampling frame in the propensity score model since they are unaffected by the choice of the mode. However, their effectiveness depends on how strongly they predict selection process to a given mode, which is most unlikely. It is possible that the assumption that socio-demographic variables are unaffected by the choice of the mode and are measured without error may be wrong. Third, the differences between the two online surveys suggest that it is necessary for survey designers to further explore the sources of these data quality differences since the mode is the same.

While different formulations of propensity score models and the estimation of mode effects using APDs have been extensively explored, this study has some limitations. First, the pattern of the results obtained is difficult to interpret because there is no reference criterion against which to benchmark accuracy. Although it can be assumed that online surveys provide superior measurement quality for attitudinal and behavioural questions due to lower social desirability effects, this is just an assumption. It would require additional empirical evidence to properly justify the conclusion that the ABOS survey provides equally good, or even better-quality data than the face-to-face survey. Therefore, future studies should assess whether it is possible to predict the size of the mode effect after matching as a function of question characteristics, such as their susceptibility to social desirability bias and satisficing. It might equally be the case that the mean squared errors (MSE) in the online samples are larger than the face-to-face sample due to a combination of nonresponse bias and measurement errors, and these errors are similar to one another in the online surveys. In summary, the pattern of results in this study are not only consistent with Williams (2017b) conclusion, but they are also consistent with a conclusion that the face-to-face data has lower MSE than both online samples.

Second, this analysis focuses on a single study that asked attitudinal and behavioural questions from the target population of United Kingdom residents. To generalise the conclusions of this study into other contexts and countries would need more evidence. Therefore, this study may be replicated using mixed-mode surveys investigating different societal issues, surveys from other countries, and with better predictors relative to those applied in this study. Additionally, the scope of this study may be limited because of the unobserved characteristics not controlled for in the propensity score model. It would be important if future research attempts to identify reliable baseline covariates on which to base propensity scores that can fully account for any selection effects in mixed-mode surveys. Third, to counter for the higher number of unmatched respondents in some modes, matching with replacement approach should be applied rather than the matching without replacement approach used in this study. Fourth, the direction of measurement errors produced by different modes was not considered. The knowledge of the direction of measurement effects may inform whether errors in one mode can offset those in another leading to the overall reduction of mode effects. Fourth, matching without replacement resulted in some respondents not within defined callipers been discarded after matching. This may be fixed by using matching with replacement approach which result to matched samples with less variability between groups. These limitations present other potential areas of future research.

Chapter 5 Conclusion

In recent years the survey landscape has been transforming rapidly because of a combination of declining response rates, increasing numbers of survey requests, technological change, and increasing survey costs. This may have adversely affected survey quality because known error sources are becoming more complex and new error structures are emerging. Nevertheless, survey research remains the bedrock of social scientific research in different areas through which key public policies and business decisions are made. Therefore, an investigation aimed at improving understanding of factors that influence survey quality is of crucial importance.

Survey quality can be considered in the context of the Total Survey Error (TSE) framework (Biemer, 2010; Groves, 1989). The TSE framework evaluates survey quality by identifying major sources of errors at each stage of the survey process and allocating survey resources to reduce such errors within budgetary and time constraints. However, survey errors are inter-related and a reduction in one error may actually increase other errors. Also, it is difficult to adopt a single strategy for reducing TSE because the relative importance of errors varies across surveys and uses. This has promoted survey researchers to conduct evaluation studies with the aim of understanding and quantifying the various sources of errors depending on users' requirements and the changing survey environment. This thesis has focused its attention on the two main survey errors: nonresponse error and measurement error. It also assessed factors that influence TSE such as interviewers, incentives and modes of data collection.

This thesis made both methodological and substantive contributions in three distinct but related papers. Paper 1 explored whether the predictions of survey response propensity models may be improved by using informative priors in a Bayesian framework, derived from previous wave data in a longitudinal context. Paper 2 investigated the role of interviewers in determining whether incentives are effective in improving response and cooperation rates. Lastly, Paper 3 provided both methodological and substantive contributions by assessing whether a low response rate online survey produces data of equivalent or better quality to a face-to-face survey, while adjusting for selection effects using propensity score matching.

This chapter provides a summary of the main findings from the three papers, discusses their implications for survey practise, and presents suggestions for future work.

5.1 Summary of Key Findings

A variety of data sources and methods were employed to address the main research questions of the thesis. Five datasets were used for the analysis. Paper 1 used data from the first five waves of UK household longitudinal study (Understanding Society). Three datasets were used in Paper 2 namely: 2015 National Survey for Wales Field Test (NSW 2015), 2016 National Survey for Wales Incentive Experiment (NSW 2016), and Wave 1 of the UK Household Longitudinal Study Innovation Panel (UKHLS-IP). Paper 3 used a data from the Community Life Survey (CLS) which had three different samples namely: face-to-face, online (follow up) and address based online surveying (ABOS). The methodologies employed to examine and understand survey quality include: response propensity models using Bayesian approach in Paper 1, multilevel modelling in Paper 2, and propensity score matching in Paper 3.

This thesis addressed gaps in the existing literature on survey quality by seeking to answer the following research questions:

1. Does the use of informative priors based on previous wave data in a Bayesian framework improve predictions of survey response propensities in a longitudinal survey?
2. Do interviewers moderate the effect of monetary incentives on response and cooperation rates in household interview surveys?
3. Do low response rate online surveys provide better quality data than high response face-to-face interviews?

The first research question was addressed in Paper 1 and provided a methodological contribution by evaluating whether specification of informative priors based on previous wave information in longitudinal data improves the predictive and discrimination power of survey response predictions. The performance of different response propensity models was assessed using a range of evaluation criteria.

The analyses provided a clearer understanding of the use of informative priors derived from previous wave data in response propensity models for longitudinal surveys and made a methodological contribution to the literature. In the analysis, vague priors were used as the benchmark in which no previous wave information was incorporated. The results showed only a slight improvement in model fit when previous wave information was incorporated in response propensity models as informative priors. In addition, measures of classification, discrimination, prediction and AUC showed minimal gains in terms of discriminating the accuracy of survey response predictions. The gain in predictive and discriminative power in survey response predictions was more evident in earlier waves of the analyses and diminished in later waves. This indicates that timeliness of the previous wave data is of crucial importance when used to derive informative priors. It was also observed that altering the variance components of the informative priors with an aim of moderating the strength of information borrowed from previous wave data did not have any impact in the range of the predictive and discrimination measures obtained. This indicated that information borrowed from previous wave data was less relevant and dominant compared to the amount of contribution made by the data in the likelihood component of the model. The low predictive strength of borrowed information from previous wave data may also be informed by the extent to which auxiliary variables were correlated with key survey response outcomes in response propensity models. In most instances auxiliary variables are not strongly correlated with key survey outcomes which may negatively impact the strength of priors derived from previous wave data.

In order to address the second research question, multilevel modelling was applied to response outcome data across three face-to-face surveys in paper 2. The findings suggested that interviewers vary significantly in how effective they were at using incentives to increase response and cooperation rates in face-to-face interviews. Surprisingly, none of the interviewer characteristics considered (age, gender, and experience) significantly explained the between interviewer variability in the effectiveness of incentives observed across the three surveys. The cross-level interactions of interviewer characteristics and incentives were also not significant indicating that they did not moderate interviewer variability in incentive deployment. This shows that other factors such as interviewer attitudes, personalities and behaviours which were not included in the models due to data not being available may be influencing interviewers' variability in deployment of incentives. The results also showed that exchanging interviewers from the top to the bottom of decile performance in terms of incentives performance will on average increase the effect of incentive relative to no incentive condition by a 15-percentage point on response rates. It was also found that interviewers who performed better in gaining response and cooperation rates were not more effective in deployment of incentives. Lastly, there was no evidence of differential effect of interviewers on cooperation relative to nonresponse.

The third research question had two complementary objectives in Paper 3, one methodological and one substantive. The first substantive objective showed that the majority of total mode effects between the online and face-to-face surveys is due to measurement rather than selection effects. The larger differences found between face-to-face and online surveys and the fact that matching made little difference reinforced the conclusion that differences were due to measurement rather than selection effects. This finding was consistent with results obtained by Williams (2017b). However, this is conclusion should be taken cautiously because matching did not remove all selection effects as evidenced by the smaller differences in the two online samples. The results also showed that neither of the two online surveys was similar to the face-to-face interview after matching. This implies that it was not possible to conclude that the online surveys provided equal or better data quality than higher response rate face-to-face interviews. The assumption that online surveys can provide superior measurement quality due to lower social desirability than face-to-face surveys, is just an assumption which requires additional empirical evidence to properly justify. The second objective was addressed by assessing how effective propensity score matching approach was in removing selection effects and whether different formulations of survey weights in propensity score models and outcome analysis had an impact in the estimation of mode effects. The results showed that propensity score matching cannot be assumed to be a completely effective method for removing selection effects in surveys with different modes of data collection. This can be explained by differences that remained even after matching the two online surveys. Normally, the differences between the data collected using similar modes is expected to be close to zero, due to minimal measurement differences. It was not possible to ascertain the degree of selection effects that remained after matching because they were confounded with measurement effects. Specification of different formulations of survey weights in propensity score models and outcome analysis were found to have no impact on the estimates of mode effects. This may be explained by the fact that survey weights added no additional power in the estimation of propensity scores and outcome analysis because they were estimated using the same socio-demographic variables controlled for in propensity score models.

5.2 Survey Practice Implications

The results presented in the thesis have important survey practice implications. The results from Paper 1 contribute to a better understanding of the use of previous wave data as informative priors in response propensity models. This is especially useful in adaptive and responsive survey designs where survey data collection process entails regular monitoring and adjustments. In addition, the results indicate that it may be important to consider timeliness and the amount of previous wave data when eliciting priors based on previous wave data. In summary, these findings open a new framework for exploration of other sources of informative priors for response propensity models.

The results from Paper 2 had two main implications for survey practice. First, the approach implemented in this study can be used to identify underperforming interviewers. This approach could be applied in responsive design as a way of identifying interviewers who miss their fieldwork targets and as means of providing an understanding of the strategies employed by more successful interviewers. This will help improve recruitment and training of interviewers especially on approaches of recognising and heightening the saliency of incentives in surveys. Secondly, interviewers at the top end of the performance distribution may be encouraged to share their ideas and practises with poorly performing interviewers. This will steer underperforming interviewers in the right direction in terms of mediating the effects of incentives on survey response and cooperation and reduce wastage on money spent on incentives.

The results of the analysis in the third paper had three main substantive implications. First, survey designers need to be careful when switching from costly face-to-face interviews to more affordable online surveys. This is because of the substantial mode differences observed between online and face-to-face surveys. For online surveys to continue fulfilling the mission of contributing to public policy and business decisions that has over the years relied on face-to-face surveys it is important to first reduce the observed differences between the two modes before any switch. This may be achieved by optimising the strengths of each mode to the minimum affordable difference given the budgetary and time constraints. Once compelling evidence is attained that the required data quality has been achieved between the two different modes, then a switch to the lower-cost alternative is merited and data obtained will be deemed reliable. Second, propensity score matching requires further optimisation and improvement to effectively remove selection effects in mixed-mode surveys. The first step of making propensity matching approach more effective may involve accounting for variables that are robust at correcting for selection bias in propensity score models. Therefore, survey methodologists should aim to explore ways of obtaining robust auxiliary variables that are unaffected by the choice of the mode and predictive of the selection process to control for in propensity score models. Third, the findings suggest that data collected using online surveys may be susceptible to data quality issues because of the differences observed when the two online surveys were compared. Therefore, survey designers should explore the sources of these differences in online surveys and how to control for them.

5.3 Limitation of the Research

Although a detailed exploration of the use of previous wave data as informative priors in the context of longitudinal studies was conducted in Paper 1, this study had some limitations. Firstly, the analysis did not account for correlations among regression parameters because it is computationally demanding to incorporate the covariance structure for informative priors due to the many covariates controlled for in the models, resulting in a higher dimensionality. Accounting for correlation structures could have allowed a wide range of sensitivity analyses on the impact of informative priors on predictions of survey response. Secondly, the data generating mechanism in Understanding Society was not constant over time due to responsive designs which may have negatively impacted the strength and relevance of informative priors in later waves. It is also crucial to note that sequential Bayesian updating only allows for inclusion of auxiliary variables in subsequent waves if they were controlled for in the first wave analysis. Therefore, it was not possible to control for additional survey auxiliary variables available in subsequent waves. Updating response propensity models with new auxiliary variables available in later waves could have potentially influenced the power of response predictions.

The data used for the analysis in Paper 2 only allowed a limited range of characteristics in sample units, areas, and interviewers to be controlled for. It is likely that interviewers' variability in deployment of incentives may have reduced if there were stronger controls of differences between sample units, interviewers and areas. Secondly, caution should be exercised when generalising the findings from this study further because the surveys considered had a narrow range of incentives values which were administered to all households in the incentive condition. This may have reduced the effect of the overall incentives in terms of improving response and cooperation rates. The relevance of this study findings applies only to face-to-face surveys and not to online surveys which have been on the rise in recent years.

Despite the thorough investigation of data quality between face-to-face interviews and online surveys in Paper 3, this study also had some limitations. First, it was not possible to tell whether online surveys provide data of equivalent or better quality than face-to-face surveys. This is because the pattern of the results obtained were quite difficult to interpret since there was no reference criterion against which to assess the most accurate mode. Secondly, more evidence is required before generalising the analyses from this study into other contexts and countries. The results from this study are based on one study that asked attitudinal and behavioural questions of the target population of United Kingdom residents only. This study would be of more benefit if more variables that have direct effect on the mode of data collection assignment are either collected or obtained from other sources such as administrative data and included into the analysis. Finally, this study did not investigate whether errors in different modes (i.e. direction of measurement effects) offset each other leading to overall reduction of the mode effects.

5.4 Future Research

The work presented here can be expanded in a number of ways. Response propensity models fitted using Bayesian approach in Paper 1 were restricted to incorporating informative priors derived from previous wave data. Future studies should consider using informative priors derived from monthly or quarterly data which may be less susceptible to changes in data generating mechanism and timeliness of the data. In addition, other sources of informative priors based on experts' knowledge should be considered in future work.

For Paper 2, the following issues should be addressed in future studies. First, the analysis approach considered in Paper 2 should be replicated using a broad range of incentive values and in other countries. This would provide a clear understanding of how interviewers influence deployment of incentives in a more generalisable way. Second, future studies should consider controlling for sample units' characteristics obtained from external sources such as registers and administrative data. Controlling for such variables may reduce the magnitude of the interviewer effects in deployment of incentives observed. Third, future studies should collect variables measuring interviewer attitudes, beliefs, behaviours, and personalities. Inclusion of these variables into models might explain why some interviewers are more effective in deploying incentives than others.

As shown in paper 3 mode differences exist between face-to-face interviews and online probability surveys when only focusing on behavioural and attitudinal questions from target population of United Kingdome residents. Therefore, this study should be replicated in different contexts to obtain more conclusive and generalisable results. Also, to justify that online surveys produce data of better quality due to lower social desirability bias than face-to-face surveys for attitudinal and behavioural questions; future studies should attempt to predict the size of the mode effect as a function of question characteristics after matching. Future studies may also consider identifying baseline covariates on which to base propensity scores that can fully account for any selection effects in mixed-mode surveys and are unaffected by the choice of the mode. These baseline covariates can be obtained from sources such as sampling, frame, administrative data, and registers. Their inclusion in propensity score models may lead to an increase in the precision of the estimated mode effects without increasing bias. In addition, to reduce the variability that may arise due to the number of unmatched respondents between modes, future studies should consider using matching with replacement approach. Different modes of data collection may produce measurement errors with different directions which may affect the overall mode effects. Thus, future studies should consider estimating mode effects while incorporating the direction of measurement errors as this may inform whether errors in one mode can offset those in another leading to overall reduction in mode effects.

Appendix A [Paper 1]

A.1 Descriptive

Timetable for data collection for Wave 1 to 5 by quarter (Q) 2009-2014

2009				2010				2011				2012				2013				2014			
Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Wave 1	🟩	🟩	🟩	🟩	🟩	🟩	🟩																
				Wave 2	🟩	🟩	🟩	🟩	🟩	🟩	🟩												
								Wave 3	🟩	🟩	🟩	🟩	🟩	🟩	🟩								
												Wave 4	🟩	🟩	🟩	🟩	🟩	🟩	🟩				
																Wave 5	🟩	🟩	🟩	🟩	🟩	🟩	🟩

A year is made up of four quarters: January-March is (Q1); April-June is (Q2); July-September is (Q3); and October-December is (Q4)

Variable names and their corresponding categorical levels considered in the analysis

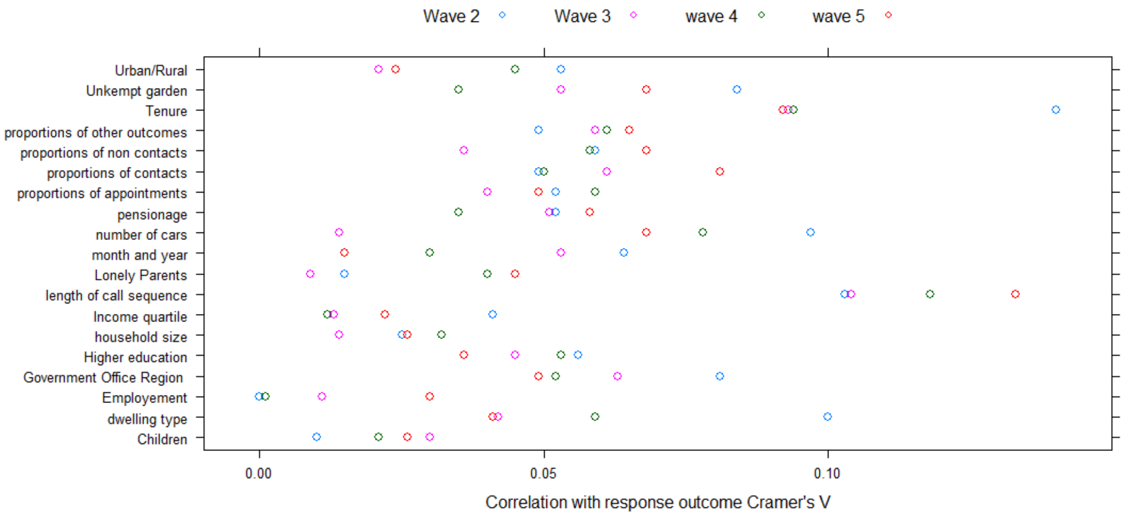
Categories	*Variable name*	*Response Categories*
Geographical Information and Design Variables	Government Office Region	East Midlands (reference)
		East of England
		London
		North East
		North West
		Scotland
		South East
		South West
		Wales
		West Midlands
		Yorkshire and the Humber
	Urban indicator	Urban(reference)
		Rural
	Month and year of household issues	January-June Year 1 (reference)
		July-December Year 1
		January-June Year 2
		July- December Year 2
Survey Variables	Lone parents	lone parents in household (reference)
		no lone parents in household
	Pensioners in household	no pension age people in household (reference)
		pension age people in household
	Employment status	employed people in household (reference)
		No employed people in household
	Number of cars	no car (Reference)
		one car
		two cars
		three or more cars
	Highest education qualification	Higher degree & Degree (reference)
		A level & GCSE
		Others or none qualification
	Household income	1st Quartile (reference)

		2nd Quartile
		3rd Quartile
		4th Quartile
	Tenure	Owned (reference)
		Rented from employer privately and other
		Rented from LA or housing association
	Household size	One person (reference)
		2 to 3 persons
		More than 4 persons
Interviewer observations	Type of accommodation/ dwelling type	house/bungalow (reference)
		flat/maisonette
		other
	Relative condition of property	Mainly good (reference)
		Mainly fair
		Mainly bad and others
	Presence of unkempt garden in address	yes (Reference)
		no
		No obvious garden
		Don’t know
	Condition of surrounding houses (vicinity 1)	Mentioned (reference)
		Not mentioned
	Trash, litter or junk in street or road (vicinity 2)	Mentioned (reference)
		Not mentioned
	Heavy traffic on street or road (Vicinity 3)	Mentioned (reference)
		Not mentioned
	Presence of car or van	No (reference)
		yes, probably belonging to this address
		yes, unsure whether belonging to this address
		don't know
	Presence of children in a household	children in household (reference)
		no children in household
Call record variables	Length of call sequence	short (Reference)
		long
	Proportion of noncontacts	0-25% (Reference)
		26-50%
		51-75%
		76-100%
	Proportion of appointments	0-25% (Reference)
		26-50%
		51-100%
	Proportion of contact calls	0-25% (Reference)
		26-50%
		51-75%
		76-100%
	Proportion of other call outcomes (ineligibles and refusals)	0-25% (Reference)
		26-50%
		51-75%
		Any other status
	Proportion of interviews	0-25% (Reference)
		26-50%
		51-75%
		76-100%

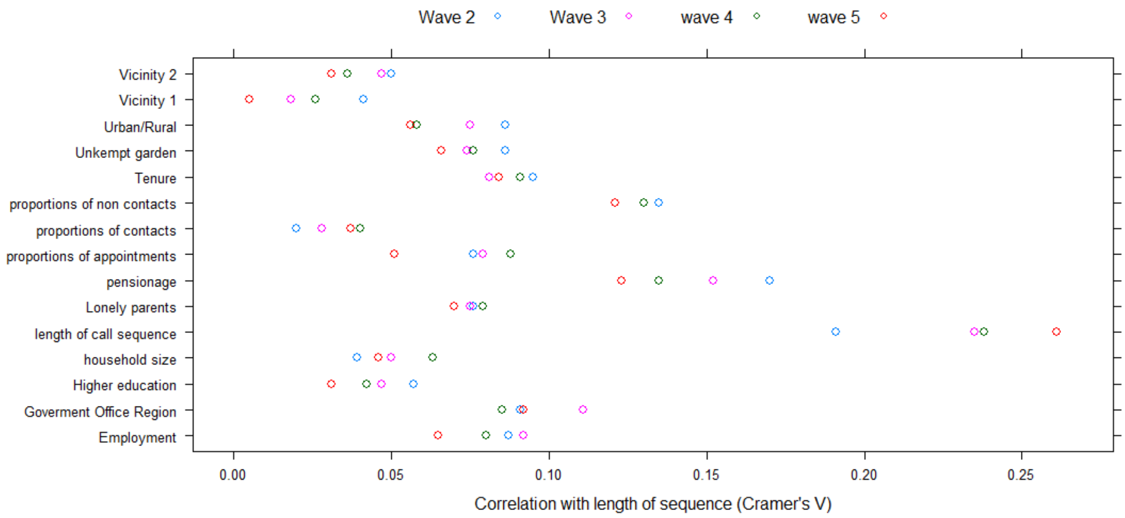
Explanatory variables for the final call outcome and length of call sequence models

Final call outcome	*Length of call sequence*
Government Office Region	Government Office Region
Urban indicator	Urban indicator
Month and year of household issues	Pensioners in household
Pensioners in household	Lone parents
Employment status	Employment status
Highest education qualification	Highest education qualification
Household income	Tenure
Tenure	Presence of unkempt garden in address
Type of dwelling type	Condition of surrounding houses (vicinity 1)
Presence of children in a household	Trash, litter or junk in street or road (vicinity 2)
Number of cars	Household size
Household size	Proportion of contacts
Proportion of other call outcomes (ineligibles and refusals)	Proportion of noncontacts
Proportion of contact calls	
Proportion of appointments	
Proportion of noncontacts	
Length of call sequence	

A.2 Bivariate Correlations between Response Outcomes and Auxiliary Variables for main dataset



Bivariate correlations between the final call outcome and auxiliary variables



Bivariate correlations between length of call sequence and auxiliary variables

A.3 Length of Call Sequence Results

Evaluation criteria for frequentist models using (Akaike information Criteria (AIC), Nagelkerke's pseudo R2 and Watanabe Akaike Information Criteria (WAIC)) for Bayesian models

Wave	*Model*	*AIC*	*Nagelkerke R^2 (%)*	*WAIC*
1 and 2	frequentist	9766.50	12.30	-
	M1	-	-	9767.10
2 and 3	frequentist	6702.10	13.10	
	M1	-	-	6849.61
	M2	-	-	6700.66
	M3	-	-	7105.74
	M4	-	-	6695.09
	M5	-	-	6699.91
	M6	-	-	6701.53
	M7	-	-	6711.57
3 and 4	frequentist	5352.30	13.50	
	M1	-	-	5353.33
	M2	-	-	5381.91
	M3	-	-	5816.85
	M4	-	-	5346.97
	M5	-	-	5349.50
	M6	-	-	5351.75
	M7	-	-	5353.30
4 and 5	frequentist	4487.60	12.73	
	M1	-		4489.53
	M2	-	-	4478.02
	M3	-	-	4483.82
	M4	-	-	4475.08
	M5	-	-	4483.82
	M6	-	-	4487.57
	M7	-	-	4489.51

Classification table and AUC of ROC curves for the length of call sequence

Wave	*Model*	*Classification (%)*	*AUC (%)*
1 and 2	frequentist	83.1	69.1
	M1	83.1	69.1
2 and 3	frequentist	87.8	71.8
	M1	87.8	71.8
	M2	87.8	72.0
	M3	87.9	64.4
	M4	87.8	72.1
	M5	87.8	72.1
	M6	87.8	72.1
	M7	87.8	71.8
3 and 4	frequentist	89.1	71.8
	M1	89.1	71.7
	M2	89.1	70.8
	M3	89.3	58.7
	M4	89.1	71.8
	M5	89.1	71.8
	M6	89.1	71.8
	M7	89.1	71.8
4 and 5	frequentist	90.7	74.3
	M1	90.7	74.3
	M2	90.7	74.1
	M3	90.7	74.3
	M4	90.7	74.3
	M5	90.7	74.3
	M6	90.7	74.3
	M7	90.7	74.3

Results of classification table and AUC of ROC curves, sensitivity, specifity, positive predictive values (PPV) and negative predictive values (NPV) for the length of call sequence

Wave	*Modelling approach*	*Sensitivity*	*Specificity*	*PPV*	*NPV*
1 and 2	frequentist	39.0	83.0	2.0	99.0
	M1	40.0	83.0	2.0	99.0
2 and 3	frequentist	43.0	88.0	4.0	99.0
	M1	43.0	88.0	4.0	99.0
	M2	43.0	88.0	3.0	99.0
	M3	NaN	88.0	0.0	100.0
	M4	46.0	88.0	4.0	99.0
	M5	44.0	88.0	4.0	99.0
	M6	43.0	88.0	4.0	99.0
	M7	46.0	88.0	4.0	99.0
3 and 4	frequentist	38.0	90.0	3.0	99.0
	M1	38.0	90.0	3.0	99.0
	M2	40.0	90.0	3.0	99.0
	M3	Nan	0.89	0.0	100.0
	M4	40.0	90.0	3.0	99.0
	M5	38.0	90.0	3.0	99.0
	M6	38.0	90.0	3.0	99.0
	M7	38.0	90.0	3.0	99.0
4 and 5	frequentist	53.0	91.0	5.0	100.0
	M1	52.0	91.0	5.0	100.0
	M2	48.0	91.0	4.0	100.0
	M3	52.0	91.0	5.0	100.0
	M4	51.0	91.0	4.0	100.0
	M5	52.0	91.0	5.0	100.0
	M6	53.0	91.0	5.0	100.0
	M7	53.0	91.0	5.0	100.0

A.4 A subsample consisting of 10% of main sample

Wave		**Final Call Outcome**		**Length Call Sequence**		**Total**
		At least one interview	*No interview*	*Short Sequence (1-6 calls)*	*Long sequence*	
2	Frequency	1,882	563	2,077	368	2445
	Percentage	77.0	23.0	84.9	15.1	
3	Frequency	1553	365	1675	243	1918
	Percentage	81.0	19.0	87.3	12.7	
4	Frequency	1507	230	1541	196	1737
	Percentage	86.8	13.2	88.7	11.3	
5	Frequency	1425	188	1453	160	1613
	Percentage	88.3	11.7	90.1	9.9	

Evaluation criteria for frequentist and Bayesian models (Akaike information Criteria (AIC), Nagelkerke's pseudo R^2 and Watanabe Akaike Information Criteria (WAIC))

Wave	Model	Final call outcome			Length of call sequence		
		AIC	*Nagelkerke R^2 (%)*	*WAIC*	*AIC*	*Nagelkerke R^2 (%)*	*WAIC*
1 and 2	frequentist	1221.1	8.31	-	865.4	11.9	-
	M1	-	-	1221.85	-	-	865.28
2 and 3	frequentist	952.77	3.98		697.4	10.2	
	M1	-	-	953.56	-	-	697.77
	M2	-	-	937.58	-	-	691.81
	M3			944.33	-	-	786.32
	M4	-	-	941.67	-	-	693.94
	M5	-	-	948.95	-	-	696.81
	M6	-	-	952.06	-	-	697.51
	M7	-	-	953.55	-	-	697.77
3 and 4	frequentist	686.41	6.49		557.08	12.26	
	M1			687.91	-	-	557.41
	M2			677.94	-	-	558.23
	M3			703.00	-	-	714.22
	M4			675.92	-	-	554.34
	M5	-	-	676.38	-	-	555.73
	M6	-	-	676.23	-	-	556.88
	M7	-	-	684.05	-	-	557.41
4 and 5	frequentist	556.2	11.08	-			-
	M1	-	-	557.96	-	-	516.16
	M2	-	-	544.63	-	-	513.04
	M3	-		545.42	-	-	512.80
	M4	-	-	544.48	-	-	512.80
	M5	-	-	550.20	-	-	515.01
	M6	-	-	554.75	-	-	515.82
	M7	-	-	557.93	-	-	516.65

Appendix A

Classification table and AUC of ROC curves for the final call outcome and length of call sequence.

Wave	**Modelling approach**	**Final call outcome**		**Length of call sequence**	
		Classification (%)	*AUC (%)*	*Classification (%)*	*AUC (%)*
1 and 2	frequentist	78.1	64.6	84.3	68.7
	M1	78.0	64.6	84.3	68.7
2 and 3	frequentist	80.8	59.2	87.5	68.7
	M1	80.8	59.2	87.5	68.7
	M2	80.8	56.4	87.2	68.2
	M3	80.8	55.1	87.4	57.6
	M4	80.8	58.1	87.3	68.6
	M5	80.8	59.1	87.5	68.7
	M6	80.8	59.1	87.5	68.7
	M7	80.8	59.2	87.5	68.7
3 and 4	frequentist	86.8	62.0	88.1	73.6
	M1	86.8	61.9	88.1	73.6
	M2	86.8	61.9	88.1	73.8
	M3	86.3	61.3	88.1	59.7
	M4	86.6	62.9	88.1	74.4
	M5	86.6	62.4	88.0	73.9
	M6	86.6	62.1	88.1	73.7
	M7	86.8	61.9	88.1	73.6
4 and 5	frequentist	88.0	64.4	90.6	70.9
	M1	88.0	64.4	90.6	71.1
	M2	88.0	60.9	91.2	73.3
	M3	87.7	58.7	91.2	72.7
	M4	87.7	63.5	91.2	72.7
	M5	87.7	64.5	90.7	71.6
	M6	87.7	64.3	90.6	71.2
	M7	88.0	64.4	90.6	71.1

Results of sensitivity, specifity, positive predictive values (PPV) and negative predictive values (NPV) of the two binary responses.

Wave	Modelling approach	Final call outcome				Length of call sequence			
		Sensitivity	*Specificity*	*PPV*	*NPV*	*Sensitivity*	*Specificity*	*PPV*	*NPV*
1 and 2	frequentist	52.0	79.0	4.0	99.0	46.0	85.0	3.0	99.0
	M1	50.0	79.0	4.0	99.0	46.0	85.0	3.0	99.0
2 and 3	frequentist	25.0	81.0	1.0	100.0	60.0	88.0	2.0	100.0
	M1	25.0	81.0	1.0	100.0	60.0	88.0	2.0	100.0
	M2	0.0	81.0	1.0	100.0	25.0	87.0	1.0	100.0
	M3	15.0	81.0	1.0	99.0	NaN	87.0	0.0	100.0
	M4	0.0	81.0	0.0	100.0	33.0	87.0	0.0	100.0
	M5	0.0	81.0	0.0	100.0	60.0	88.0	2.0	100.0
	M6	0.0	81.0	0.0	100.0	60.0	88.0	2.0	100.0
	M7	25.0	81.0	1.0	100.0	60.0	88.0	2.0	100.0
3 and 4	frequentist	100.0	87.0	1.0	100.0	50.0	88.0	2.0	100.0
	M1	100.0	87.0	1.0	100.0	50.0	88.0	2.0	100.0
	M2	NaN	87.0	0.0	100.0	NaN	88.0	0.0	100.0
	M3	NaN	87.0	2.0	100.0	50.0	88.0	0.0	100.0
	M4	NaN	87.0	2.0	100.0	50.0	88.0	0.0	100.0
	M5	NaN	87.0	2.0	100.0	33.0	88.0	0.0	100.0
	M6	NaN	87.0	2.0	100.0	50.0	88.0	0.0	100.0
	M7	100.0	87.0	2.0	100.0	50.0	88.0	0.0	100.0
4 and 5	frequentist	75.0	88.0	3.0	100.0	40.0	92.0	8.0	99.0
	M1	75.0	88.0	3.0	100.0	40.0	92.0	8.0	99.0
	M2	50.0	88.0	3.0	100.0	62.0	91.0	7.0	100.0
	M3	50.0	88.0	1.0	100.0	60.0	92.0	8.0	100.0
	M4	50.0	88.0	1.0	100.0	60.0	92.0	8.0	100.0
	M5	67.0	88.0	2.0	100.0	43.0	92.0	8.0	100.0
	M6	57.0	88.0	2.0	100.0	40.0	92.0	8.0	100.0
	M7	75.0	88.0	3.0	100.0	40.0	92.0	8.0	100.0

A.5 A subsample of 5% of main sample

Distributions of the Two Response Variables in the Final Analysis Sample.

Wave		**Final Call Outcome**		**Length Call Sequence**		**Total**
		At least one interview	*No interview*	*Short Sequence (1-6 calls)*	*Long sequence*	
2	Frequency	949	274	1032	191	1223
	Percentage	77.6	22.4	84.4	15.6	
3	Frequency	790	169	833	126	959
	Percentage	82.4	17.6	86.9	13.1	
4	Frequency	754	115	765	104	869
	Percentage	86.8	13.2	88.0	12.0	
5	Frequency	730	77	729	78	807
	Percentage	90.5	9.5	90.3	9.7	

Evaluation criteria for frequentist and Bayesian models (Akaike information Criteria (AIC), Nagelkerke's pseudo R^2 and Watanabe Akaike Information Criteria (WAIC))

Wave	**Model**	**Final call outcome**			**Length of call sequence**		
		AIC	*Nagelkerke R^2 (%)*	*WAIC*	*AIC*	*Nagelkerke R^2 (%)*	*WAIC*
1 and 2	frequentist	1009.6	10.0		812.55	11.37	-
	M1	-	-	1435.25	-	-	812.57
2 and 3	frequentist	698.0	4.5		544.34	12.69	-
	M1	-	-	698.1	-	-	544.35
	M2	-	-	692.2	-	-	556.40
	M3			697.2	-	-	591.94
	M4	-	-	693.2	-	-	544.72
	M5	-	-	696.4	-	-	543.80
	M6	-	-	697.6	-	-	544.18
	M7	-	-	698.1	-	-	544.35
3 and 4	frequentist	538.94	6.4		495.79	11.6	-
	M1			539.49	-	-	496.11
	M2			537.84	-	-	504.13
	M3			541.78	-	-	533.01
	M4			535.93	-	-	495.82
	M5	-	-	537.04	-	-	495.65
	M6	-	-	538.39	-	-	495.67
	M7	-	-	539.48	-	-	496.11
4 and 5	frequentist	407.04	8.23	-	387.6	10.58	-
	M1	-	-	407.26	-	-	387.67
	M2	-	-	407.68	-	-	393.26
	M3	-		411.09	-	-	386.86
	M4	-	-	405.74	-	-	386.55
	M5	-	-	406.08	-	-	387.61
	M6	-	-	406.36	-	-	387.67
	M7	-	-	407.25	-	-	387.67

Classification table and AUC of ROC curves for the final call outcome and length of call sequence.

Wave	Model	Final call outcome		Length of call sequence	
		Classification (%)	*AUC (%)*	*Classification (%)*	*AUC (%)*
1 and 2	frequentist	78.8	56.0	86.9	71.5
	M1	78.8	56.0	86.9	71.5
2 and 3	frequentist	79.2	57.1	85.4	64.8
	M1	79.2	57.1	85.4	64.9
	M2	79.2	58.8	85.4	67.3
	M3	79.2	55.2	85.4	61.0
	M4	79.2	57.1	85.4	67.5
	M5	79.2	57.5	85.4	65.6
	M6	79.2	57.3	85.4	65.4
	M7	79.2	57.1	85.4	64.9
3 and 4	frequentist	86.2	58.6	90.8	73.1
	M1	86.2	58.0	90.8	73.1
	M2	86.2	55.9	90.8	74.7
	M3	86.2	52.8	90.8	67.0
	M4	86.2	56.1	90.8	73.2
	M5	86.2	58.2	90.8	73.1
	M6	86.2	58.3	90.8	73.1
	M7	86.2	58.0	90.8	73.1
4 and 5	frequentist	89.4	67.4	90.1	67.4
	M1	89.4	67.2	90.1	67.4
	M2	90.7	66.2	90.1	67.0
	M3	90.7	58.9	90.1	67.0
	M4	90.7	67.9	90.1	67.0
	M5	90.7	67.0	90.1	67.5
	M6	90.1	67.3	90.1	71.0
	M7	89.4	67.3	90.1	67.4

Results of sensitivity, specifity, positive predictive values (PPV) and negative predictive values (NPV) of the two binary responses

Wave	Model	Final call outcome				Length of call sequence			
		Sensitivity	Specificity	PPV	NPV	Sensitivity	Specificity	PPV	NPV
1 and 2	frequentist	25.0	80.0	2.0	98.0	NaN	0.87	0.0	100.0
	M1	25.0	80.0	2.0	98.0	NaN	0.87	0.0	100.0
2 and 3	frequentist	NaN	79.0	0.0	100.0	NAN	0.85	0.0	100.0
	M1	NaN	79.0	0.0	100.0	NaN	0.85	0.0	100.0
	M2	NaN	79.0	0.0	100.0	NaN	0.85	0.0	100.0
	M3	NaN	79.0	0.0	100.0	NaN	0.85	0.0	100.0
	M4	NaN	79.0	0.0	100.0	NaN	0.85	0.0	100.0
	M5	NaN	79.0	0.0	100.0	NaN	0.85	0.0	100.0
	M6	NaN	79.0	0.0	100.0	NaN	85.4	0.0	100.0
	M7	NaN	79.0	0.0	100.0	NaN	85.4	0.0	100.0
3 and 4	frequentist	NaN	86.0	0.0	100.0	NaN	91.0	0.0	100.0
	M1	NaN	86.0	0.0	100.0	NaN	91.0	0.0	100.0
	M2	NaN	86.0	0.0	100.0	NaN	91.0	0.0	100.0
	M3	NaN	86.0	0.0	100.0	NaN	91.0	0.0	100.0
	M4	NaN	86.0	0.0	100.0	NaN	91.0	0.0	100.0
	M5	NaN	86.0	0.0	100.0	NaN	91.0	0.0	100.0
	M6	NaN	86.0	0.0	100.0	NaN	91.0	0.0	100.0
	M7	NaN	86.0	0.0	100.0	NaN	91.0	0.0	100.0
4 and 5	frequentist	0.0	91.0	0.0	99.0	NaN	90.0	0.0	100.0
	M1	0.0	91.0	0.0	99.0	NaN	90.0	0.0	100.0
	M2	NaN	91.0	0.0	99.0	NaN	90.0	0.0	100.0
	M3	NaN	91.0	0.0	99.0	NaN	90.0	0.0	100.0
	M4	NaN	91.0	0.0	99.0	NaN	90.0	0.0	100.0
	M5	NaN	91.0	0.0	99.0	NaN	90.0	0.0	100.0
	M6	0.0	91.0	0.0	99.0	NaN	90.0	0.0	100.0
	M7	0.0	91.0	0.0	99.0	NaN	90.0	0.0	100.0

A.6 A subsample of 2 % of main sample

Distributions of the Two Response Variables in the Final Analysis Sample.

Wave		Final Call Outcome		Length Call Sequence		Total
		At least one interview	*No interview*	*Short Sequence (1-6 calls)*	*Long sequence*	
2	Frequency	390	99	411	78	489
	Percentage	79.8	20.2	84.0	16.0	
3	Frequency	320	64	338	46	384
	Percentage	83.3	16.7	88.0	12.0	
4	Frequency	306	41	310	37	347
	Percentage	88.2	11.8	89.3	10.7	
5	Frequency	286	37	294	29	323
	Percentage	88.5	11.5	91.0	9.0	

Evaluation criteria for frequentist and Bayesian models (Akaike information Criteria (AIC), Nagelkerke's pseudo R^2 and Watanabe Akaike Information Criteria (WAIC))

Wave	**Model**	**Final call outcome**			**Length of call sequence**		
		AIC	Nagelkerke R^2 (%)	WAIC	AIC	Nagelkerke R^2 (%)	WAIC
1 and 2	frequentist	412.06	7.37	-	316.01	21.3	-
	M1	-	-	412.6	-	-	316.91
2 and 3	frequentist	289.89	10.21	-	218.76	25.16	-
	M1	-	-	290.74	-	-	219.43
	M2	-	-	278.88	-	-	223.51
	M3			279.04	-	-	297.50
	M4	-	-	283.75	-	-	219.26
	M5	-	-	288.76	-	-	218.06
	M6	-	-	290.18	-	-	218.89
	M7	-	-	290.76	-	-	219.46
3 and 4	frequentist	207.69	15.45	-	195.72	10.05	-
	M1			611.71	-	-	197.21
	M2			205.90	-	-	187.39
	M3			207.45	-	-	185.21
	M4			204.82	-	-	191.54
	M5	-	-	205.66	-	-	193.70
	M6	-	-	206.81	-	-	195.43
	M7	-	-	2112.28	-	-	196.59
4 and 5	frequentist	189.85	14.55	-	164.49	27.09	-
	M1	-	-	192.39	-	-	820.3
	M2	-	-	179.02	-	-	166.32
	M3	-		178.25	-	-	162.98
	M4	-	-	182.24	-	-	164.71
	M5	-	-	188.49	-	-	159.34
	M6	-	-	191.14	-	-	157.53
	M7	-	-	192.41	-	-	9658.18

Appendix A

Classification table and AUC of ROC curves for the final call outcome and length of call sequence.

Wave	Model	Final call outcome		Length of call sequence	
		Classification (%)	AUC (%)	Classification (%)	AUC (%)
1 and 2	frequentist	83.7	66.9	79.6	55.5
	M1	83.7	66.7	79.6	55.2
2 and 3	frequentist	83.7	67.9	90.5	77.7
	M1	83.8	67.8	90.5	78.6
	M2	82.4	70.4	89.2	67.2
	M3	82.4	65.8	89.2	62.7
	M4	82.4	69.7	90.5	75.5
	M5	83.8	68.7	90.5	79.3
	M6	83.8	68.2	90.5	78.9
	M7	83.8	67.8	90.5	78.6
3 and 4	frequentist	87.0	75.9	85.5	76.1
	M1	87.0	76.4	85.5	75.9
	M2	87.0	53.3	85.5	75.4
	M3	87.0	54.8	85.5	44.5
	M4	87.0	66.2	85.5	75.7
	M5	87.0	74.4	85.5	76.2
	M6	87.0	75.8	85.5	76.0
	M7	87.0	76.0	85.5	75.9
4 and 5	frequentist	86.2	46.19	95.3	69.7
	M1	86.2	46.89	95.4	69.3
	M2	86.2	57.29	95.4	67.5
	M3	86.2	65.26	95.4	69.2
	M4	86.2	47.79	95.4	67.2
	M5	86.2	47.04	95.4	69.5
	M6	86.2	47.23	95.4	68.2
	M7	86.2	46.89	95.4	69.3

Results of sensitivity, specifity, positive predictive values (PPV) and negative predictive values (NPV) of the two binary responses

Wave	Model	Final call outcome				Length of call sequence			
		Sensitivity	Specificity	PPV	NPV	Sensitivity	Specificity	PPV	NPV
1 and 2	frequentist	100.0	83.0	11.0	100.0	25.0	82.0	6.0	96.0
	M1	100.0	83.0	11.0	100.0	25.0	82.0	6.0	96.0
2 and 3	frequentist	100.0	83.0	11.0	100.0	100.0	90.0	12.0	100.0
	M1	100.0	84.0	8.0	100.0	100.0	90.0	12.0	100.0
	M2	NaN	82.0	0.0	100.0	NaN	89.0	0.0	100.0
	M3	NaN	82.0	0.0	100.0	NaN	89.0	0.0	100.0
	M4	NaN	82.0	0.0	100.0	100.0	90.0	12.0	100.0
	M5	100.0	84.0	8.0	100.0	100.0	90.0	12.0	100.0
	M6	100.0	84.0	8.0	100.0	100.0	90.0	12.0	100.0
	M7	100.0	84.0	8.0	100.0	100.0	90.0	12.0	100.0
3 and 4	frequentist	NaN	87.0	0.0	100.0	NaN	86.0	0.0	100.0
	M1	NaN	87.0	0.0	100.0	NaN	86.0	0.0	100.0
	M2	NaN	87.0	0.0	100.0	NaN	86.0	0.0	100.0
	M3	NaN	87.0	0.0	100.0	NaN	86.0	0.0	100.0
	M4	NaN	87.0	0.0	100.0	NaN	86.0	0.0	100.0
	M5	NaN	87.0	0.0	100.0	NaN	86.0	0.0	100.0
	M6	NaN	87.0	0.0	100.0	NaN	86.0	0.0	100.0
	M7	NaN	87.0	0.0	100.0	NaN	86.0	0.0	100.0
4 and 5	frequentist	NaN	86.0	0.0	100.0	NaN	95.0	0.0	100.0
	M1	NaN	86.0	0.0	100.0	NaN	95.0	0.0	100.0
	M2	NaN	86.0	0.0	100.0	NaN	95.0	0.0	100.0
	M3	NaN	86.0	0.0	100.0	NaN	95.0	0.0	100.0
	M4	NaN	86.0	0.0	100.0	NaN	95.0	0.0	100.0
	M5	NaN	86.0	0.0	100.0	NaN	95.0	0.0	100.0
	M6	NaN	86.0	0.0	100.0	NaN	95.0	0.0	100.0
	M7	NaN	86.0	0.0	100.0	NaN	95.0	0.0	100.0

Classification table and AUC of ROC curves for the final call outcome and length of call sequence.

Wave	Model	Final call outcome		Length of call sequence	
		Classification (%)	AUC (%)	Classification (%)	AUC (%)
1 and 2	frequentist	83.7	66.9	79.6	55.5
	M1	83.7	66.7	79.6	55.2
2 and 3	frequentist	83.7	67.9	90.5	77.7
	M1	83.8	67.8	90.5	78.6
	M2	82.4	70.4	89.2	67.2
	M3	82.4	65.8	89.2	62.7
	M4	82.4	69.7	90.5	75.5
	M5	83.8	68.7	90.5	79.3
	M6	83.8	68.2	90.5	78.9
	M7	83.8	67.8	90.5	78.6
3 and 4	frequentist	87.0	75.9	85.5	76.1
	M1	87.0	76.4	85.5	75.9
	M2	87.0	53.3	85.5	75.4
	M3	87.0	54.8	85.5	44.5
	M4	87.0	66.2	85.5	75.7
	M5	87.0	74.4	85.5	76.2
	M6	87.0	75.8	85.5	76.0
	M7	87.0	76.0	85.5	75.9
4 and 5	frequentist	86.2	46.19	95.3	69.7
	M1	86.2	46.89	95.4	69.3
	M2	86.2	57.29	95.4	67.5
	M3	86.2	65.26	95.4	69.2
	M4	86.2	47.79	95.4	67.2
	M5	86.2	47.04	95.4	69.5
	M6	86.2	47.23	95.4	68.2
	M7	86.2	46.89	95.4	69.3

A.7 Investigating Effect of Correlation Between Outcome and Auxiliary Variable

Predicting employment using Income

This analysis aims to investigate whether strength of relationships between response and explanatory variables influences borrowed information from previous wave using household income and employment variables that are highly correlated with each other.

The response variable for analysis is the employment status and is defined as:

$$y_i = \begin{cases} 1 & \text{Employed} \\ 0 & \text{Not Employed} \end{cases}$$

Household Income is the explanatory variable.

Correlation for employment and household income across three waves

Wave	*Contingency coefficient*	*Cramer's V*
2	0.5	0.6
3	0.5	0.6
4	0.5	0.6

Evaluation criteria for frequentist and Bayesian models (Akaike information Criteria (AIC), Nagelkerke's pseudo R^2 and Watanabe Akaike Information Criteria (WAIC)) for employment

Wave	Model	AIC	Nagelkerke R^2 (%)	WAIC
2	frequentist	10195.0	45.3	-
	M1	-	-	10915.1
3	frequentist	8860.0	42.2	-
	M1	-	-	8860.0
	M2	-	-	9416.7
	M3	-	-	9030.7
	M4	-	-	8922.3
	M5	-	-	8872.6
	M6	-	-	9187.7
	M7	-	-	9030.7
4	frequentist	8183.9	41.0	-
	M1	-	-	8183.9
	M2	-	-	8575.3
	M3	-	-	8292.1
	M4	-	-	8190.9
	M5	-	-	8402.2
	M6	-	-	8328.6
	M7	-	-	8240.5

Appendix A

Classification table and AUC of ROC values for employment.

Wave	Model	Classification (%)	AUC (%)
2	frequentist	79.2	87.1
	M1	79.2	87.1
3	frequentist	77.7	83.0
	M1	77.7	85.0
	M2	77.7	85.0
	M3	77.7	85.0
	M4	77.7	85.0
	M5	77.7	85.0
	M6	77.7	85.0
	M7	77.7	85.0
4	frequentist	77.5	84.7
	M1	77.5	84.7
	M2	77.5	84.7
	M3	77.5	84.7
	M4	77.5	84.7
	M5	77.4	84.7
	M6	77.4	84.7
	M7	77.4	84.7

Results of sensitivity, specifity, positive predictive values (PPV) and negative predictive values (NPV) of employment

Wave	Model	Sensitivity	Specificity	PPV	NPV
2	frequentist	79.0	79.0	55.0	92.0
	M1	79.0	79.0	55.0	92.0
3	frequentist	76.0	78.0	53.0	91.0
	M1	76.0	78.0	53.0	91.0
	M2	76.0	78.0	53.0	91.0
	M3	76.0	78.0	53.0	91.0
	M4	76.0	78.0	53.0	91.0
	M5	76.0	78.0	53.0	91.0
	M6	76.0	78.0	53.0	91.0
	M7	76.0	78.0	53.0	91.0
4	frequentist	75.0	78.0	54.0	90.0
	M1	75.0	78.0	54.0	90.0
	M2	75.0	78.0	54.0	90.0
	M3	75.0	78.0	54.0	90.0
	M4	75.0	78.0	54.0	90.0
	M5	75.0	78.0	54.0	90.0
	M6	75.0	78.0	54.0	90.0
	M7	75.0	78.0	54.0	90.0

A.8 Estimated coefficients for the Final Call Outcome Models

Posterior parameter estimates for wave 3 in final call outcome in uninformative prior (M1) and informative priors (M2, M3 and M7) models

Fixed Effects		M1		M2		M3		M7	
		Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
	Intercept	2.23	0.23	1.80	0.15	1.96	0.19	2.07	0.21
Government Office Region	East Midlands (reference)								
	East of England	-0.19	0.14	0.05	0.07	0.01	0.10	-0.06	0.11
	London	-0.44	0.14	-0.12	0.07	-0.22	0.10	-0.30	0.11
	North East	-0.21	0.16	0.03	0.09	-0.01	0.12	-0.08	0.14
	North West	-0.22	0.13	0.03	0.07	-0.02	0.09	-0.09	0.11
	Scotland	-0.55	0.14	-0.16	0.07	-0.30	0.10	-0.39	0.11
	South East	-0.39	0.13	-0.08	0.07	-0.17	0.09	-0.25	0.10
	South West	-0.15	0.14	0.08	0.08	0.04	0.10	-0.02	0.12
	Wales	-0.26	0.16	0.01	0.08	-0.05	0.11	-0.12	0.13
	West Midlands	-0.23	0.14	0.02	0.07	-0.03	0.10	-0.10	0.11
	Yorkshire and the Humber	-0.22	0.14	0.03	0.08	-0.02	0.10	-0.09	0.12
Rural/Urban	Urban(Reference)								
	Rural	0.04	0.07	0.06	0.06	0.05	0.07	0.05	0.07
Month	January-June Year 1 (reference)								
	July-December Year 1	0.06	0.08	0.00	0.06	0.03	0.07	0.05	0.07
	January-June Year 2	0.17	0.08	0.09	0.06	0.13	0.07	0.15	0.07
	July- December Year 2	0.39	0.08	0.25	0.06	0.34	0.07	0.37	0.08
Pensionage	no pension age people in HH								
	pension age people in HH	0.21	0.08	0.14	0.06	0.18	0.07	0.20	0.08
Employed	employed people in HH								
	No employed people in HH	-0.08	0.08	-0.03	0.06	-0.06	0.08	-0.07	0.08
Highest educational qualification in the household	Higher degree & Degree (reference)								
	A level & GCSE	-0.24	0.07	-0.13	0.05	-0.20	0.06	-0.22	0.06
	Other & No qualification	-0.43	0.09	-0.23	0.07	-0.35	0.08	-0.40	0.09
Income	1st Quartile (Reference)								
	2nd Quartile	0.02	0.09	0.04	0.06	0.03	0.08	0.02	0.08
	3rd Quartile	-0.10	0.10	-0.02	0.06	-0.06	0.08	-0.08	0.09
	4th Quartile	-0.17	0.11	-0.05	0.07	-0.12	0.09	-0.15	0.10
Tenure	Owned								
	Rented from employer privately and other	-0.60	0.08	-0.41	0.07	-0.54	0.08	-0.58	0.08
	Rented from LA or housing association	-0.06	0.09	-0.05	0.06	-0.07	0.08	-0.07	0.08
Dwelling type	house/bungalow (reference)								
	flat/maisonette	-0.02	0.09	-0.05	0.07	-0.04	0.08	-0.03	0.09
	other	-0.21	0.24	-0.03	0.10	-0.10	0.17	-0.15	0.21
Children	children in HH(Reference)								
	no children in HH	-0.19	0.08	-0.06	0.06	-0.13	0.07	-0.16	0.08
Number of cars	no car (Reference)								
	one car	0.22	0.08	0.15	0.06	0.19	0.07	0.21	0.08
	two cars	0.24	0.10	0.15	0.07	0.20	0.09	0.22	0.10
	three or more car	0.17	0.14	0.04	0.08	0.10	0.11	0.13	0.13
Household size	One person								
	2 to 3 persons	-0.22	0.08	-0.10	0.06	-0.17	0.07	-0.20	0.08
	More than 4 people	-0.23	0.12	-0.05	0.07	-0.15	0.10	-0.19	0.11
Proportion of other call outcomes	0-25% (Reference)								
	26-50%	-0.07	0.22	-0.03	0.10	-0.06	0.16	-0.07	0.19
	51-75%	0.11	0.59	0.02	0.11	0.03	0.21	0.05	0.31
	Any other status	0.27	0.13	0.15	0.08	0.21	0.11	0.23	0.12
Proportion of Contacts	0-25% (Reference)								
	26-50%	-0.28	0.08	-0.17	0.07	-0.23	0.08	-0.26	0.08
	51-75%	-0.58	0.24	-0.07	0.10	-0.24	0.17	-0.39	0.21
	76-100%	-0.27	0.86	0.02	0.11	0.00	0.22	-0.02	0.34
Proportion of appointments	0-25%								
	26-50%	-0.08	0.07	-0.02	0.05	-0.04	0.06	-0.06	0.06
	51-100%	-0.37	0.22	-0.05	0.10	-0.17	0.16	-0.26	0.19
Proportion of	0-25% (Reference)								
	26-50%	0.00	0.07	0.04	0.05	0.03	0.06	0.02	0.07
	51-75%	-0.05	0.11	0.02	0.07	0.00	0.09	-0.02	0.10

noncontacts	76-100%	0.14	0.19	0.11	0.09	0.16	0.14	0.16	0.16
Call length	short (Reference)								
	long	-0.07	0.01	-0.08	0.01	-0.08	0.01	-0.07	0.01

Posterior parameter estimates for wave 4 in final call outcome in uninformative prior (M1) and informative priors (M2, M3 and M7) models

Fixed Effects		M1		M2		M3		M7	
		Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
	Intercept	1.85	0.26	1.94	0.15	1.91	0.20	1.90	0.23
Government Office Region	East Midlands (reference)								
	East of England	0.09	0.15	0.07	0.07	0.09	0.10	0.09	0.12
	London	-0.05	0.15	-0.05	0.07	-0.06	0.10	-0.06	0.12
	North East	-0.07	0.18	-0.02	0.08	-0.04	0.12	-0.05	0.14
	North West	-0.02	0.14	0.00	0.07	-0.01	0.10	-0.01	0.11
	Scotland	-0.16	0.15	-0.06	0.08	-0.11	0.10	-0.14	0.12
	South East	-0.10	0.14	-0.04	0.07	-0.07	0.09	-0.08	0.11
	South West	0.21	0.16	0.11	0.08	0.17	0.11	0.19	0.13
	Wales	0.21	0.18	0.09	0.08	0.15	0.12	0.19	0.15
	West Midlands	-0.08	0.15	-0.02	0.07	-0.05	0.10	-0.06	0.12
	Yorkshire and the Humber	-0.07	0.15	-0.02	0.07	-0.05	0.10	-0.06	0.12
Rural/Urban	Urban(Reference)								
	Rural	0.21	0.09	0.16	0.06	0.20	0.08	0.21	0.08
Month	January-June Year 1 (reference)								
	July-December Year 1	0.11	0.09	0.03	0.06	0.07	0.08	0.09	0.08
	January-June Year 2	0.15	0.09	0.06	0.06	0.11	0.08	0.14	0.08
	July- December Year 2	0.20	0.09	0.10	0.06	0.15	0.08	0.18	0.09
Pensionage	no pension age people in HH								
	pension age people in HH	0.02	0.09	0.04	0.06	0.03	0.08	0.02	0.09
Employed	employed people in HH								
	No employed people in HH	0.10	0.10	0.02	0.06	0.06	0.08	0.08	0.09
Highest educational qualification in the household	Higher degree & Degree (reference)								
	A level & GCSE	-0.22	0.08	-0.10	0.06	-0.17	0.07	-0.20	0.07
	Other & No qualification	-0.50	0.10	-0.24	0.07	-0.39	0.09	-0.45	0.10
Income	1st Quartile (Reference)								
	2nd Quartile	0.03	0.10	0.02	0.06	0.02	0.08	0.02	0.09
	3rd Quartile	0.00	0.12	0.03	0.07	0.01	0.09	0.01	0.10
	4th Quartile	-0.11	0.13	-0.01	0.07	-0.06	0.10	-0.09	0.11
Tenure	Owned								
	Rented from employer privately and other	-0.40	0.10	-0.24	0.07	-0.36	0.09	-0.39	0.09
	Rented from LA or housing association	-0.09	0.10	-0.11	0.06	-0.12	0.08	-0.11	0.09
Dwelling type	house/bungalow (reference)								
	flat/maisonette	-0.13	0.10	-0.11	0.07	-0.12	0.09	-0.12	0.09
	other	-0.66	0.26	-0.07	0.09	-0.23	0.16	-0.40	0.21
Children	children in HH(Reference)								
	no children in HH	-0.09	0.10	0.02	0.06	-0.02	0.08	-0.05	0.09
Number of cars	no car (Reference)								
	one car	0.47	0.09	0.22	0.06	0.34	0.08	0.41	0.08
	two cars	0.61	0.12	0.24	0.07	0.40	0.09	0.50	0.11
	three or more cars	0.36	0.16	0.03	0.08	0.13	0.12	0.23	0.14
Household size	One person								
	2 to 3 persons	-0.31	0.10	-0.05	0.06	-0.16	0.08	-0.23	0.09
	More than 4 people	-0.48	0.14	-0.10	0.07	-0.25	0.11	-0.36	0.13
Proportion of other call outcomes	0-25% (Reference)								
	26-50%	0.23	0.24	0.00	0.09	0.04	0.15	0.11	0.19
	51-75%	0.64	0.77	0.02	0.10	0.05	0.19	0.11	0.29
	Any other status	0.36	0.14	0.13	0.08	0.22	0.11	0.27	0.12
Proportion of Contacts	0-25% (Reference)								
	26-50%	-0.19	0.10	-0.10	0.07	-0.15	0.09	-0.17	0.10
	51-75%	-0.28	0.30	-0.02	0.09	-0.08	0.17	-0.15	0.22
	76-100%	4.84	13.04	0.02	0.10	0.03	0.20	0.06	0.32
Proportion of appointments	0-25%								
	26-50%	0.09	0.08	0.09	0.06	0.10	0.07	0.10	0.07
	51-100%	0.15	0.29	0.02	0.09	0.06	0.16	0.09	0.21
Proportion of noncontacts	0-25% (Reference)								
	26-50%	0.00	0.08	0.02	0.06	0.02	0.07	0.01	0.08
	51-75%	0.00	0.13	0.04	0.07	0.04	0.10	0.03	0.11
	76-100%	-0.10	0.20	0.02	0.09	0.00	0.14	-0.04	0.17
Call length	short (Reference)								
	long	-0.08	0.02	-0.09	0.01	-0.09	0.01	-0.08	0.01

Posterior parameter estimates for wave 5 in final call outcome in uninformative prior (M1) and informative priors (M2, M3 and M7) models

Fixed Effects		M1		M2		M3		M7	
		Mean	*SD*	*Mean*	*SD*	Mean	*SD*	*Mean*	*SD*
	Intercept	2.65	0.30	1.91	0.16	2.15	0.22	2.34	0.25
Government Office Region	East Midlands (reference)								
	East of England	-0.37	0.17	0.03	0.08	-0.06	0.11	-0.14	0.13
	London	-0.27	0.18	0.05	0.08	0.00	0.12	-0.06	0.14
	North East	-0.29	0.21	0.08	0.09	0.03	0.14	-0.05	0.17
	North West	-0.35	0.17	0.03	0.08	-0.05	0.11	-0.13	0.13
	Scotland	-0.32	0.18	0.06	0.08	-0.01	0.12	-0.09	0.14
	South East	-0.23	0.17	0.11	0.08	0.06	0.11	-0.01	0.13
	South West	-0.18	0.18	0.12	0.08	0.10	0.12	0.03	0.14
	Wales	-0.66	0.19	-0.08	0.09	-0.27	0.13	-0.40	0.15
	West Midlands	-0.23	0.18	0.10	0.08	0.06	0.12	-0.01	0.14
	Yorkshire and the Humber	-0.07	0.18	0.15	0.08	0.16	0.12	0.11	0.14
Rural/Urban	Urban(Reference)								
	Rural	-0.02	0.09	0.04	0.07	0.04	0.07	-0.01	0.09
Month	January-June Year 1 (reference)								
	July-December Year 1	-0.10	0.10	-0.02	0.07	-0.06	0.09	-0.08	0.09
	January-June Year 2	0.17	0.10	0.17	0.07	0.18	0.09	0.18	0.10
	July- December Year 2	-0.07	0.10	0.00	0.07	-0.04	0.09	-0.06	0.10
Pensionage	no pension age people in HH								
	pension age people in HH	0.13	0.10	0.13	0.07	0.13	0.09	0.13	0.10
Employed	employed people in HH								
	No employed people in HH	0.06	0.11	0.08	0.07	0.08	0.09	0.08	0.10
Highest educational qualification in the household	Higher degree & Degree (reference)								
	A level & GCSE	-0.20	0.08	-0.09	0.06	-0.15	0.08	-0.18	0.08
	Other & No qualification	-0.14	0.12	0.00	0.08	-0.08	0.10	-0.11	0.11
Income	1st Quartile (Reference)								
	2nd Quartile	-0.20	0.12	-0.03	0.07	-0.10	0.09	-0.15	0.10
	3rd Quartile	-0.18	0.13	0.03	0.07	-0.06	0.10	-0.12	0.12
	4th Quartile	-0.20	0.15	0.08	0.08	-0.03	0.11	-0.11	0.13
Tenure	Owned								
	Rented from employer privately and other	-0.49	0.11	-0.19	0.08	-0.19	0.08	-0.44	0.10
	Rented from LA or housing association	-0.29	0.11	-0.13	0.07	-0.13	0.07	-0.27	0.10
Dwelling type	house/bungalow (reference)								
	flat/maisonette	0.11	0.12	0.04	0.08	0.06	0.10	0.08	0.11
	other	0.15	0.30	0.08	0.10	0.09	0.18	0.11	0.23
Children	children in HH(Reference)								
	no children in HH	-0.18	0.11	0.01	0.07	-0.07	0.09	-0.13	0.10
Number of cars	no car (Reference)								
	one car	0.52	0.10	0.29	0.07	0.39	0.08	0.45	0.09
	two cars	0.43	0.13	0.18	0.07	0.28	0.10	0.35	0.12
	three or more cars	0.48	0.18	0.14	0.09	0.25	0.13	0.35	0.16
Household size	One person								
	2 to 3 persons	-0.17	0.11	-0.01	0.07	-0.08	0.09	-0.13	0.10
	More than 4 people	-0.23	0.16	0.02	0.08	-0.09	0.12	-0.16	0.14
Proportion of other call outcomes	0-25% (Reference)								
	26-50%	-0.19	0.28	0.02	0.10	-0.07	0.17	-0.13	0.22
	51-75%	0.09	0.79	0.09	0.11	0.10	0.21	0.10	0.33
	Any other status	0.44	0.18	0.20	0.09	0.29	0.13	0.35	0.15
Proportion of Contacts	0-25% (Reference)								
	26-50%	-0.56	0.11	-0.16	0.08	-0.32	0.09	-0.42	0.10
	51-75%	-0.76	0.32	0.04	0.10	-0.09	0.18	-0.29	0.25
	76-100%	-2.09	0.63	0.05	0.11	-0.08	0.21	-0.34	0.33
Proportion of appointments	0-25%								
	26-50%	-0.14	0.09	0.07	0.06	0.01	0.08	-0.04	0.08
	51-100%	-0.82	0.25	0.00	0.10	-0.21	0.17	-0.44	0.21
Proportion of noncontacts	0-25% (Reference)								
	26-50%	-0.20	0.09	0.03	0.06	-0.03	0.08	-0.09	0.08
	51-75%	-0.12	0.15	0.19	0.08	0.17	0.12	0.08	0.13
	76-100%	-0.93	0.22	-0.05	0.10	-0.30	0.15	-0.54	0.18
Call length	short (Reference)								
	long	-0.07	0.02	-0.11	0.01	-0.10	0.02	-0.09	0.02

A.9 Estimated coefficients for the length of call sequence models

Posterior parameter estimates for wave 3 in length of call sequence in uninformative prior (M1) and informative priors (M2, M3 and M7) models

Fixed Effects		**M1**		**M2**		**M3**		**M7**	
		Mean	*SD*	*Mean*	*SD*		*Mean*	*SD*	*Mean*
	Intercept	1.97	0.33	1.38	0.20	1.64	0.26	1.80	0.30
Government	East Midlands (reference)								
Office Region	East of England	0.31	0.16	0.31	0.09	0.35	0.12	0.34	0.13
	London	-0.31	0.14	-0.09	0.08	-0.20	0.11	-0.25	0.12
	North East	-0.38	0.17	-0.03	0.10	-0.20	0.13	-0.29	0.15
	North West	-0.12	0.14	0.05	0.08	-0.02	0.10	-0.07	0.12
	Scotland	0.13	0.16	0.21	0.09	0.20	0.12	0.17	0.14
	South East	0.24	0.14	0.29	0.08	0.29	0.11	0.28	0.12
	South West	0.42	0.16	0.37	0.09	0.44	0.12	0.44	0.14
	Wales	0.22	0.18	0.25	0.10	0.27	0.14	0.26	0.16
	West Midlands	0.05	0.15	0.17	0.09	0.14	0.11	0.10	0.13
	Yorkshire and the Humber	0.14	0.15	0.22	0.09	0.21	0.12	0.18	0.13
Rural/Urban	Urban(reference)								
	Rural	0.31	0.09	0.31	0.07	0.31	0.08	0.31	0.09
Lone Parents	Lone parents in household (reference)								
	No lone parents in household	0.38	0.11	0.35	0.08	0.37	0.10	0.38	0.11
Pensionage	no pension age people in HH (reference)								
	pension age people in HH	0.79	0.11	0.59	0.08	0.71	0.10	0.76	0.10
Employed	employed people in HH (Reference)								
	No employed people in HH	0.25	0.10	0.24	0.07	0.24	0.09	0.25	0.09
Highest educational	Higher degree & Degree (reference)								
qualification in	A level & GCSE	-0.14	0.07	-0.05	0.06	-0.05	0.06	-0.13	0.07
the household	Other & No qualification	-0.05	0.11	0.09	0.08	0.09	0.08	-0.02	0.10
Tenure	Owned (reference)								
	Rented from employer privately and other	-0.07	0.10	0.02	0.08	-0.04	0.09	-0.06	0.09
	Rented from LA or housing association	-0.22	0.09	-0.12	0.07	-0.19	0.08	-0.21	0.09
Garden	Yes (reference)								
	no	0.09	0.12	0.21	0.08	0.14	0.10	0.10	0.11
	No obvious garden	-0.08	0.13	0.03	0.08	-0.04	0.11	-0.07	0.12
	Don’t know	6.03	11.92	0.22	0.13	0.29	0.25	0.42	0.39
Vicinity 1	Mentioned (reference)								
	Not mentioned	-0.19	0.26	0.11	0.11	0.00	0.18	-0.09	0.21
Vicinity 2	Mentioned (reference)								
	Not mentioned	0.13	0.17	0.16	0.10	0.13	0.14	0.12	0.16
Household size	One person								
	2 to 3 persons	0.23	0.09	0.22	0.07	0.22	0.08	0.23	0.09
	More than 4 people	0.15	0.10	0.14	0.08	0.15	0.09	0.15	0.10
Proportion of	0-25% (reference)								
contacts	26-50%	0.10	0.10	0.13	0.08	0.11	0.09	0.11	0.10
	51-75%	0.37	0.32	0.21	0.12	0.26	0.20	0.30	0.25
	76-100%	-0.61	0.86	0.18	0.13	0.13	0.25	0.05	0.39
Proportion of	0-25% (reference)								
noncontacts	26-50%	0.02	0.08	0.06	0.06	0.03	0.07	0.02	0.08
	51-75%	0.07	0.11	0.10	0.08	0.07	0.10	0.07	0.11
	76-100%	0.35	0.19	0.27	0.10	0.31	0.15	0.34	0.17

Posterior parameter estimates for wave 4 in length of call sequence in uninformative prior (M1) and informative priors (M2, M3 and M7) models

Fixed Effects		**M1**		**M2**		**M3**		**M7**	
		SD	*Mean*	*SD*		*Mean*	*SD*	*Mean*	*SD*
	Intercept	1.79	0.46	0.70	0.19	1.10	0.28	1.39	0.35
Government Office Region	East Midlands (reference)								
	East of England	0.16	0.17	0.31	0.08	0.26	0.11	0.23	0.14
	London	0.03	0.16	0.24	0.08	0.16	0.11	0.10	0.13
	North East	0.29	0.20	0.35	0.09	0.36	0.13	0.35	0.16
	North West	0.14	0.16	0.29	0.08	0.25	0.10	0.21	0.12
	Scotland	0.44	0.18	0.41	0.08	0.46	0.12	0.47	0.15
	South East	0.14	0.15	0.30	0.08	0.24	0.10	0.21	0.12
	South West	0.64	0.19	0.48	0.08	0.59	0.13	0.64	0.15
	Wales	-0.12	0.19	0.23	0.09	0.09	0.13	0.00	0.15
	West Midlands	0.30	0.17	0.37	0.08	0.37	0.12	0.36	0.14
	Yorkshire and the Humber	0.17	0.17	0.31	0.08	0.27	0.11	0.24	0.14
Rural/Urban	Urban(reference)								
	Rural	0.27	0.10	0.33	0.07	0.29	0.09	0.28	0.10
Lone Parents	Lone parents in household (reference)								
	No lone parents in household	0.45	0.12	0.47	0.08	0.48	0.11	0.46	0.12
Pensionage	no pension age people in HH (reference)								
	pension age people in HH	0.61	0.12	0.50	0.07	0.57	0.10	0.59	0.11
Employed	employed people in HH (Reference)								
	No employed people in HH	0.29	0.11	0.28	0.07	0.27	0.09	0.28	0.10
Highest educational qualification in the household	Higher degree & Degree (reference)								
	A level & GCSE	-0.10	0.08	0.08	0.06	-0.02	0.08	-0.06	0.08
	Other & No qualification	-0.15	0.12	0.16	0.08	0.00	0.11	-0.08	0.12
Tenure	Owned (reference)								
	Rented from employer privately and other	-0.32	0.11	0.05	0.07	-0.16	0.09	-0.25	0.10
	Rented from LA or housing association	-0.41	0.10	-0.05	0.07	-0.26	0.09	-0.35	0.10
Garden	Yes (reference)								
	no	0.31	0.13	0.39	0.07	0.35	0.10	0.32	0.11
	No obvious garden	0.27	0.14	0.29	0.08	0.28	0.11	0.27	0.12
	Don’t know	0.08	1.09	0.34	0.10	0.33	0.20	0.31	0.33
Vicinity 1	Mentioned (reference)								
	Not mentioned	0.11	0.36	0.33	0.10	0.29	0.18	0.24	0.25
Vicinity 2	Mentioned (reference)								
	Not mentioned	0.06	0.18	0.30	0.09	0.20	0.13	0.13	0.16
Household size	One person								
	2 to 3 persons	0.17	0.10	0.31	0.07	0.25	0.09	0.21	0.09
	More than 4 people	0.13	0.12	0.29	0.07	0.23	0.10	0.19	0.11
Proportion of contacts	0-25% (reference)								
	26-50%	-0.17	0.11	0.15	0.08	0.00	0.10	-0.08	0.11
	51-75%	-0.51	0.32	0.28	0.10	0.14	0.18	-0.07	0.25
	76-100%	-1.57	1.19	0.33	0.10	0.30	0.21	0.25	0.34
Proportion of noncontacts	0-25% (reference)								
	26-50%	-0.14	0.09	0.15	0.06	0.03	0.08	-0.05	0.09
	51-75%	-0.16	0.13	0.23	0.08	0.09	0.10	-0.02	0.12
	76-100%	-0.21	0.20	0.30	0.09	0.17	0.14	0.02	0.17

Appendix A

Posterior parameter estimates for wave 5 in length of call sequence in uninformative prior (M1) and informative priors (M2, M3 and M7) models

Fixed Effects		**M1**		**M2**		**M3**		**M7**	
		mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
	Intercept	2.41	0.57	1.67	0.23	1.95	0.35	2.15	0.43
Government Office Region	East Midlands (reference)								
	East of England	0.32	0.19	0.29	0.10	0.32	0.13	0.33	0.16
	London	-0.32	0.17	-0.07	0.09	-0.22	0.12	-0.27	0.14
	North East	0.13	0.22	0.20	0.11	0.18	0.16	0.16	0.19
	North West	0.34	0.18	0.31	0.10	0.35	0.13	0.35	0.15
	Scotland	0.22	0.20	0.24	0.10	0.25	0.14	0.24	0.16
	South East	0.20	0.17	0.25	0.09	0.23	0.12	0.22	0.14
	South West	0.65	0.21	0.43	0.10	0.56	0.15	0.62	0.18
	Wales	-0.20	0.20	0.07	0.11	-0.07	0.15	-0.13	0.17
	West Midlands	0.20	0.19	0.23	0.10	0.23	0.14	0.22	0.16
	Yorkshire and the Humber	0.30	0.19	0.28	0.10	0.31	0.14	0.31	0.16
Rural/Urban	Urban(reference)								
	Rural	0.22	0.11	0.25	0.08	0.23	0.10	0.22	0.11
Lone Parents	Lone parents in household (reference)								
	No lone parents in household	0.25	0.15	0.30	0.10	0.28	0.13	0.27	0.14
Pensionage	no pension age people in HH (reference)								
	pension age people in HH	0.75	0.13	0.53	0.09	0.66	0.11	0.71	0.12
Employed	employed people in HH (Reference)								
	No employed people in HH	0.13	0.12	0.19	0.09	0.16	0.11	0.14	0.12
Highest educational qualification in the household	Higher degree & Degree (reference)								
	A level & GCSE	-0.03	0.09	0.07	0.07	0.07	0.07	-0.01	0.09
	Other & No qualification	-0.10	0.14	0.09	0.09	0.09	0.09	-0.07	0.13
Tenure	Owned (reference)								
	Rented from employer privately and other	0.08	0.13	0.15	0.09	0.11	0.11	0.09	0.12
	Rented from LA or housing association	-0.29	0.11	-0.12	0.08	-0.23	0.10	-0.27	0.11
Garden	Yes (reference)								
	no	0.18	0.15	0.27	0.09	0.23	0.12	0.21	0.13
	No obvious garden	0.13	0.16	0.18	0.09	0.16	0.13	0.15	0.14
	Don’t know	-0.20	0.43	0.19	0.13	0.11	0.23	0.01	0.32
Vicinity 1	Mentioned (reference)								
	Not mentioned	0.03	0.47	0.22	0.13	0.18	0.23	0.13	0.32
Vicinity 2	Mentioned (reference)								
	Not mentioned	0.01	0.22	0.18	0.11	0.10	0.17	0.05	0.20
Household size	One person								
	2 to 3 persons	0.08	0.11	0.17	0.08	0.12	0.10	0.10	0.11
	More than 4 people	0.07	0.13	0.14	0.09	0.10	0.11	0.08	0.12
Proportion of contacts	0-25% (reference)								
	26-50%	0.00	0.12	0.08	0.09	0.03	0.11	0.01	0.12
	51-75%	0.14	0.38	0.22	0.13	0.20	0.22	0.18	0.29
	76-100%	-0.61	0.84	0.21	0.13	0.16	0.26	0.06	0.41
Proportion of noncontacts	0-25% (reference)								
	26-50%	-0.08	0.10	0.02	0.07	-0.04	0.09	-0.06	0.09
	51-75%	0.20	0.15	0.24	0.09	0.22	0.12	0.21	0.14
	76-100%	0.39	0.24	0.30	0.11	0.35	0.17	0.37	0.21

Appendix B [Paper 2]

B.1 Descriptive

Distributions of explanatory variables for NSW Field Test 2015 based on response and cooperation

		Response	**Cooperation**
Variable	*Levels*	*Frequency (%)*	*Frequency (%)*
Incentive	No Incentive	2,460 (48.6%)	1,898 (48.4%)
	Incentive	2,597 (51.4%)	2,027 (51.6%)
Interviewer age	Less than 50 years	612 (12.1%)	457 (11.6%)
	51 to 59 years	1,807 (35.7%)	1,399 (35.6%)
	61 to 69 years	1,953 (38.6%)	1,540 (39.2%)
	Greater than 70 years	685 (13.5%)	529 (13.4%)
Interviewer experience	Less than 2 years	493 (9.7%)	383 (9.7%)
	3 to 6 years	1,894 (37.5%)	1,450 (36.9%)
	7 to 10 years	964 (19.1%)	745 (19.0%)
	Greater than 10 years	1,706 (33.7%)	1,347 (34.3%)
Gender	Female	1,969 (38.9%)	1,551 (39.5%)
	Male	3,088 (61.1%)	2,374 (60.5%)
Area Variable			
Rural/Urban	Village	812 (16.1%)	653 (16.6%)
	Hamlet & isolated dwellings	402 (7.9%)	311 (7.9%)
	Town and Fringe	982 (19.4%)	767 (19.5%)
	Urban >10k	2,861 (56.6%)	2,194 (55.9%)

Appendix B

Distributions of explanatory variables for NSW Incentive Experiment 2016 based on response and cooperation

		Response	Cooperation
Variable	*Levels*	*Frequency*	*Frequency (%)*
Incentive	No Incentive	3,011 (49.2%)	2,539 (49.1%)
	Incentive	3,111 (50.8%)	2,636 (50.9%)
Interviewer age	Up to 55 years	3,665 (59.9%)	3,030 (58.6%)
	55 years and older	2,457 (40.1%)	2,145 (41.4%)
Interviewer experience	Less than 1 to 5 years	4,473 (78.0%)	3,714 (71.8%)
	5 to 10 years	841 (13.7%)	759 (14.7%)
	Greater than 10 years	808 (13.2%)	702 (13.5%)
Gender	Female	2,936 (48.0%)	2,461 (47.6%)
	Male	3,186 (52.0%)	2,714 (52.2%)
Area Variable			
Rural/Urban	Village	775 (12.7%)	685 (13.2%)
	Hamlet & isolated dwellings	497 (8.1%)	444 (8.6%)
	Town and Fringe	1,141(18.6%)	1,004 (19.4%)
	Urban >10k	3,709 (60.6%)	3,042 (58.8%)

Distributions of categorical explanatory variables for Innovation Panel (wave 1) based on response and cooperation

		Response	**Cooperation**
Variable	*Levels*	*Frequency (%)*	*Frequency (%)*
Incentive	£5 per adult	714 (33.8%)	600 (32.5%)
	£10 per adult and £5 rising to £10 per adult	1,399 (66.2%)	1,247 (67.5%)
Interviewer age	less than 40 years	127 (6.0%)	106 (5.7%)
	41 to 50 years	279 (13.2%)	237 (12.8%)
	51 to 60 years	925 (43.8%)	806 (43.6%)
	Greater than 60 years	782 (37.0%)	698 (37.8%)
Interviewer experience	Less than 2 years	762 (36.1%)	659 (35.7%)
	3 to 6 years	816 (38.6%)	714 (38.7%)
	7 to 10 years	344 (16.3%)	307 (16.6%)
	Greater than 10 years	191 (8.2%)	167 (9.0%)
Race	Majority	1,418 (67.1%)	1,238 (67.0%)
	Others	57 (2.7%)	53 (2.9%)
	Refused	638 (30.2%)	168 (9.0%)
Gender	Female	1,115 (52.8%)	972 (52.6%)
	Male	998 (47.2%)	875 (47.4%)
Rural/Urban	Rural	1,667(78.9%)	1,446 (78.3%)
	Urban	446(21.1%)	401(21.7%)

Distributions of continuous explanatory variables for Innovation Panel (Wave 1) based on response and cooperation

Variable	**Response**		**Cooperation**	
	Mean	*Standard deviation*	*Mean*	*Standard deviation*
Socio-economic disadvantage	0.023	1.003	-0.022	0.957
Urbanicity	-0.091	0.895	-0.110	0.902
Population Mobility	-0.132	0.831	-0.174	0.782
Age Profile	-0.097	1.055	-0.117	1.05
Housing structure	-0.062	0.962	-0.038	0.894
Crime rate	-0.103	0.790	-0.132	0.779

B.2 Estimated Coefficients for Response Models

Estimated coefficients for the final standard multilevel models 5 and 6 for National Survey for Wales Field Test 2015

Appendix B

Variable *{Reference Category}*	*Category*	Model 5				Model 6			
		β	*SD*	*0.025 Quantile*	*0.975 Quantile*	*β*	*SD*	*0.025 Quantile*	*0.975 Quantile*
Intercept		-0.074	0.266	-0.608	0.443	-0.124	0.283	-0.678	0.443
Incentive *{no incentive}*	£10 Incentive	0.163	0.067	0.033	0.296	0.349	0.296	-0.232	0.915
Urban/Rural {Village}	Hamlet & isolated dwellings	0.077	0.130	-0.178	0.329	0.072	0.130	-0.182	0.327
	Town and Fringe	-0.184	0.106	-0.392	0.024	-0.187	0.106	-0.397	0.020
	Urban >10k	-0.237	0.094	-0.421	-0.052	-0.240	0.095	-0.426	-0.054
Interviewer age *{young}*	Lower middle	0.017	0.184	-0.334	0.385	0.005	0.204	-0.383	0.413
	Upper Middle	0.191	0.182	-0.157	0.560	0.211	0.205	-0.181	0.626
	Old	0.214	0.217	-0.208	0.644	0.366	0.235	-0.084	0.833
Interviewer Experience {*less*}	Lower middle	0.030	0.196	-0.351	0.411	0.090	0.218	-0.344	0.519
	Upper middle	0 .305	0.217	-0.125	0.735	0.439	0.242	-0.047	0.913
	Highest	0.375	0.207	-0.024	0.779	0.389	0.228	-0.065	0.836
Interviewer Sex *{Female}*	Male	-0.112	0.125	-0.359	0.132	-0.148	0.134	-0.415	0.111
Incentive *{£10 per adult}**Gender *{Female}*	£10 per adult *Male					0.064	0.147	-0.223	0.350
Incentive *{£10 per adult}* * Age *{young}*	£10* Lower middle					-0.031	0.228	-0.414	0.480
	£10* Upper Middle					-0.050	0.221	-0.488	0.380
	£10* Old					-0.430	0.259	-0.933	0.076
Incentive *{£5}* * Experience *{less}*	£10*Lower Middle					-0.174	0.236	-0.641	0.286
	£10*Upper Middle					-0.396	0.260	-0.911	0.106
	£10*Highest					-0.053	0.248	-0.547	0.425
$\sigma^2_{\mu 0} = var(\mu_{oj})$		0.139	0.047	0.069	0.244	0.138	0.045	0.068	0.242
$\sigma^2_{\mu 1} = var(\mu_{1j})$		0.063	0.031	0.022	0.139	0.065	0.032	0.022	0.144
$\sigma_{\mu 01} = cov(\mu_{0j}, \mu_{1j})$		0.027	0.028	-0.032	0.079	0.029	0.027	-0.029	0.081
DIC		6751.556				6750.275			

Estimated coefficients for the final standard multilevel models 5 and 6 for NSW Incentive Experiment 2016

		Model 5				*Model 6*			
Variable *{Reference Category}*	*Category*	*β*	*SD*	*0.025 Quantile*	*0.975 Quantile*	*β*	*SD*	*0.025 Quantile*	*0.975 Quantile*
Intercept		0.419	0.117	0.185	0.644	0.410	0.004	0.167	0.649
Incentive *{no incentive}*	£5 Incentive	0.079	0.061	-0.042	0.199	0.166	0.099	-0.026	0.360
Population density of area {Village}	Hamlet and isolated dwellings	0.165	0.126	-0.083	0.412	0.160	0.126	-0.085	0.406
	Town and Fringe	-0.297	0.103	-0.495	-0.090	-0.305	0.103	-0.507	-0.101
	Urban	-0.351	0.098	-0.538	-0.16	-0.363	0.095	-0.550	-0.175
Interviewer age {young}	Upper Middle	0.060	0.172	-0.282	0.394	0.218	0.149	-0.074	0.511
Interviewer Experience {less}	Upper middle	-0.192	0.188	-0.557	0.177	-0.156	0.190	-0.528	0.219
	Highest	0.187	0.129	-0.071	0.437	-0.226	0.212	-0.638	0.195
Interviewer Sex *{Female}*	Male					-0.084	0.124	-0.326	0.156
Incentive *{£10 per adult}**Gender *{Female}*	£10 per adult *Male					-0.297	0.131	-0.555	-0.043
Incentive *{£10 per adult}* * Age *{young}*	£10* Upper Middle					-0.028	0.155	-0.334	0.276
Incentive *{£5}* * Experience *{less}*	£10*Upper Middle					0.495	0.202	0.101	0.894
	£10*Highest					0.068	0.221	-0.368	0.494
$\sigma^2_{\mu 0} = var(\mu_{oj})$		0.128	0.042	0.071	0.234	0.132	0.041	0.068	0.226
$\sigma^2_{\mu 1} = var(\mu_{1j})$		0.068	0.032	0.023	0.148	0.062	0.030	0.022	0.137
$\sigma_{\mu 01} = cov(\mu_{0j}, \mu_{1j})$		-0.005	0.030	-0.083	0.038	-0.007	0.028	-0.071	0.040
DIC		8161.125				8156.666			

Appendix C [Paper 3]

C.1 Results for weighted Propensity Score Model

SMD for baseline covariates for face-to-face and online (follow up) samples before and after matching (weighted model)-Greedy Nearest Neighbour Matching

Variable {Ref}	Categories	Before Matching			After Matching (weighted model)		
		Face-to-face	*Online follow up*	*P-value (SMD)*	*Face-to- Face*	*Online follow up*	*P-value (SMD)*
		Freq (%)	*Freq (%)*		*Freq (%)*	*Freq (%)*	
Age	16 to 34 years	156 (23.4)	226 (16.0)	**0.001**	143 (22.1)	146 (22.6)	0.983
	35 to 49 years	142 (21.3)	387 (27.4)	(0.273)	142 (21.9)	142 (21.9)	(0.035)
	50 to 64 years	168 (25.2)	415 (29.4)		167 (25.8)	167 (25.8)	
	65 to 74 years	107 (16.1)	255 (18.1)		105 (16.2)	109 (16.8)	
	Over 75 years	93 (14.0)	127 (9.0)		90 (13.9)	83 (12.8)	
Race {Others}	White	579 (86.9)	1297 (92.0)	**0.001** (0.165)	573 (88.6)	563 (87.0)	0.445 (0.047)
Number of adults in household	1	228 (34.2)	349 (24.8)	**0.001**	220 (34.0)	217 (33.5)	0.997
	2	331 (49.7)	817 (57.9)	(0.218)	325 (50.2)	328 (50.7)	(0.013)
	3	72 (10.8)	149 (10.6)		67 (10.4)	68 (10.5)	
	4 or more	35 (5.3)	95 (6.7)		35 (5.4)	34 (5.3)	
Income	0 to < £15K	302 (45.3)	596 (42.3)	**0.001**	294 (45.4)	281 (43.4)	0.897
	£15K to <£40K	206 (30.9)	529 (37.5)	(0.158)	204 (31.5)	209 (32.3)	(0.043)
	>£40K	62 (9.3)	163 (11.6)		62 (9.6)	67 (10.4)	
	No data	96 (14.4)	122 (8.7)		87 (13.4)	90 (13.9)	
Education	No Qualifications	255 (38.3)	380 (27.0)	**0.001**	241 (37.2)	231 (35.7)	0.518
	Other Qualifications	284 (42.6)	645 (45.7)	(0.269)	280 (43.3)	300 (46.4)	(0.064)
	Degree or above	127 (19.1)	385 (27.3)		126 (19.5)	116 (17.9)	
GOR	London	90 (13.5)	142 (10.1)	**0.007**	80 (12.4)	84 (13.0)	0.954
	East Midlands	53 (8.0)	103 (7.3)	(0.218)	53 (8.2)	67 (10.4)	(0.091)
	East of England	81 (12.2)	165 (11.7)		81 (12.5)	76 (11.7)	
	North East	39 (5.9)	76 (5.4)		39 (6.0)	42 (6.5)	
	North West	88 (13.2)	197 (14.0)		87 (13.4)	84 (13.0)	
	South East	87 (13.1)	266 (18.9)		86 (13.3)	78 (12.1)	
	South West	56 (8.4)	154 (10.9)		56 (8.7)	53 (8.2)	
	West Midlands	92 (13.8)	154 (10.9)		85 (13.1)	87 (13.4)	
	Yorkshire and Humberside	80 (12.0)	153 (10.9)		80 (12.4)	76 (11.7)	
Number of children	0	491 (73.7)	1014 (71.9)	0.755 (0.052)	480 (74.2)	463 (71.6)	0.539 (0.082)
	1	76 (11.4)	184 (13.0)		74 (11.4)	91 (14.1)	
	2	71 (10.7)	151 (10.7)		67 (10.4)	65 (10.0)	
	3 or more	28 (4.2)	61 (4.3)		26 (4.0)	28 (4.3)	
Paid work {No}	Yes	339 (50.9)	781 (55.4)	0.062 (0.090)	333 (51.5)	323 (49.9)	0.617 (0.031)
Tenure	private rent	150 (22.5)	243 (17.2)	**0.001**	142 (21.9)	144 (22.3)	0.987
	Mortgaged	172 (25.8)	462 (32.8)	(0.267)	170 (26.3)	168 (26.0)	(0.020)
	Outright ownership	226 (33.9)	551 (39.1)		224 (34.6)	228 (35.2)	
	Social rent	118 (17.7)	154 (10.9)		111 (17.2)	107 (16.5)	
Language {Other}	English	627 (94.1)	1358 (96.3)	**0.033** (0.102)	614 (94.9)	609 (94.1)	0.625 (0.034)
Gender {Female}	Male	287 (43.1)	615 (43.6)	0.859 (0.011)	280 (43.3)	281 (43.4)	1.000 (0.003)
Marital Status {married}	Single	197 (29.6)	308 (21.8)	**0.001** (0.178)	188 (29.1)	178 (27.5)	0.579 (0.034)

Before matching: face-to-face = 666 and Online (follow up) =1,410 respondents
After matchin: face-to-face = 649 and Online (follow up) = 649 respondents

SMD for baseline covariates for face-to-face and ABOS samples before and after matching (weighted model) -Greedy Nearest Neighour Matching

Variable {Ref}	Categories	Before Matching			After Matching (weighted model)		
		Face-to-face	*Online (ABOS)*	*P-value (SMD)*	*Face-to-Face*	*Online (ABOS)*	*P-value (SMD)*
		Freq (%)	Freq (%)		Freq (%)	Freq (%)	
Age	16 to 34 years	156 (23.4)	175 (22.4)	**0.001**	118 (24.0)	110 (22.4)	0.696
	35 to 49 years	142 (21.3)	203 (26.0)	(0.243)	118 (24.0)	107 (21.7)	(0.095)
	50 to 64 years	168 (25.2)	206 (26.4)		125 (25.4)	141 (28.7)	
	65 to 74 years	107 (16.1)	142 (18.2)		87 (17.7)	84 (17.1)	
	Over 75 years	93 (14.0)	127 (9.0)		44 (8.9)	50 (10.2)	
Race {Others}	White	579 (86.9)	712 (91.2)	**0.012** (0.136)	444 (90.2)	438 (89.0)	0.601 (0.040)
Number of adults in household	1	228 (34.2)	120 (15.4)	**0.001**	102 (20.7)	114 (23.2)	0.777
	2	331 (49.7)	461 (59.0)	(0.477)	289 (58.7)	285 (57.9)	(0.067)
	3	72 (10.8)	110 (14.1)		66 (13.4)	59 (12.0)	
	4 or more	35 (5.3)	90 (11.5)		35 (7.1)	34 (6.9)	
Income	0 to < £15K	302 (45.3)	305 (39.1)	**0.002**	201 (40.9)	204 (41.5)	0.976
	£15K to <£40K	206 (30.9)	308 (39.4)	(0.205)	170 (34.6)	173 (35.2)	(0.029)
	>£40K	62 (9.3)	83 (10.6)		54 (11.0)	52 (10.6)	
	No data	96 (14.4)	85 (10.9)		67 (13.6)	63 (12.8)	
Education	No Qualifications	255 (38.3)	199 (25.5)	**0.001**	148 (30.1)	156 (31.7)	0.857
	Other Qualifications	284 (42.6)	341 (43.7)	(0.332)	227 (46.1)	221 (44.9)	(0.035)
	Degree or above	127 (19.1)	241 (30.9)		117 (23.8)	115 (23.4)	
GOR	London	90 (13.5)	85 (10.9)	**0.001**	56 (11.4)	68 (13.8)	0.890
	East Midlands	53 (8.0)	65 (8.3)	(0.308)	40 (8.1)	42 (8.5)	(0.121)
	East of England	81 (12.2)	77 (9.9)		62 (12.6)	59 (12.0)	
	North East	39 (5.9)	30 (3.8)		23 (4.7)	26 (5.3)	
	North West	88 (13.2)	128 (16.4)		72 (14.6)	70 (14.2)	
	South East	87 (13.1)	160 (20.5)		82 (16.7)	68 (13.8)	
	South West	56 (8.4)	87 (11.1)		51 (10.4)	48 (9.8)	
	West Midlands	92 (13.8)	67 (8.6)		46 (9.3)	54 (11.0)	
	Yorkshire and Humberside	80 (12.0)	82 (10.5)		60 (12.2)	57 (11.6)	
Number of children	0	491 (73.7)	594 (76.1)	**0.023** (0.160)	362 (73.6)	376 (76.4)	0.462 (0.102)
	1	76 (11.4)	91 (11.7)		60 (12.2)	59 (12.0)	
	2	71 (10.7)	84 (10.8)		61 (12.4)	46 (9.3)	
	3 or more	28 (4.2)	12 (1.5)		9 (1.8)	11 (2.2)	
Paid work {No}	Yes	339 (50.9)	443 (56.7)	**0.031** (0.117)	288 (58.5)	272 (55.3)	0.334 (0.066)
Tenure	private rent	150 (22.5)	176 (22.5)	**0.001** (0.271)	112 (22.8)	105 (21.3)	0.354 (0.115)
	Mortgaged	172 (25.8)	238 (30.5)		142 (28.9)	146 (29.7)	
	Outright ownership	226 (33.9)	298 (38.2)		177 (36.0)	195 (39.6)	
	Social rent	118 (17.7)	69 (8.8)		61 (12.4)	46 (9.3)	
Language {Other}	English	627 (94.1)	758 (97.1)	**0.009** (0.142)	469 (95.3)	474 (96.3)	0.523 (0.051)
Gender {Female}	Male	287 (43.1)	371 (47.5)	0.104 (0.089)	222 (45.1)	229 (46.5)	0.701 (0.029)
Marital Status {married}	Single	197 (29.6)	197 (25.2)	0.073 (0.098)	140 (28.5)	127 (25.8)	0.390 (0.056)

Before matching: face-to-face = 666 and ABOS =781 respondents
After matching: face-to-face = 492 and ABOS = 492 respondents

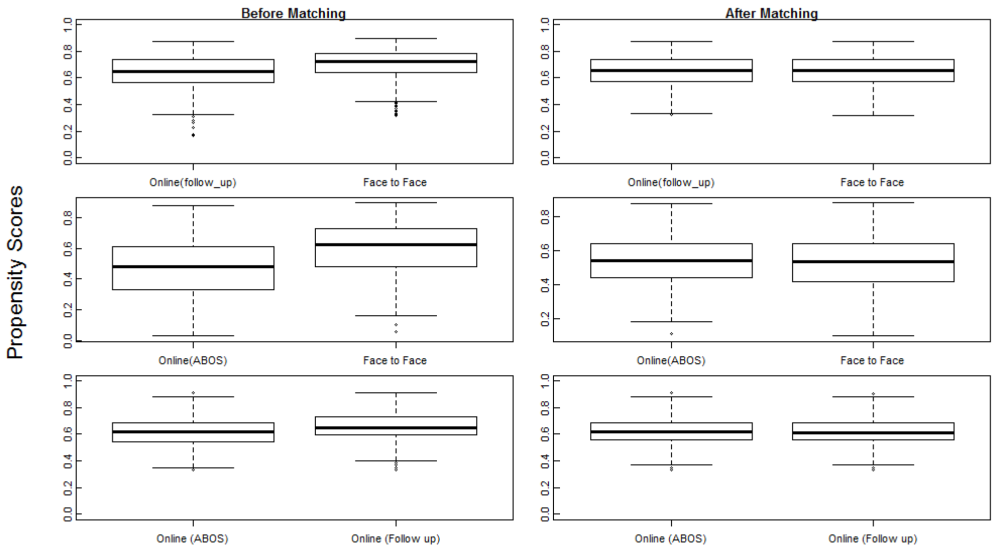
Appendix C

SMD for baseline covariates for ABOS and online (follow up) samples before and after matching (weighted model)-Greedy Nearest Neighour Matching

Variable {Ref}	Categories	Before Matching			After Matching (weighted model)		
		ABOS	*Online follow up*	*P-value (SMD)*	*ABOS*	*Online follow up*	*P-value (SMD)*
		Freq (%)	Freq (%)		Freq (%)	Freq (%)	
Age	16 to 34 year*s*	175 (22.4)	226 (16.0)	**0.004**	156 (20.5)	155 (20.4)	0.944
	35 to 49 years	203 (26.0)	387 (27.4)	(0.174)	201 (26.4)	196 (25.8)	(0.045)
	50 to 64 years	206 (26.4)	415 (29.4)		206 (27.1)	202 (26.6)	
	65 to 74 years	142 (18.2)	255 (18.1)		142 (18.7)	155 (20.4)	
	Over 75 years	55 (7.0)	127 (9.0)		156 (20.5)	155 (20.4)	
Race {Others}	White	712 (91.2)	1297 (92.0)	0.558 (0.030)	695 (91.4)	694 (91.3)	1.000 (0.005)
Number of adults in household	1	120 (15.4)	349 (24.8)	**0.001**	120 (15.8)	123 (16.2)	0.852
	2	461 (59.0)	817 (57.9)	(0.284)	461 (60.7)	455 (59.9)	(0.046)
	3	110 (14.1)	149 (10.6)		109 (14.3)	103 (13.6)	
	4 or more	90 (11.5)	95 (6.7)		70 (9.2)	79 (10.4)	
Income	0 to < £15K	305 (39.1)	596 (42.3)	0.118	290 (38.2)	332 (43.7)	**0.010**
	£15K to <£40K	308 (39.4)	529 (37.5)	(0.097)	303 (39.9)	288 (37.9)	(0.174)
	>£40K	83 (10.6)	163 (11.6)		82 (10.8)	88 (11.6)	
	No data	85 (10.9)	122 (8.7)		85 (11.2)	52 (6.8)	
Education	No Qualifications	199 (25.5)	380 (27.0)	0.211(0.078)	198 (26.1)	197 (25.9)	0.177
	Other Qualifications	341 (43.7)	645 (45.7)		329 (43.3)	360 (47.4)	(0.096)
	Degree or above	241 (30.9)	385 (27.3)		233 (30.7)	203 (26.7)	
GOR	London	85 (10.9)	142 (10.1)	0.231	85 (11.2)	79 (10.4)	0.966
	East Midlands	65 (8.3)	103 (7.3)	(0.146)	62 (8.2)	59 (7.8)	(0.079)
	East of England	77 (9.9)	165 (11.7)		75 (9.9)	70 (9.2)	
	North East	30 (3.8)	76 (5.4)		30 (3.9)	25 (3.3)	
	North West	128 (16.4)	197 (14.0)		120 (15.8)	120 (15.8)	
	South East	160 (20.5)	266 (18.9)		159 (20.9)	155 (20.4)	
	South West	87 (11.1)	154 (10.9)		81 (10.7)	95 (12.5)	
	West Midlands	67 (8.6)	154 (10.9)		67 (8.8)	73 (9.6)	
	Yorkshire and Humberside	82 (10.5)	153 (10.9)		81 (10.7)	84 (11.1)	
Number of children	0	594 (76.1)	1014 (71.9)	**0.003** (0.175)	574 (75.5)	588 (77.4)	0.750 (0.057)
	1	91 (11.7)	184 (13.0)		91 (12.0)	78 (10.3)	
	2	84 (10.8)	151 (10.7)		83 (10.9)	83 (10.9)	
	3 or more	12 (1.5)	61 (4.3)		12 (1.6)	11 (1.4)	
Paid work {No}	Yes	443 (56.7)	781 (55.4)	0.578 (0.027)	433 (57.0)	423 (55.7)	0.642 (0.027)
Tenure	private rent	176 (22.5)	243 (17.2)	**0.015** (0.143)	164 (21.6)	146 (19.2)	0.415 (0.087)
	Mortgaged	238 (30.5)	462 (32.8)		234 (30.8)	251 (33.0)	
	Outright ownership	298 (38.2)	551 (39.1)		296 (38.9)	285 (37.5)	
	Social rent	69 (8.8)	154 (10.9)		66 (8.7)	78 (10.3)	
Language {Other}	English	758 (97.1)	1358 (96.3)	0.428 (0.072)	740 (97.4)	730 (96.1)	0.196 (0.074)
Gender {Female}	Male	371 (47.5)	615 (43.6)	0.088 (0.078)	362 (47.6)	334 (43.9)	0.165(0.074)
Marital Status {married}	Single	197 (25.2)	308 (21.8)	0.081 (0.080)	182 (23.9)	185 (24.3)	0.905 (0.009)

Before matching: ABOS =781 and online (follow up) = 1,410 respondents

After matching: ABOS = 760 and online (follow up) = 760 respondents



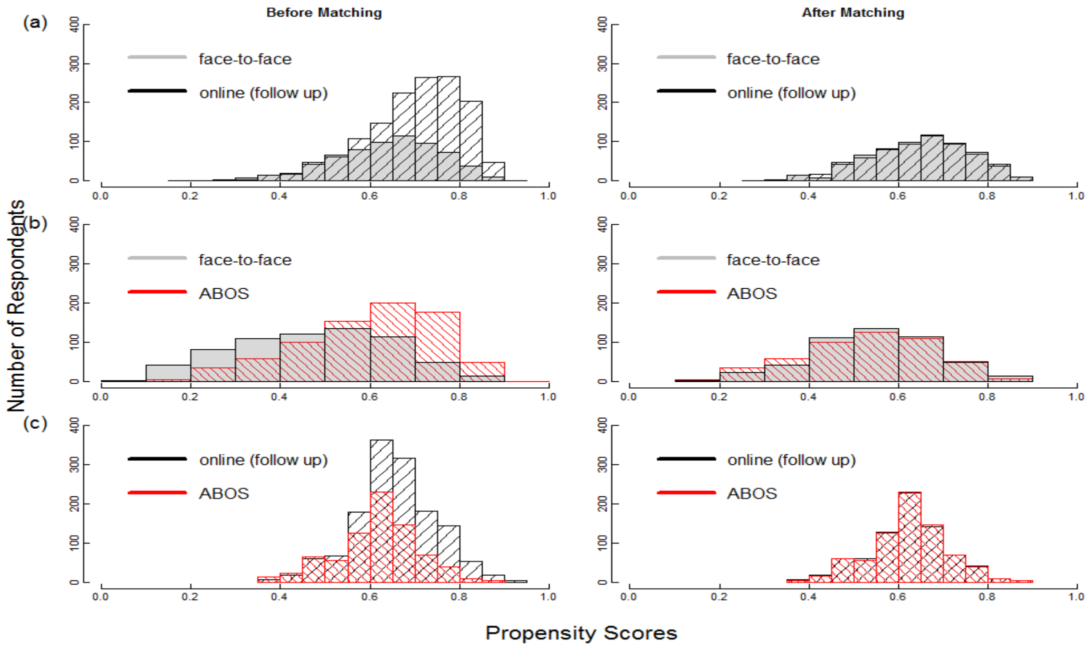
Boxplots of propensity scores distributions before and after matching for face-to-face and online (follow up) (top panel), face-to-face and ABOS (middle panel) and ABOS and online (follow up)

C.2 Results for Propensity Score Model without weights

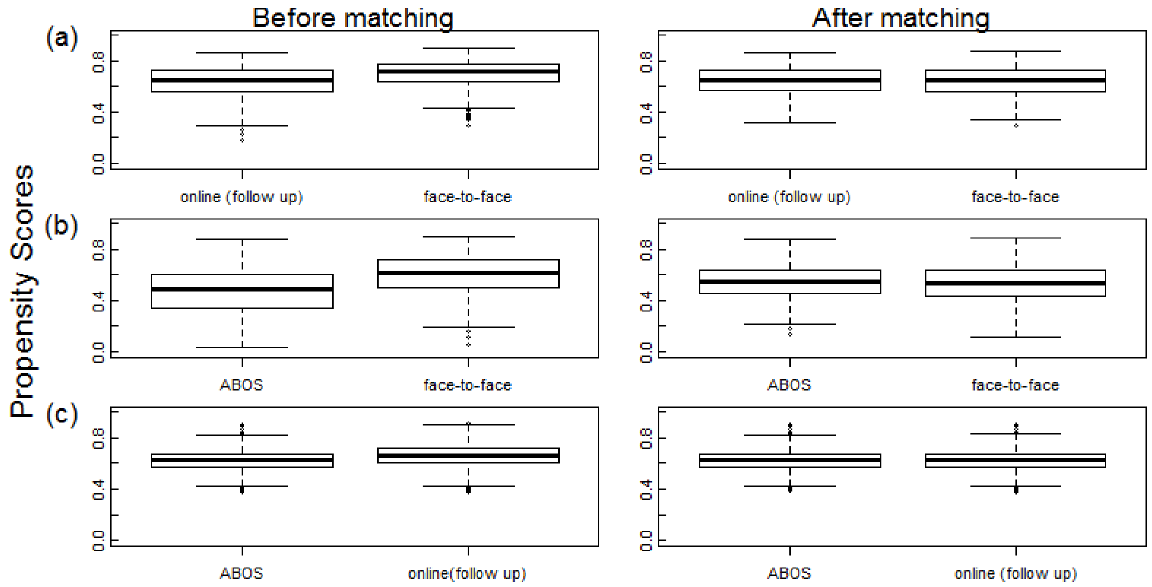
The sample sizes before and after matching (weight as covariate model)

	Face-to-face and online (follow up)		**Face-to-face and ABOS**		**ABOS and online (follow up)**	
	Face-to-face	*Online (follow up)*	*Face-to-face*	*ABOS*	*ABOS*	*Online (follow up)*
Before matching	666	1,410	666	781	781	1,410
After matching	647	647	492	492	760	760
Discarded	15(2.9%)	763(54.1%)	174(25.8%)	289(36.7%)	31(4.0%)	650(46%)

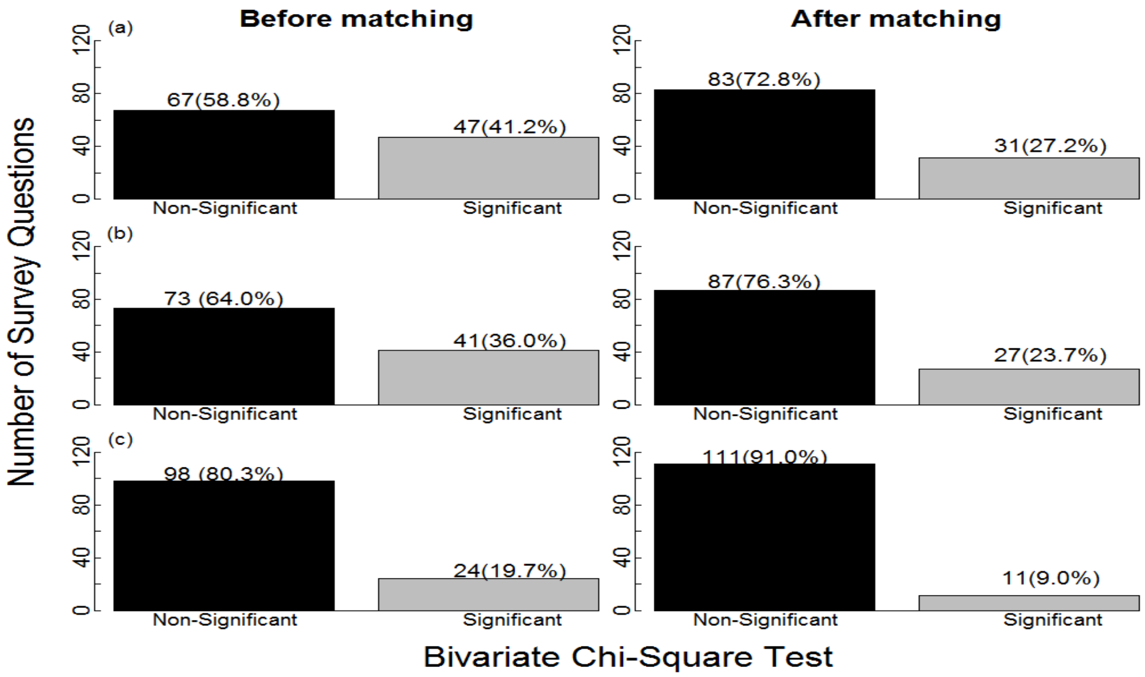
Appendix C



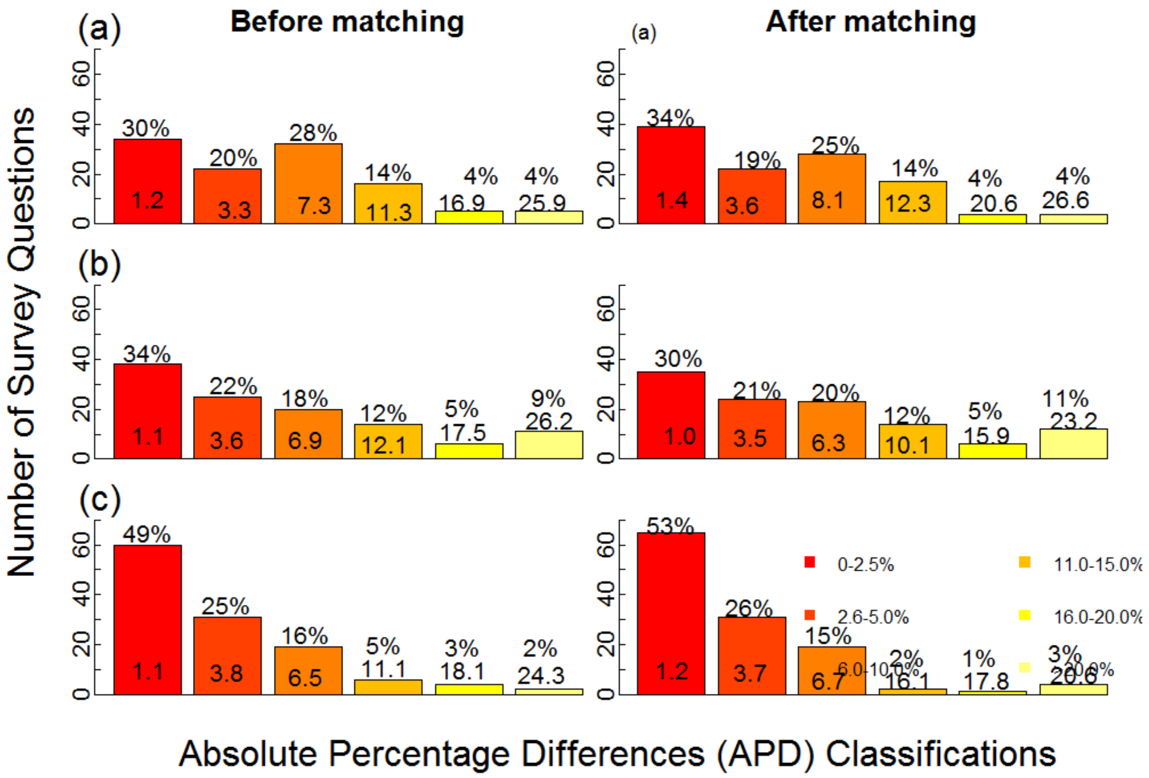
Histograms of propensity scores distributions for model without weights before and after matching for face-to-face and online (follow up) (top panel), face-to-face and ABOS (middle panel) and ABOS and online (follow up)



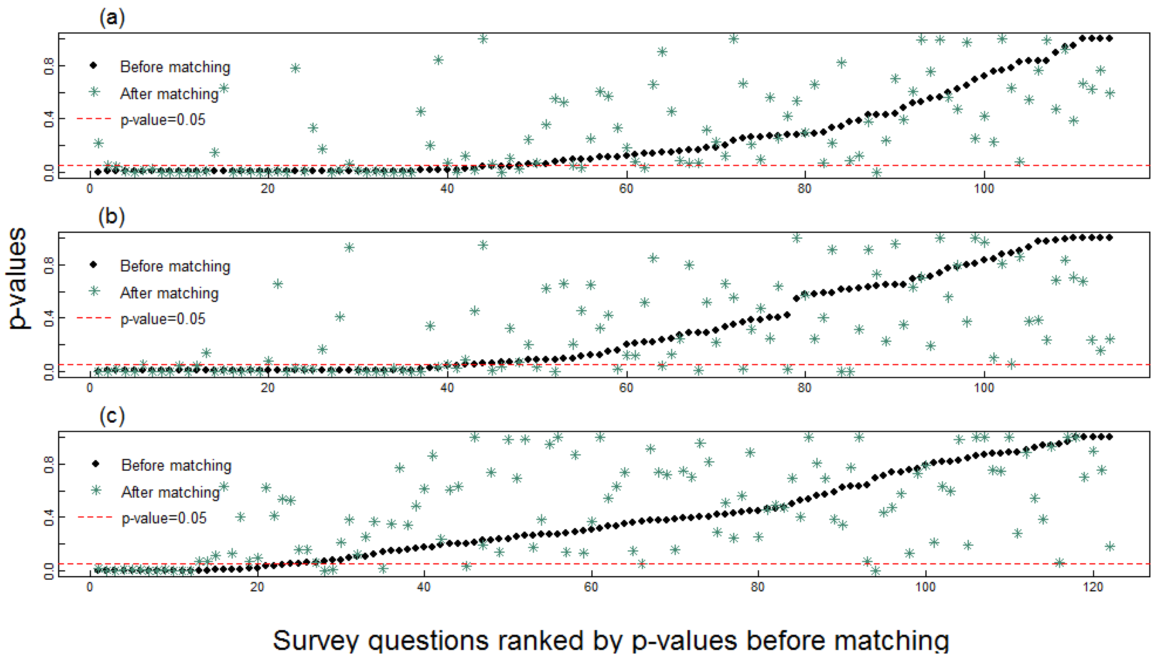
Boxplots of propensity scores distributions for model without weights before and after matching for face-to-face and online (follow up) (top panel), face-to-face and ABOS (middle panel) and ABOS and online (follow up)



Barplots of bivariate chi-square tests of the survey questions before and after matching (unweighted model) for face-to-face and online (follow up) (a), face-to-face and ABOS (b), and online (follow up) and ABOS (c).



Barplots of Absolute Percentage Differences (APD) classifications with corresponding medians and percentages before and after matching (unweighted model) for face-to-face and online (follow up) (a), face-to-face and ABOS (b), and online (follow up) and ABOS (c)



P-values by survey questions before and after matching (unweighted model) for face-to-face and online (follow up) (a), face-to-face and ABOS (b), and online (follow up) and ABOS (c)

SMD for baseline covariates for face-to-face and online (follow up) samples before and after matching (model without weights)-Greedy Nearest Neighbour Matching

Variable {Ref}	Categories	Before Matching			After Matching (weighted model)		
		Face-to-face	*Online follow up*	*P-value (SMD)*	*Face-to- Face*	*Online follow up*	*P-value (SMD)*
		Freq (%)	*Freq (%)*		*Freq (%)*	*Freq (%)*	
Age	16 to 34 years	156 (23.4)	226 (16.0)	**0.001**	143 (22.1)	146 (22.6)	0.983
	35 to 49 years	142 (21.3)	387 (27.4)	(0.273)	142 (21.9)	142 (21.9)	(0.035)
	50 to 64 years	168 (25.2)	415 (29.4)		167 (25.8)	167 (25.8)	
	65 to 74 years	107 (16.1)	255 (18.1)		105 (16.2)	109 (16.8)	
	Over 75 years	93 (14.0)	127 (9.0)		90 (13.9)	83 (12.8)	
Race {Others}	White	579 (86.9)	1297 (92.0)	**0.001** (0.165)	573 (88.6)	563 (87.0)	0.445 (0.047)
Number of adults in household	1	228 (34.2)	349 (24.8)	**0.001**	220 (34.0)	217 (33.5)	0.997
	2	331 (49.7)	817 (57.9)	(0.218)	325 (50.2)	328 (50.7)	(0.013)
	3	72 (10.8)	149 (10.6)		67 (10.4)	68 (10.5)	
	4 or more	35 (5.3)	95 (6.7)		35 (5.4)	34 (5.3)	
Income	0 to < £15K	302 (45.3)	596 (42.3)	**0.001**	294 (45.4)	281 (43.4)	0.897
	£15K to <£40K	206 (30.9)	529 (37.5)	(0.158)	204 (31.5)	209 (32.3)	(0.043)
	>£40K	62 (9.3)	163 (11.6)		62 (9.6)	67 (10.4)	
	No data	96 (14.4)	122 (8.7)		87 (13.4)	90 (13.9)	
Education	No Qualifications	255 (38.3)	380 (27.0)	**0.001**	241 (37.2)	231 (35.7)	0.518
	Other Qualifications	284 (42.6)	645 (45.7)	(0.269)	280 (43.3)	300 (46.4)	(0.064)
	Degree or above	127 (19.1)	385 (27.3)		126 (19.5)	116 (17.9)	
GOR	London	90 (13.5)	142 (10.1)	**0.007**	80 (12.4)	84 (13.0)	0.954
	East Midlands	53 (8.0)	103 (7.3)	(0.218)	53 (8.2)	67 (10.4)	(0.091)
	East of England	81 (12.2)	165 (11.7)		81 (12.5)	76 (11.7)	
	North East	39 (5.9)	76 (5.4)		39 (6.0)	42 (6.5)	
	North West	88 (13.2)	197 (14.0)		87 (13.4)	84 (13.0)	
	South East	87 (13.1)	266 (18.9)		86 (13.3)	78 (12.1)	
	South West	56 (8.4)	154 (10.9)		56 (8.7)	53 (8.2)	
	West Midlands	92 (13.8)	154 (10.9)		85 (13.1)	87 (13.4)	
	Yorkshire and Humberside	80 (12.0)	153 (10.9)		80 (12.4)	76 (11.7)	
Number of children	0	491 (73.7)	1014 (71.9)	0.755 (0.052)	480 (74.2)	463 (71.6)	0.539 (0.082)
	1	76 (11.4)	184 (13.0)		74 (11.4)	91 (14.1)	
	2	71 (10.7)	151 (10.7)		67 (10.4)	65 (10.0)	
	3 or more	28 (4.2)	61 (4.3)		26 (4.0)	28 (4.3)	
Paid work {No}	Yes	339 (50.9)	781 (55.4)	0.062 (0.090)	333 (51.5)	323 (49.9)	0.617 (0.031)
Tenure	private rent	150 (22.5)	243 (17.2)	**0.001**	142 (21.9)	144 (22.3)	0.987
	Mortgaged	172 (25.8)	462 (32.8)	(0.267)	170 (26.3)	168 (26.0)	(0.020)
	Outright ownership	226 (33.9)	551 (39.1)		224 (34.6)	228 (35.2)	
	Social rent	118 (17.7)	154 (10.9)		111 (17.2)	107 (16.5)	
Language {Other}	English	627 (94.1)	1358 (96.3)	**0.033** (0.102)	614 (94.9)	609 (94.1)	0.625 (0.034)
Gender {Female}	Male	287 (43.1)	615 (43.6)	0.859 (0.011)	280 (43.3)	281 (43.4)	1.000 (0.003)
Marital Status {married}	Single	197 (29.6)	308 (21.8)	**0.001** (0.178)	188 (29.1)	178 (27.5)	0.579 (0.034)

Before matching: face-to-face = 666 and Online (follow up) =1,410 respondents
After matchin: face-to-face = 647 and Online (follow up) = 647 respondents

Appendix C

SMD for baseline covariates for face-to-face and ABOS samples before and after matching (model without weights)-Greedy Nearest Neighour Matching

Variable {Ref}	Categories	Before Matching			After Matching (weighted model)		
		Face-to-face	*Online (ABOS)*	*P-value (SMD)*	*Face-to-Face*	*Online (ABOS)*	*P-value (SMD)*
		Freq (%)	Freq (%)		Freq (%)	Freq (%)	
Age	16 to 34 years	156 (23.4)	175 (22.4)	**0.001**	118 (24.0)	110 (22.4)	0.696
	35 to 49 years	142 (21.3)	203 (26.0)	(0.243)	118 (24.0)	107 (21.7)	(0.095)
	50 to 64 years	168 (25.2)	206 (26.4)		125 (25.4)	141 (28.7)	
	65 to 74 years	107 (16.1)	142 (18.2)		87 (17.7)	84 (17.1)	
	Over 75 years	93 (14.0)	127 (9.0)		44 (8.9)	50 (10.2)	
Race {Others}	White	579 (86.9)	712 (91.2)	**0.012** (0.136)	444 (90.2)	438 (89.0)	0.601 (0.040)
Number of adults in household	1	228 (34.2)	120 (15.4)	**0.001**	102 (20.7)	114 (23.2)	0.777
	2	331 (49.7)	461 (59.0)	(0.477)	289 (58.7)	285 (57.9)	(0.067)
	3	72 (10.8)	110 (14.1)		66 (13.4)	59 (12.0)	
	4 or more	35 (5.3)	90 (11.5)		35 (7.1)	34 (6.9)	
Income	0 to < £15K	302 (45.3)	305 (39.1)	**0.002**	201 (40.9)	204 (41.5)	0.976
	£15K to <£40K	206 (30.9)	308 (39.4)	(0.205)	170 (34.6)	173 (35.2)	(0.029)
	>£40K	62 (9.3)	83 (10.6)		54 (11.0)	52 (10.6)	
	No data	96 (14.4)	85 (10.9)		67 (13.6)	63 (12.8)	
Education	No Qualifications	255 (38.3)	199 (25.5)	**0.001**	148 (30.1)	156 (31.7)	0.857
	Other Qualifications	284 (42.6)	341 (43.7)	(0.332)	227 (46.1)	221 (44.9)	(0.035)
	Degree or above	127 (19.1)	241 (30.9)		117 (23.8)	115 (23.4)	
GOR	London	90 (13.5)	85 (10.9)	**0.001**	62 (12.6)	63 (12.8)	0.998
	East Midlands	53 (8.0)	65 (8.3)	(0.308)	41 (8.3)	37 (7.5)	(0.066)
	East of England	81 (12.2)	77 (9.9)		60 (12.1)	63 (12.8)	
	North East	39 (5.9)	30 (3.8)		24 (4.9)	26 (5.3)	
	North West	88 (13.2)	128 (16.4)		73 (14.8)	67 (13.6)	
	South East	87 (13.1)	160 (20.5)		78 (15.8)	80 (16.2)	
	South West	56 (8.4)	87 (11.1)		48 (9.7)	46 (9.3)	
	West Midlands	92 (13.8)	67 (8.6)		48 (9.7)	54 (10.9)	
	Yorkshire and Humberside	80 (12.0)	82 (10.5)		60 (12.1)	58 (11.7)	
Number of children	0	491 (73.7)	594 (76.1)	**0.023**	362 (73.6)	376 (76.4)	0.462
	1	76 (11.4)	91 (11.7)	(0.160)	60 (12.2)	59 (12.0)	(0.102)
	2	71 (10.7)	84 (10.8)		61 (12.4)	46 (9.3)	
	3 or more	28 (4.2)	12 (1.5)		9 (1.8)	11 (2.2)	
Paid work {No}	Yes	339 (50.9)	443 (56.7)	**0.031** (0.117)	289 (58.5)	271 (54.9)	0.275(0.074)
Tenure	private rent	150 (22.5)	176 (22.5)	**0.001**	112 (22.8)	105 (21.3)	0.354(0.115)
	Mortgaged	172 (25.8)	238 (30.5)	(0.271)	142 (28.9)	146 (29.7)	
	Outright ownership	226 (33.9)	298 (38.2)		177 (36.0)	195 (39.6)	
	Social rent	118 (17.7)	69 (8.8)		61 (12.4)	46 (9.3)	
Language {Other}	English	627 (94.1)	758 (97.1)	**0.009** (0.142)	469 (95.3)	474 (96.3)	0.523 (0.051)
Gender {Female}	Male	287 (43.1)	371 (47.5)	0.104 (0.089)	222 (45.1)	229 (46.5)	0.701(0.029)
Marital Status {married}	Single	197 (29.6)	197 (25.2)	0.073 (0.098)	140 (28.5)	127 (25.8)	0.390 (0.059)

Before matching: face-to-face = 666 and ABOS =781 respondents
After matching: face-to-face = 492 and ABOS = 492 respondents

SMD for baseline covariates for ABOS and online (follow up) samples before and after matching (model without weights)-Greedy Nearest Neighour Matching

Variable {Ref}	**Categories**	**Before Matching**			**After Matching (weighted model)**		
		ABOS	*Online follow up*	*P-value (SMD)*	*ABOS*	*Online follow up*	*P-value (SMD)*
		Freq (%)	Freq (%)		Freq (%)	Freq (%)	
Age	16 to 34 year*s*	175 (22.4)	226 (16.0)	**0.004**	156 (20.5)	155 (20.4)	0.944
	35 to 49 years	203 (26.0)	387 (27.4)	(0.174)	201 (26.4)	196 (25.8)	(0.045)
	50 to 64 years	206 (26.4)	415 (29.4)		206 (27.1)	202 (26.6)	
	65 to 74 years	142 (18.2)	255 (18.1)		142 (18.7)	155 (20.4)	
	Over 75 years	55 (7.0)	127 (9.0)		55 (7.2)	52 (6.8)	
Race {Others}	White	712 (91.2)	1297 (92.0)	0.558			1.000
				(0.030)	695 (91.4)	694 (91.3)	(0.005)
Number of adults in household	1	120 (15.4)	349 (24.8)	**0.001**	120 (15.8)	123 (16.2)	0.852
	2	461 (59.0)	817 (57.9)	(0.284)	461 (60.7)	455 (59.9)	(0.046)
	3	110 (14.1)	149 (10.6)		109 (14.3)	103 (13.6)	
	4 or more	90 (11.5)	95 (6.7)		120 (15.8)	123 (16.2)	
Income	0 to < £15K	305 (39.1)	596 (42.3)	0.118	290 (38.2)	332 (43.7)	**0.001**
	£15K to <£40K	308 (39.4)	529 (37.5)	(0.097)	303 (39.9)	288 (37.9)	(0.174)
	>£40K	83 (10.6)	163 (11.6)		82 (10.8)	88 (11.6)	
	No data	85 (10.9)	122 (8.7)		85 (11.2)	52 (6.8)	
Education	No Qualifications	199 (25.5)	380 (27.0)	0.211(0.078)	198 (26.1)	197 (25.9)	0.177
	Other	341 (43.7)	645 (45.7)				(0.096)
	Qualifications				329 (43.3)	360 (47.4)	
	Degree or above	241 (30.9)	385 (27.3)		233 (30.7)	203 (26.7)	
GOR	London	85 (10.9)	142 (10.1)	0.231	85 (11.2)	79 (10.4)	0.966
	East Midlands	65 (8.3)	103 (7.3)	(0.146)	62 (8.2)	59 (7.8)	(0.079)
	East of England	77 (9.9)	165 (11.7)		75 (9.9)	70 (9.2)	
	North East	30 (3.8)	76 (5.4)		30 (3.9)	25 (3.3)	
	North West	128 (16.4)	197 (14.0)		120 (15.8)	120 (15.8)	
	South East	160 (20.5)	266 (18.9)		159 (20.9)	155 (20.4)	
	South West	87 (11.1)	154 (10.9)		81 (10.7)	95 (12.5)	
	West Midlands	67 (8.6)	154 (10.9)		67 (8.8)	73 (9.6)	
	Yorkshire and	82 (10.5)	153 (10.9)				
	Humberside				574 (75.5)	588 (77.4)	
Number of children	0	594 (76.1)	1014 (71.9)	**0.003** (0.175)	91 (12.0)	78 (10.3)	0.750 (0.057)
	1	91 (11.7)	184 (13.0)		83 (10.9)	83 (10.9)	
	2	84 (10.8)	151 (10.7)		12 (1.6)	11 (1.4)	
	3 or more	12 (1.5)	61 (4.3)		574 (75.5)	588 (77.4)	
Paid work {No}	Yes	443 (56.7)	781 (55.4)	0.578 (0.027)	433 (57.0)	423 (55.7)	0.642 (0.027)
Tenure	private rent	176 (22.5)	243 (17.2)	**0.015** (0.143)	164 (21.6)	146 (19.2)	0.415 (0.087)
	Mortgaged	238 (30.5)	462 (32.8)		234 (30.8)	251 (33.0)	
	Outright ownership	298 (38.2)	551 (39.1)		296 (38.9)	285 (37.5)	
	Social rent	69 (8.8)	154 (10.9)		66 (8.7)	78 (10.3)	
Language {Other}	English	758 (97.1)	1358 (96.3)	0.428			0.196
				(0.072)	740 (97.4)	730 (96.1)	(0.074)
Gender {Female}	Male	371 (47.5)	615 (43.6)	0.088			0.165
				(0.078)	362 (47.6)	334 (43.9)	(0.074)
Marital Status {married}	Single	197 (25.2)	308 (21.8)	0.081			0.905
				(0.080)	182 (23.9)	185 (24.3)	(0.009)

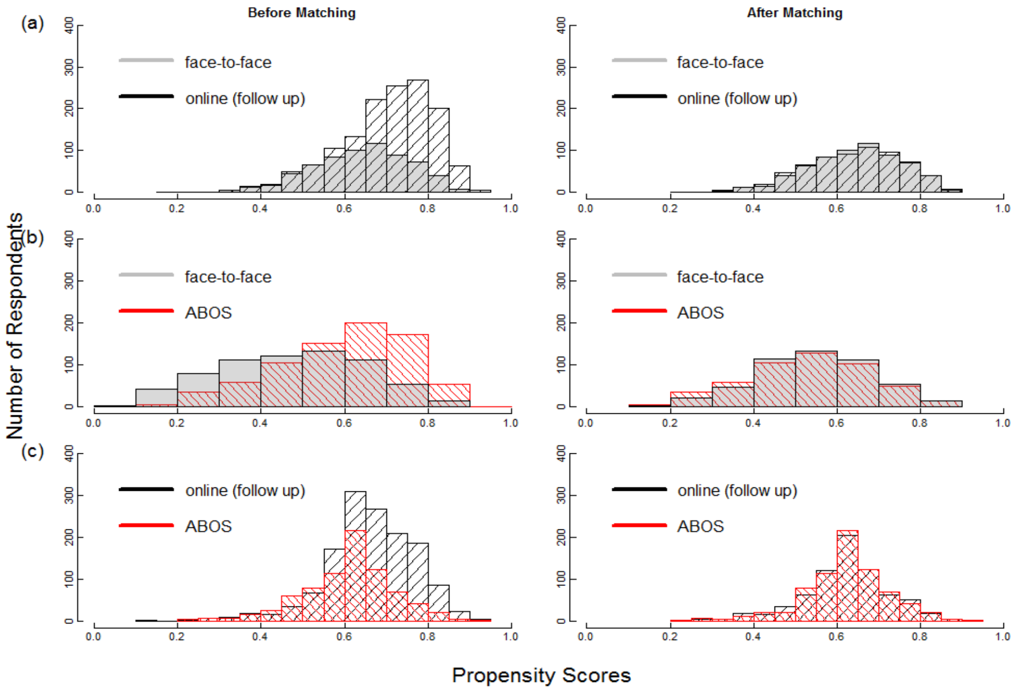
Before matching: ABOS =781 and online (follow up) = 1,410 respondents

After matching: ABOS = 760 and online (follow up) = 760 respondents

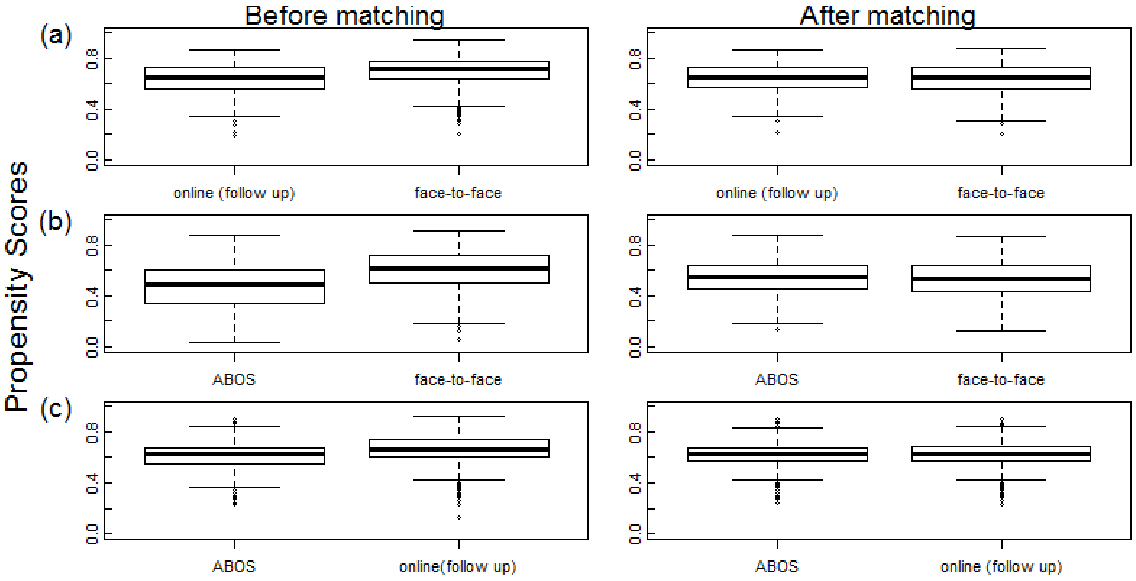
C.3 Results for Propensity Score Model with weights specified as covariate

The sample sizes before and after matching (weight as covariate model)

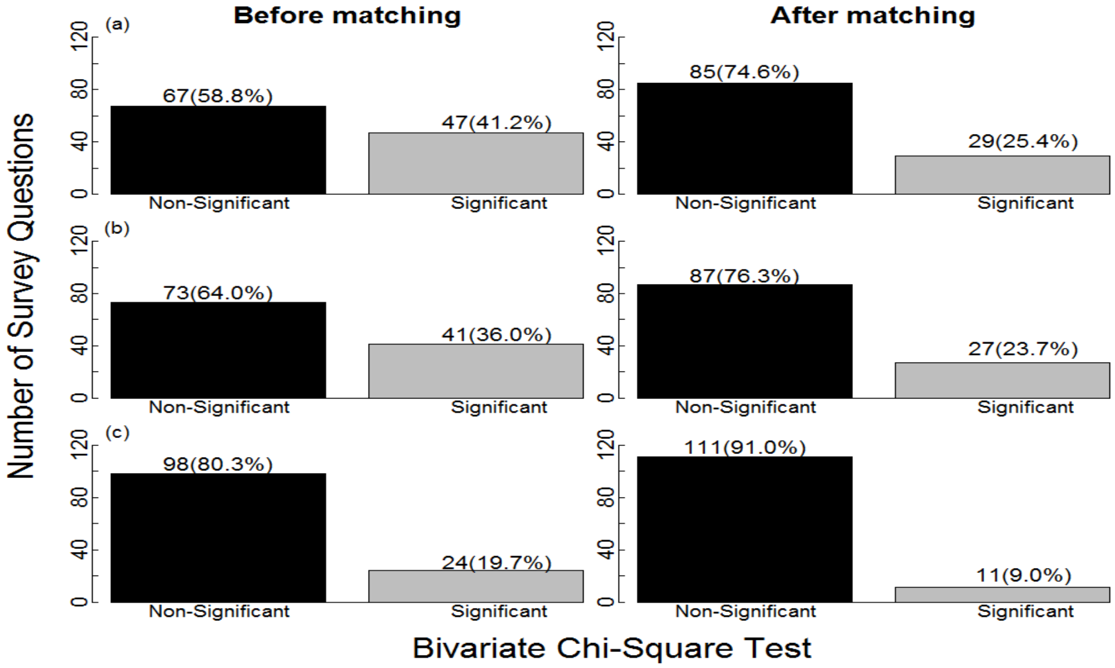
	Face-to-face and online (follow up)		Face-to-face and ABOS		ABOS and online (follow up)	
	Face-to-face	*Online (follow up)*	*Face-to-face*	*ABOS*	*ABOS*	*Online (follow up)*
Before matching	666	1,410	666	781	781	1,410
After matching	651	651	494	494	726	726
Discarded	15(2.3%)	761(53.8%)	174(25.8%)	289(36.7%)	55(7.0%)	684(48%)



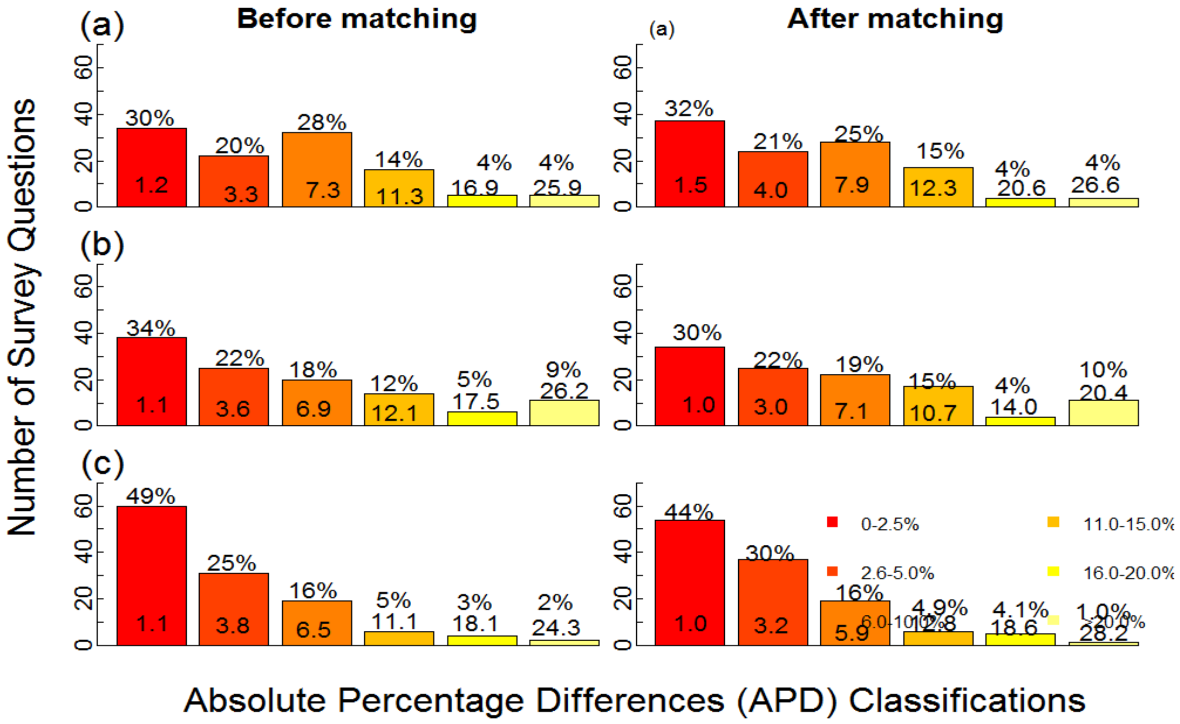
Histograms of propensity scores distributions for weight as covariate model before and after matching for face-to-face and online (follow up) (top panel), face-to-face and ABOS (middle panel) and ABOS and online (follow up)



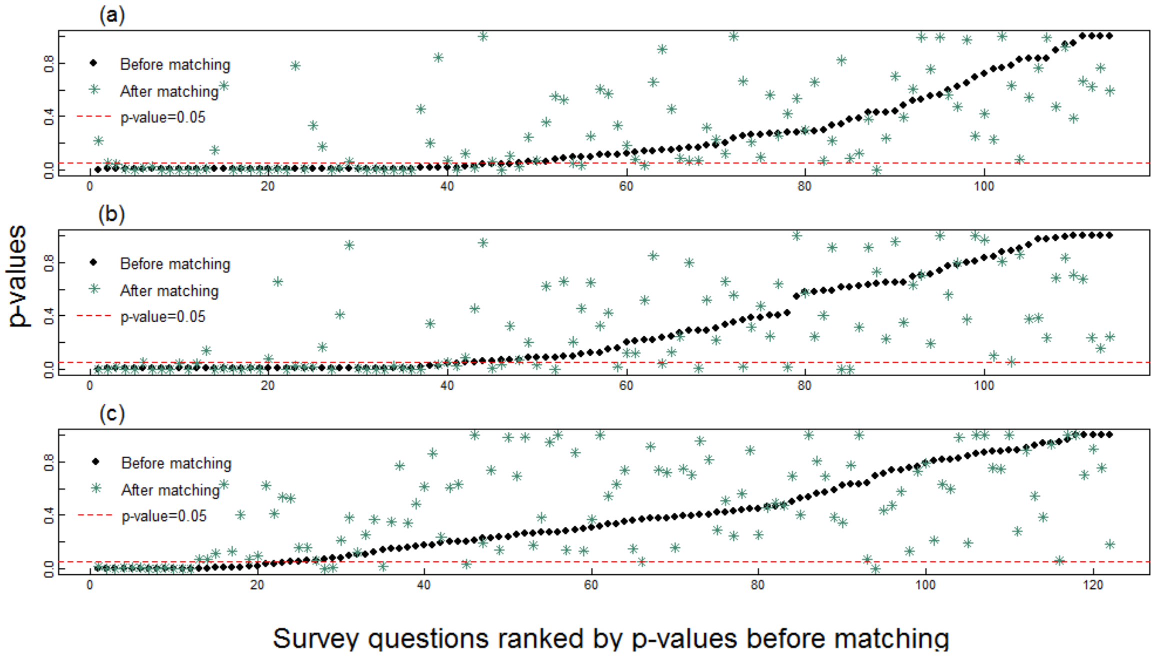
Boxplots of propensity scores distributions for weight as covariate model before and after matching for face-to-face and online (follow up) (top panel), face-to-face and ABOS (middle panel) and ABOS and online (follow up)



Barplots of bivariate chi-square tests of the survey questions before and after matching (weight as covariates model) for face-to-face and online (follow up) (a), face-to-face and ABOS (b), and online (follow up) and ABOS (c).



Barplots of Absolute Percentage Differences (APD) classifications with corresponding medians and percentages before and after matching (weight as covariates model) for face-to-face and online (follow up) (a), face-to-face and ABOS (b), and online (follow up) and ABOS (c)



P-values by survey questions before and after matching (weight as covariates model) for face-to-face and online (follow up) (a), face-to-face and ABOS (b), and online (follow up) and ABOS (c)

SMD for baseline covariates for face-to-face and online (follow up) samples before and after matching (weight as a covariate model)-Greedy Nearest Neighbour Matching

Variable {Ref}	Categories	Before Matching			After Matching (weighted model)		
		Face-to-face	*Online follow up*	*P-value (SMD)*	*Face-to- Face*	*Online follow up*	*P-value (SMD)*
		Freq (%)	*Freq (%)*		*Freq (%)*	*Freq (%)*	
Age	16 to 34 years	156 (23.4)	226 (16.0)	**0.001**	146 (22.4)	150 (23.0)	0.840
	35 to 49 years	142 (21.3)	387 (27.4)	(0.273)	142 (21.8)	130 (20.0)	(0.066)
	50 to 64 years	168 (25.2)	415 (29.4)		167 (25.7)	166 (25.5)	
	65 to 74 years	107 (16.1)	255 (18.1)		107 (16.4)	120 (18.4)	
	Over 75 years	93 (14.0)	127 (9.0)		90 (13.9)	85 (13.1)	
Race {Others}	White	579 (86.9)	1297 (92.0)	**0.001** (0.165)	573 (88.0)	572 (87.9)	1.000 (0.057)
Number of adults in household	1	228 (34.2)	349 (24.8)	**0.001**	221 (33.9)	216 (33.2)	0.728
	2	331 (49.7)	817 (57.9)	(0.218)	326 (50.1)	336 (51.6)	(0.063)
	3	72 (10.8)	149 (10.6)		69 (10.6)	72 (11.1)	
	4 or more	35 (5.3)	95 (6.7)		35 (5.4)	27 (4.1)	
Income	0 to < £15K	302 (45.3)	596 (42.3)	**0.001**	292 (44.9)	315 (48.4)	0.515(0.084)
	£15K to <£40K	206 (30.9)	529 (37.5)	(0.158)	206 (31.6)	183 (28.1)	
	>£40K	62 (9.3)	163 (11.6)		62 (9.5)	64 (9.8)	
	No data	96 (14.4)	122 (8.7)		91 (14.0)	89 (13.7)	
Education	No Qualifications	255 (38.3)	380 (27.0)	**0.001**	244 (37.5)	243 (37.3)	0.732(0.044)
	Other Qualifications	284 (42.6)	645 (45.7)	(0.269)	280 (43.0)	291 (44.7)	
	Degree or above	127 (19.1)	385 (27.3)		127 (19.5)	117 (18.0)	
GOR	London	90 (13.5)	142 (10.1)	**0.007**	86 (13.2)	86 (13.2)	0.953(0.091)
	East Midlands	53 (8.0)	103 (7.3)	(0.218)	51 (7.8)	63 (9.7)	
	East of England	81 (12.2)	165 (11.7)		80 (12.3)	70 (10.8)	
	North East	39 (5.9)	76 (5.4)		39 (6.0)	41 (6.3)	
	North West	88 (13.2)	197 (14.0)		88 (13.5)	88 (13.5)	
	South East	87 (13.1)	266 (18.9)		86 (13.2)	76 (11.7)	
	South West	56 (8.4)	154 (10.9)		56 (8.6)	57 (8.8)	
	West Midlands	92 (13.8)	154 (10.9)		86 (13.2)	89 (13.7)	
	Yorkshire and Humberside	80 (12.0)	153 (10.9)		79 (12.1)	81 (12.4)	
Number of children	0	491 (73.7)	1014 (71.9)	0.755 (0.052)	482 (74.0)	482 (74.0)	0.785(0.057)
	1	76 (11.4)	184 (13.0)		74 (11.4)	82 (12.6)	
	2	71 (10.7)	151 (10.7)		68 (10.4)	59 (9.1)	
	3 or more	28 (4.2)	61 (4.3)		27 (4.1)	28 (4.3)	
Paid work {No}	Yes	339 (50.9)	781 (55.4)	0.062 (0.090)	334 (51.3)	303 (46.5)	0.096(0.095)
Tenure	private rent	150 (22.5)	243 (17.2)	**0.001**	143 (22.0)	142 (21.8)	0.849
	Mortgaged	172 (25.8)	462 (32.8)	(0.267)	172 (26.4)	159 (24.4)	(0.050)
	Outright ownership	226 (33.9)	551 (39.1)		225 (34.6)	234 (35.9)	
	Social rent	118 (17.7)	154 (10.9)		111 (17.1)	116 (17.8)	
Language {Other}	English	627 (94.1)	1358 (96.3)	**0.033** (0.102)	615 (94.5)	617 (94.8)	0.902(0.014)
Gender {Female}	Male	287 (43.1)	615 (43.6)	0.859 (0.011)	282 (43.3)	286 (43.9)	1.000 (0.005)
Marital Status {married}	Single	197 (29.6)	308 (21.8)	**0.001** (0.178)	190 (29.2)	176 (27.0)	0.423(0.048)

Before matching: face-to-face = 666 and Online (follow up) =1,410 respondents
After matchin: face-to-face = 651 and Online (follow up) = 651 respondents

Appendix C

SMD for baseline covariates for face-to-face and ABOS samples before and after matching (weighted as a covariate model) -Greedy Nearest Neighour Matching

Variable {Ref}	**Categories**	**Before Matching**			**After Matching (weighted model)**		
		Face-to-face	*Online (ABOS)*	*P-value (SMD)*	*Face-to-Face*	*Online (ABOS)*	*P-value (SMD)*
		Freq (%)	Freq (%)		Freq (%)	Freq (%)	
Age	16 to 34 years	156 (23.4)	175 (22.4)	**0.001**	114 (23.1)	112 (22.7)	0.850
	35 to 49 years	142 (21.3)	203 (26.0)	(0.243)	115 (23.3)	110 (22.3)	(0.074)
	50 to 64 years	168 (25.2)	206 (26.4)		130 (26.3)	137 (27.7)	
	65 to 74 years	107 (16.1)	142 (18.2)		93 (18.8)	85 (17.2)	
	Over 75 years	93 (14.0)	127 (9.0)		42 (8.5)	50 (10.1)	
Race {Others}	White	579 (86.9)	712 (91.2)	**0.012** (0.136)	443 (89.7)	440 (89.1)	0.836 (0.020)
Number of adults in household	1	228 (34.2)	120 (15.4)	**0.001**	100 (20.2)	111 (22.5)	0.743
	2	331 (49.7)	461 (59.0)	(0.477)	292 (59.1)	289 (58.5)	(0.071)
	3	72 (10.8)	110 (14.1)		68 (13.8)	59 (11.9)	
	4 or more	35 (5.3)	90 (11.5)		34 (6.9)	35 (7.1)	
Income	0 to < £15K	302 (45.3)	305 (39.1)	**0.002**	203 (41.1)	204 (41.3)	0.937
	£15K to <£40K	206 (30.9)	308 (39.4)	(0.205)	167 (33.8)	168 (34.0)	(0.041)
	>£40K	62 (9.3)	83 (10.6)		52 (10.5)	56 (11.3)	
	No data	96 (14.4)	85 (10.9)		72 (14.6)	66 (13.4)	
Education	No Qualifications	255 (38.3)	199 (25.5)	**0.001**	151 (30.6)	159 (32.2)	0.785
	Other Qualifications	284 (42.6)	341 (43.7)	(0.332)	224 (45.3)	224 (45.3)	(0.044)
	Degree or above	127 (19.1)	241 (30.9)		119 (24.1)	111 (22.5)	
GOR	London	90 (13.5)	85 (10.9)	**0.001**	62 (12.6)	63 (12.8)	0.998
	East Midlands	53 (8.0)	65 (8.3)	(0.308)	41 (8.3)	37 (7.5)	(0.066)
	East of England	81 (12.2)	77 (9.9)		60 (12.1)	63 (12.8)	
	North East	39 (5.9)	30 (3.8)		24 (4.9)	26 (5.3)	
	North West	88 (13.2)	128 (16.4)		73 (14.8)	67 (13.6)	
	South East	87 (13.1)	160 (20.5)		78 (15.8)	80 (16.2)	
	South West	56 (8.4)	87 (11.1)		48 (9.7)	46 (9.3)	
	West Midlands	92 (13.8)	67 (8.6)		48 (9.7)	54 (10.9)	
	Yorkshire and Humberside	80 (12.0)	82 (10.5)		60 (12.1)	58 (11.7)	
Number of children	0	491 (73.7)	594 (76.1)	**0.023**	362 (73.6)	376 (76.4)	0.462
	1	76 (11.4)	91 (11.7)	(0.160)	60 (12.2)	59 (12.0)	(0.102)
	2	71 (10.7)	84 (10.8)		61 (12.4)	46 (9.3)	
	3 or more	28 (4.2)	12 (1.5)		9 (1.8)	11 (2.2)	
Paid work {No}	Yes	339 (50.9)	443 (56.7)	**0.031** (0.117)	289 (58.5)	271 (54.9)	0.275(0.074)
Tenure	private rent	150 (22.5)	176 (22.5)	**0.001**	114 (23.1)	113 (22.9)	0.773(0.067)
	Mortgaged	172 (25.8)	238 (30.5)	(0.271)	140 (28.3)	145 (29.4)	
	Outright ownership	226 (33.9)	298 (38.2)		181 (36.6)	187 (37.9)	
	Social rent	118 (17.7)	69 (8.8)		59 (11.9)	49 (9.9)	
Language {Other}	English	627 (94.1)	758 (97.1)	**0.009** (0.142)	470 (95.1)	478 (96.8)	0.259 (0.082)
Gender {Female}	Male	287 (43.1)	371 (47.5)	0.104 (0.089)	223 (45.1)	227 (46.0)	0.848 (0.016)
Marital Status {married}	Single	197 (29.6)	197 (25.2)	0.073 (0.098)	137 (27.7)	126 (25.5)	0.472 (0.050)

Before matching: face-to-face = 666 and ABOS =781 respondents
After matching: face-to-face = 494 and ABOS = 494 respondents

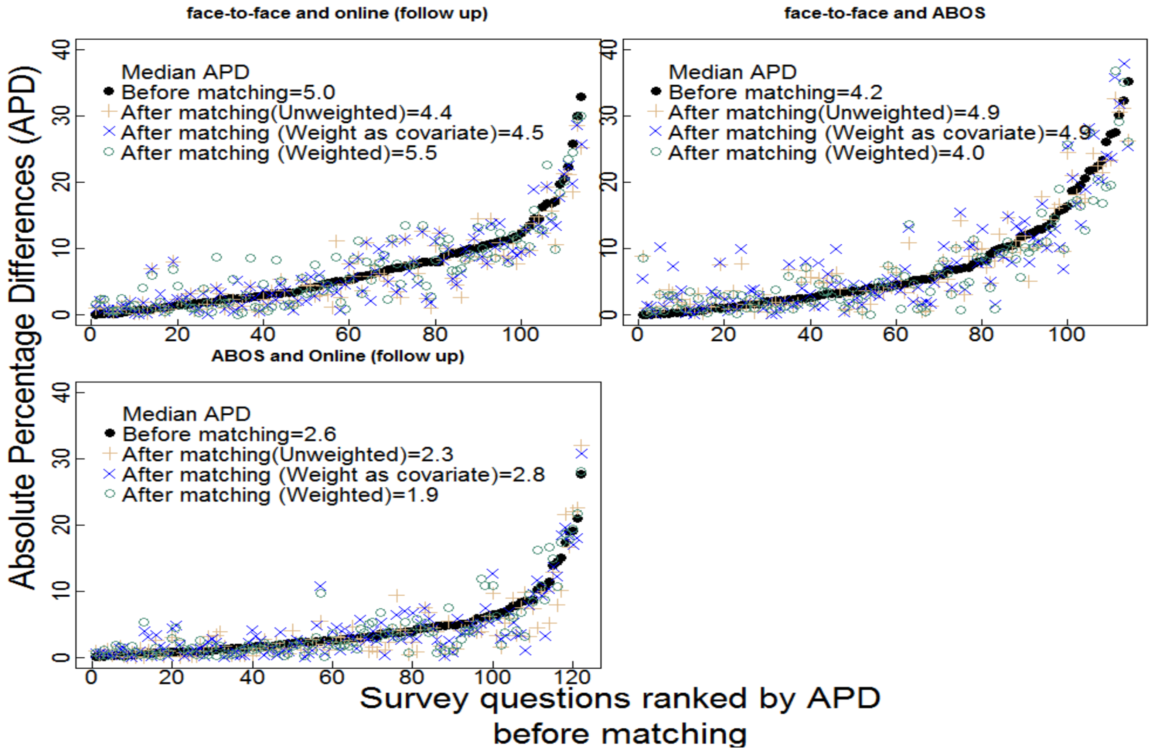
SMD for baseline covariates for ABOS and online (follow up) samples before and after matching (weight as covariate model)-Greedy Nearest Neighour Matching

Variable {Ref}	Categories	Before Matching			After Matching (weighted model)		
		ABOS	*Online follow up*	*P-value (SMD)*	*ABOS*	*Online follow up*	*P-value (SMD)*
		Freq (%)	Freq (%)		Freq (%)	Freq (%)	
Age	16 to 34 years	175 (22.4)	226 (16.0)	**0.004**	140 (19.3)	141 (19.4)	0.966
	35 to 49 years	203 (26.0)	387 (27.4)	(0.174)	193 (26.6)	193 (26.6)	(0.040)
	50 to 64 years	206 (26.4)	415 (29.4)		202 (27.8)	191 (26.3)	
	65 to 74 years	142 (18.2)	255 (18.1)		136 (18.7)	144 (19.8)	
	Over 75 years	55 (7.0)	127 (9.0)		55 (7.6)	57 (7.9)	
Race {Others}	White	712 (91.2)	1297 (92.0)	0.558			0.441
				(0.030)	663 (91.3)	672 (92.6)	(0.046)
Number of adults in household	1	120 (15.4)	349 (24.8)	**0.001**	120 (16.5)	104 (14.3)	0.452
	2	461 (59.0)	817 (57.9)	(0.284)	440 (60.6)	469 (64.6)	(0.085)
	3	110 (14.1)	149 (10.6)		97 (13.4)	91 (12.5)	
	4 or more	90 (11.5)	95 (6.7)		69 (9.5)	62 (8.5)	
Income	0 to < £15K	305 (39.1)	596 (42.3)	0.118	279 (38.4)	303 (41.7)	0.196
	£15K to <£40K	308 (39.4)	529 (37.5)	(0.097)	284 (39.1)	283 (39.0)	(0.114)
	>£40K	83 (10.6)	163 (11.6)		80 (11.0)	80 (11.0)	
	No data	85 (10.9)	122 (8.7)		83 (11.4)	60 (8.3)	
Education	No Qualifications	199 (25.5)	380 (27.0)	0.211(0.078)	191 (26.3)	190 (26.2)	0.313
	Other	341 (43.7)	645 (45.7)				(0.080)
	Qualifications				310 (42.7)	335 (46.1)	
	Degree or above	241 (30.9)	385 (27.3)		225 (31.0)	201 (27.7)	
GOR	London	85 (10.9)	142 (10.1)	0.231	80 (11.0)	83 (11.4)	0.978
	East Midlands	65 (8.3)	103 (7.3)	(0.146)	57 (7.9)	68 (9.4)	(0.076)
	East of England	77 (9.9)	165 (11.7)		76 (10.5)	72 (9.9)	
	North East	30 (3.8)	76 (5.4)		30 (4.1)	29 (4.0)	
	North West	128 (16.4)	197 (14.0)		106 (14.6)	105 (14.5)	
	South East	160 (20.5)	266 (18.9)		147 (20.2)	150 (20.7)	
	South West	87 (11.1)	154 (10.9)		82 (11.3)	73 (10.1)	
	West Midlands	67 (8.6)	154 (10.9)		67 (9.2)	61 (8.4)	
	Yorkshire and	82 (10.5)	153 (10.9)				
	Humberside				546 (75.2)	545 (75.1)	
Number of children	0	594 (76.1)	1014 (71.9)	**0.003** (0.175)	88 (12.1)	93 (12.8)	0.902 (0.040)
	1	91 (11.7)	184 (13.0)		80 (11.0)	79 (10.9)	
	2	84 (10.8)	151 (10.7)		12 (1.7)	9 (1.2)	
	3 or more	12 (1.5)	61 (4.3)		546 (75.2)	545 (75.1)	
Paid work	Yes	443 (56.7)	781 (55.4)	0.578	433 (57.0)	423 (55.7)	0.642
{No}				(0.027)			(0.027)
Tenure	private rent	176 (22.5)	243 (17.2)	**0.015** (0.143)	164 (21.6)	146 (19.2)	0.415 (0.087)
	Mortgaged	238 (30.5)	462 (32.8)		234 (30.8)	251 (33.0)	
	Outright ownership	298 (38.2)	551 (39.1)		296 (38.9)	285 (37.5)	
	Social rent	69 (8.8)	154 (10.9)		66 (8.7)	78 (10.3)	
Language	English	758 (97.1)	1358 (96.3)	0.428			0.767
{Other}				(0.072)	704 (97.0)	701 (96.6)	(0.023)
Gender	Male	371 (47.5)	615 (43.6)	0.088			0.371
{Female}				(0.078)	343 (47.2)	325 (44.8)	(0.050)
Marital Status	Single	197 (25.2)	308 (21.8)	0.081			0.281
{married}				(0.080)	168 (23.1)	150 (20.7)	(0.060)

Before matching: ABOS =781 and online (follow up) = 1,410 respondents

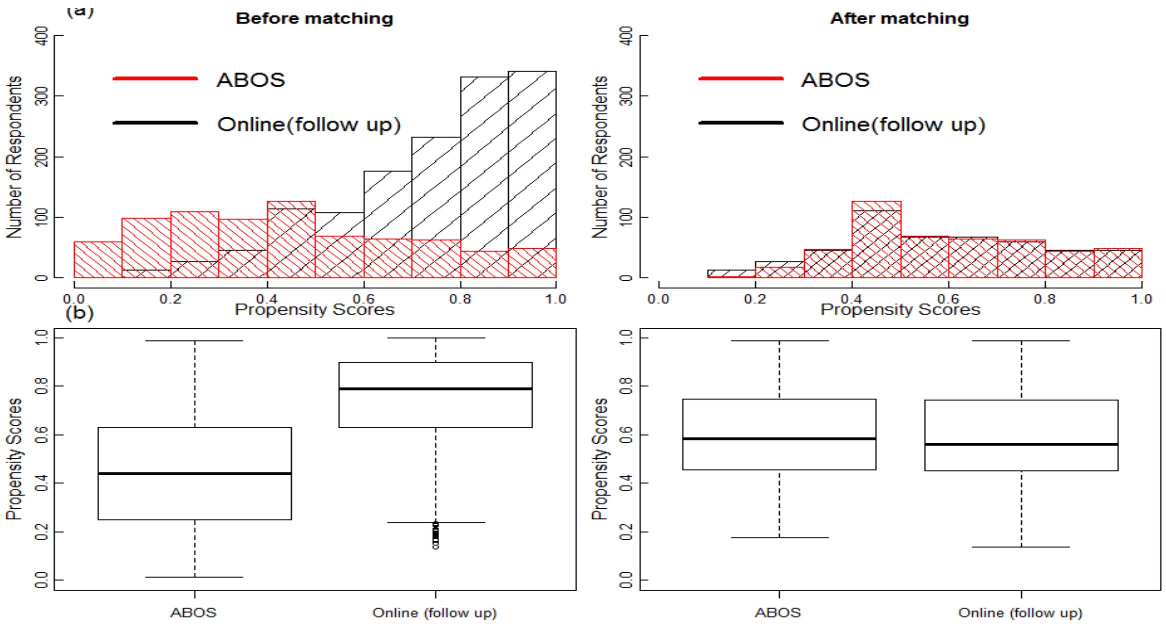
After matching: ABOS = 726 and online (follow up) = 726 respondents

C.4 Estimated Mode Effects Based on Three Different formulations of propensity Score Model



Estimated mode effects based on three different formulations of propensity score models by Question before and after matching for face-to-face and online (follow up) (top left panel), face-to-face and ABOS (top right panel), and ABOS and online (follow up) (bottom left panel)-Greedy Nearest Neighbour Matching

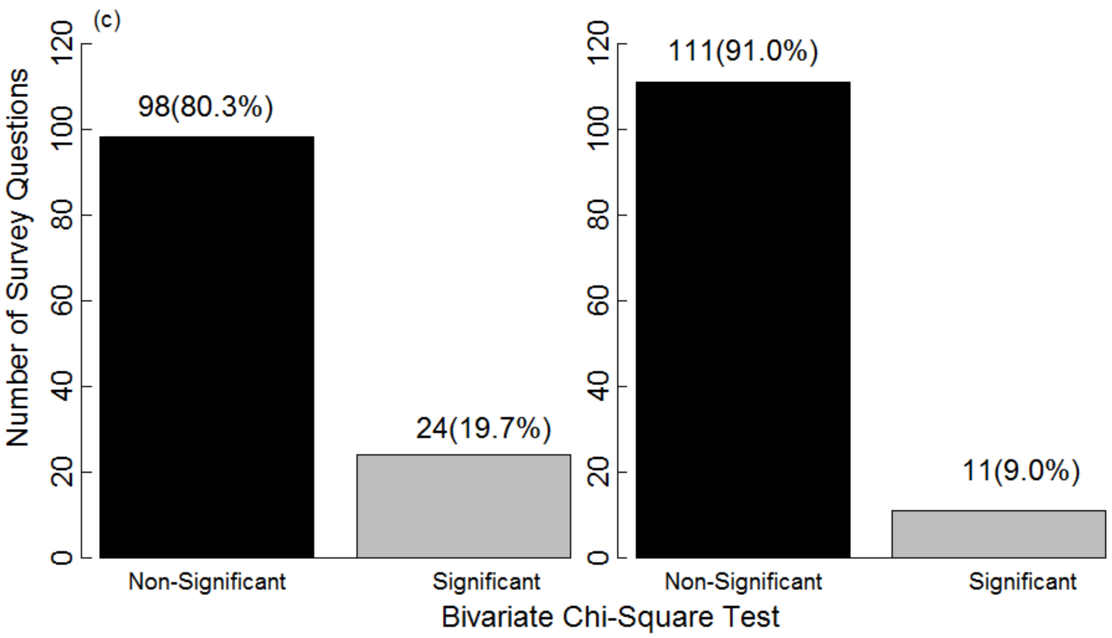
C.5 Propensity Score Model with both Socio-Demographic, Attitudinal and Behavioural Variables: Online (follow up) and ABOS



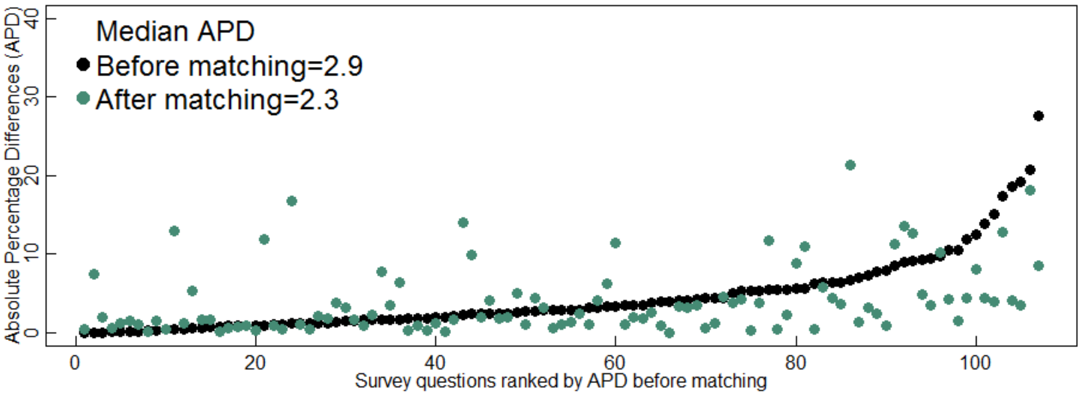
Propensity scores distributions before and after matching presented using histograms (a) and boxplots (b)

Standardised Mean Differences (SMD) for baseline covariates used in propensity score model for Online (follow up) and ABOS samples

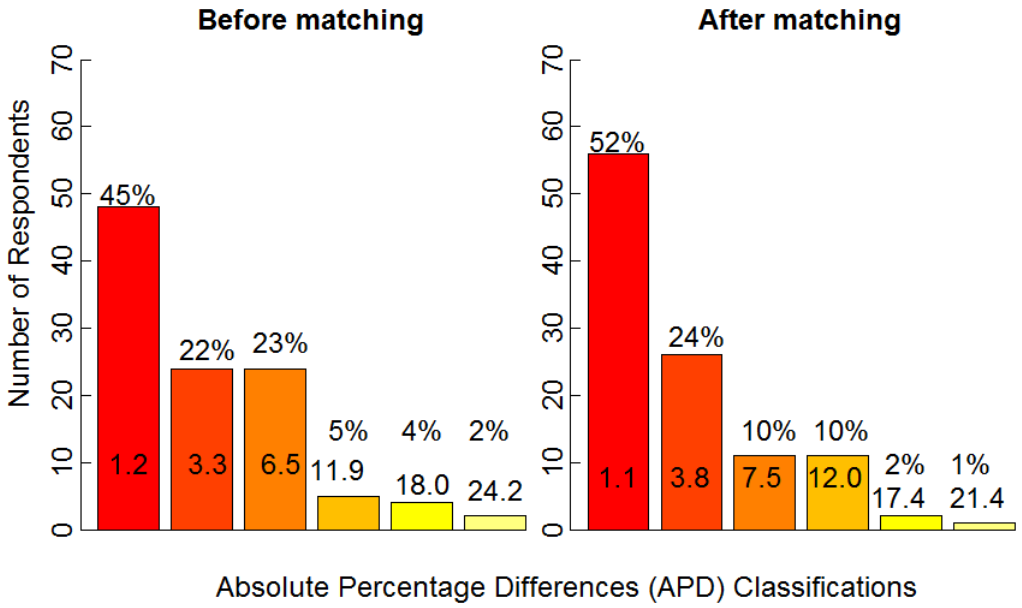
Variable {Ref}	*categories*	*Weighted model*
Propensity scores		0.04
Age	16 to 34 years	0.04
	35 to 49 years	0.06
	50 to 64 years	0.09
	65 to 74 years	0.05
	Over 75 years	0.14
Number of adults in household {1}	2	0.00
	3	0.07
	4 or more	0.01
Belong to neighbourhood		0.05
Satisfied with local area		0.03
Wellbeing		0.04
Bad Health {No}	yes	0.07
lonely{No)	Yes	0.04
Age {16 to 34 years} *Wellbeing	35 to 49 years	0.08
	50 to 64 years	0.02
	65 to 74 years	0.02
	Over 75 years	0.01
Number of adults in household {1} * Satisfied with local area	1	0.01
	2	0.04
	3 or more	0.02
Belong to neighbourhood*satisfied with local area		0.05



Barplots of bivariate chi-square tests of the survey questions before and after matching (weighted model with attitudinal variables) for online (follow up) and ABOS



Estimated mode effects based by Question before and after matching (weighted model with attitudinal variables) for ABOS and online (follow up).



Barplots of Absolute Percentage Differences (APD) classifications with corresponding medians and percentages before and after matching (weighted model with attitudinal variables) for online (follow up) and ABOS

Appendix C

SMD for baseline covariates for ABOS and online (follow up) samples before and after matching (weighted model)

Variable {Ref}	**Categories**	**Before Matching**			**After Matching (weighted model)**		
		ABOS	Online (follow up)	*P-value (SMD)*	ABOS	Online (follow up)	*P-value (SMD)*
		Freq (%)	*Freq (%)*		*Freq (%)*	*Freq (%)*	
Age	16 to 34	175 (22.4)	226 (16.0)	0.004(0.174)	101 (21.0)	88 (18.3)	0.154(0.167)
	35 to 49	203 (26.0)	387 (27.4)		152 (31.6)	133 (27.7)	
	50 to 64	206 (26.4)	415 (29.4)		117 (24.3)	127 (26.4)	
	65 to 74	142 (18.2)	255 (18.1)		74 (15.4)	100 (20.8)	
	75 plus	55 (7.0)	127 (9.0)		37 (7.7)	33 (6.9)	
Gender{Female}	Male	371 (47.5)	615 (43.6)	0.088(0.078)	234 (48.6)	211 (43.9)	0.155(0.096)
Marital Status {Married}	Single	197 (25.2)	308 (21.8)	0.081(0.080)	133 (27.7)	108 (22.5)	0.074(0.120)
Number of Children	0	594 (76.1)	1014 (71.9)	0.003(0.175)	345 (71.7)	355 (73.8)	0.725(0.074)
	1	91 (11.7)	184 (13.0)		63 (13.1)	55 (11.4)	
	2	84 (10.8)	151 (10.7)		64 (13.3)	59 (12.3)	
	3ormore	12 (1.5)	61 (4.3)		9 (1.9)	12 (2.5)	
Paid work {No}	Yes	443 (56.7)	781 (55.4)	0.578(0.027)	288 (59.9)	272 (56.5)	0.327(0.067)
Income	£15<40k	308 (39.4)	529 (37.5)	0.188(0.097)	189 (39.3)	180 (37.4)	0.415(0.109)
	£40k+	83 (10.6)	163 (11.6)		47 (9.8)	62 (12.9)	
	No data	85 (10.9)	122 (8.7)		47 (9.8)	40 (8.3)	
	Under £15k or nothing	305 (39.1)	596 (42.3)		198 (41.2)	199 (41.4)	
Race{Others}	White	712 (91.2)	1297 (92.0)	0.558(0.030)	435 (90.4)	448 (93.1)	0.159(0.099)
Language{Other}	White	758 (97.1)	1358 (96.3)	0.428(0.042)	465 (96.7)	465 (96.7)	1.000(0.001)
Number of adults in household	1	120 (15.4)	349 (24.8)	0.001(0.284)	94 (19.5)	80 (16.6)	0.537(0.095)
	2	461 (59.0)	817 (57.9)		292 (60.7)	293 (60.9)	
	3	110 (14.1)	149 (10.6)		55 (11.4)	66 (13.7)	
	4 or more	90 (11.5)	95 (6.7)		40 (8.3)	42 (8.7)	
Education	no qualifications	199 (25.5)	380 (27.0)	0.211(0.078)	119 (24.7)	121 (25.2)	0.740(0.050)
	Other qualification	341 (43.7)	645 (45.7)		201 (41.8)	210 (43.7)	
	Degree or above	241 (30.9)	385 (27.3)		161 (33.5)	150 (31.2)	
Tenure	Other (mainly private rent)	176 (22.5)	243 (17.2)	0.015(0.143)	107 (22.2)	89 (18.5)	0.442(0.106)
	Mortgaged	238 (30.5)	462 (32.8)		165 (34.3)	162 (33.7)	
	Outright ownership	298 (38.2)	551 (39.1)		166 (34.5)	182 (37.8)	
	Social rent	69 (8.8)	154 (10.9)		43 (8.9)	48 (10.0)	
GOR	London	85 (10.9)	142 (10.1)	0.231(0.146)	55 (11.4)	46 (9.6)	0.334(0.195)
	East Midlands	65 (8.3)	103 (7.3)		39 (8.1)	40 (8.3)	
	East of England	77 (9.9)	165 (11.7)		45 (9.4)	59 (12.3)	
	North East	30 (3.8)	76 (5.4)		22 (4.6)	27 (5.6)	
	North West	128 (16.4)	197 (14.0)		70 (14.6)	67 (13.9)	
	South East	160 (20.5)	266 (18.9)		98 (20.4)	80 (16.6)	
	South West	87 (11.1)	154 (10.9)		55 (11.4)	44 (9.1)	
	West Midlands	67 (8.6)	154 (10.9)		42 (8.7)	58 (12.1)	
	Yorkshire and Humberside	82 (10.5)	153 (10.9)		55 (11.4)	60 (12.5)	
Internet Use {No}	Yes	748 (95.8)	1359 (96.4)	0.552(0.031)	465 (96.7)	469 (97.5)	0.565(0.049)
Rate of Internet Usage	2 to 3 times a week	57 (7.3)	116 (8.2)	0.857(0.039)	34 (7.1)	37 (7.7)	0.899(0.049)
	less than 2-3 times a week, not at all, refused	64 (8.2)	116 (8.2)		34 (7.1)	36 (7.5)	
	more than once a day	555 (71.1)	999 (70.9)		348 (72.3)	350 (72.8)	
	once a day	105 (13.4)	179 (12.7)		65 (13.5)	58 (12.1)	
Meeting family {No}	Yes	204 (26.1)	363 (25.7)	0.888(0.009)	130 (27.0)	117 (24.3)	0.376(0.062)
Belong to neighbourhood		461 (59.0)	927 (65.7)	0.002(0.139)	291 (60.5)	280 (58.2)	0.512(0.047)
Satisfied with local area		0.81 (0.39)	0.86 (0.35)	0.012(0.110)	0.83 (0.37)	0.82 (0.39)	0.611(0.033)
Voting{No}	Yes	517 (66.2)	987 (70.0)	0.074(0.082)	316 (65.7)	341 (70.9)	0.096(0.112)
Volunteer{No}	Yes	586 (75.0)	1062 (75.3)	0.922(0.007)	365 (75.9)	360 (74.8)	0.765(0.024)
Wellbeing		7.08 (2.08)	5.34 (1.76)	0.001(0.907)	6.13 (1.94)	6.22 (1.60)	0.457(0.048)
lonely{No)	Yes	146 (18.7)	278 (19.7)	0.600(0.026)	91 (18.9)	83 (17.3)	0.558(0.043)
Bad Health {No}	Yes	45 (5.8)	82 (5.8)	1.000(0.002)	37 (7.7)	29 (6.0)	0.372(0.066)

Civic participation{No}	Yes	297 (38.0)	577 (40.9)	0.201(0.059)	195 (40.5)	201 (41.8)	0.743(0.025)
Care responsibility{No}	Yes	122 (15.6)	244 (17.3)	0.341(0.045)	79 (16.4)	76 (15.8)	0.861(0.017)
Volunteer{No}	Yes	586 (75.0)	1062 (75.3)	0.922(0.007)	365 (75.9)	360 (74.8)	0.765(0.024)

Before matching: ABOS=779 and Online (follow up) =1,386 respondents
After matching: ABOS=481 and Online (follow up) = 481 respondents

Bibliography

APPOR (2016). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 9th edition*. Retrieved from https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf

Agostino, R. B. D. (1998). Tutorial in Biostatistics: Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Non-randomized Control Group. *Statistics in Medicine*, *17*, 2265–2281.

Agresti, A. (2013). *Categorical Data Analysis* (3rd ed.). New Jersey: John Wiley& Sons.

Alizamini, F. G., Pedram, M. M., Alishahi, M., & Badie, K. (2010). Data quality improvement using fuzzy association rules. *International Conference on Electronics and Information Engineering (ICEIE)*, *1*(Iceie), 468–472. https://doi.org/10.1109/ICEIE.2010.5559676

Allen, D. S. (2016). *BYU ScholarsArchive The Impact of Shortening a Long Survey on Response Rate and Response Quality BYU ScholarsArchive Citation*. Brigham Young University. Retrieved from https://scholarsarchive.byu.edu/etd/5968

Anderson, R. E., Kasper, J., & Frankel, M. R. (1979). *Total Survey Error: Applications to Improve Health Surveys*. San Francisco,CA: Jossey-Bass.

Andrews, I., & Oster, E. (2017). *Weighting for External Validity* (No. 23826). Cambridge, MA. https://doi.org/10.3386/w23826

Armstrong, J. S. (1977). Estimating Nonresponse Bias in Mail Surveys The Wharton School , University of Pennsylvania Terry S . Overton Marketing Scientist , Merck , Sharp and Dohme. *Journal of Marketing*, *14*(3), 396–402. Retrieved from http://cogprints.org/5205/

Aumeyr, M., Brown, Z., Doherty, R., Fallows, A., Pegg, T., Perez-dominguez, R., … Shemwell, L. (2017). *National Survey for Wales 2016-17 Technical Report*. Cardiff. Retrieved from http://www.gov.wales/nationalsurvey

Austin, P. C. (2008a). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, *27*(12), 2037–2049. https://doi.org/10.1002/sim.3150

Austin, P. C. (2008b). Primer on Statistical Interpretation or Methods Report Card on Propensity-Score Matching in the Cardiology Literature From 2004 to 2006. *Circulation:*

Cardiovascular Quality and Outcomes, *1*(1), 62–67. https://doi.org/10.1161/CIRCOUTCOMES.108.790634

Austin, P. C. (2009a). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, *28*, 3083–3107. https://doi.org/10.1002/sim

Austin, P. C. (2009b). Some Methods of Propensity-Score Matching had Superior Performance to Others : Results of an Empirical Investigation and Monte Carlo simulations. *Biometrika*, *51*, 171–184. https://doi.org/10.1002/bimj.200810488

Austin, P. C. (2011a). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, *46*, 399–424. https://doi.org/10.1080/00273171.2011.568786

Austin, P. C. (2011b). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, *10*(2), 150–161. https://doi.org/10.1002/pst.433

Austin, P. C. (2012). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*, *33*(6), 1057–1069. https://doi.org/10.1002/sim.6004

Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects : a Monte Carlo study. *STATISTICS IN MEDICINE*, *26*, 734–753. https://doi.org/10.1002/sim.2580

Austin, P. C., Jembere, N., & Chiu, M. (2018). Propensity score matching and complex surveys. *Statistical Methods in Medical Research*, *27*(4), 1240–1257. https://doi.org/10.1177/0962280216658920

Baghal, T. Al, Jäckle, A., Burton, J., & Lynn, P. (2016). *Understanding Society: Innovation Panel, Waves 1-8, 2008-2015. UK Data Service.* (Vol. SN:7083).

Ballou, J., & DelBoca, F. K. (1980). Gender interaction effects on survey measures in telephone interview. In *Gender interaction effects on survey measures in telephone interviews. Paper presented at the annual meeting of the American Association of Public Opinion Research*. Mason Ohio.

Bayes, M., & Price, M. (1763). An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London*, *53*(0), 370–418. https://doi.org/10.1098/rstl.1763.0053

Beaumont, J. F. (2005). On the use of data collection process information for the treatment of unit nonresponse through weight adjustment. *Survey Methodology*, *31*(2), 227–231.

Beebe, T. J., Mika, T., Harrison, P. A., Anderson, R. E., & Fulkerson, J. A. (1997). Computerized school surveys. *Social Science Computer Review*, *15*(270), 159–169.

Berlin, M., Mohadjer, L., Waksberg, J., Kolstad, A., Kirsch, I., D.Rock, & Yamamoto, K. (1992). An experiment in monetary incentives. In *Proceedings of the Section on Survey Research Methods* (pp. 393–398). Alexandria, VA: American Statistical Association.

Bethlehem, J. (2002). Weighting Nonresponse Adjustments Based on Auxiliary Information. In Groves, Robert M., D. A. Dillman, E. John L., & R. J. A. Little (Eds.), *Survey Nonresponse* (pp. 275–288). New York: Wiley.

Bethlehem, J., & Biffignandi, S. (2011). *Handbook of Web Surveys*. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118121757

Bethlehem, J., Cobben, F., & Schouten, B. (2011). *Handbook of Nonresponse in Household Surveys*. *Handbook of Nonresponse in Household Surveys*. https://doi.org/10.1002/9780470891056

Biemer, P. P. (2010). Total Survey Error: Design, Implementation, and Evaluation. *Public Opinion Quarterly*, *74*(5), 817–848. https://doi.org/10.1093/poq/nfq058

Biemer, P. P. (2016). Total Survey Error Paradigm: Theory and Practice. In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *The SAGE Handbook of Survey Methodology* (pp. 122–141). 1 Oliver’s Yard, 55 City Road London EC1Y 1SP: SAGE Publications Ltd. https://doi.org/10.4135/9781473957893.n10

Biemer, P. P., Chen, P., & Wang, K. (2013). Using level-of-effort paradata in non-response adjustments with application to field surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176*(1), 147–168. https://doi.org/10.1111/j.1467-985X.2012.01058.x

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to Survey Quality*. Hoboken,NJ: John Wiley & Sons.

Bijak, J., & Bryant, J. (2016). Bayesian demography 250 years after Bayes. *Population Studies*, *4728*(August), 1–19. https://doi.org/10.1080/00324728.2015.1122826

Biner, P., & Kidd, H. J. (1994). The interactive effects of monetary incentive justification and questionnaire length on mail survey response rates. *Psychology and Marketing*, *11*(5), 483–492. https://doi.org/10.1002/mar.4220110505

Blangiardo, M., & Cameletti, M. (2015). *Spatial and Spatio-temporal Bayesian Models with R-INLA*. Wiley. https://doi.org/10.1002/9781118950203.ch2

Blasius, J., & Brandt, M. (2010). Representativeness in Online Surveys through Stratified Samples. *Bulletin de Méthodologie Sociologique*, *107*(1), 5–21. https://doi.org/10.1177/0759106310369964

Blau, P. (1964). *Exchange and Power in Social Life*. New York: Wiley.

Blom, A. G. (2009). Nonresponse Bias Adjustments : What Can Process Data Contribute ? *Social Sciences*.

Blom, A. G., de Leeuw, E. D., & Hox, J. J. (2011). Interviewer Effects on Nonresponse in the European Social Survey. *Journal of Official Statistics*, *27*(2), 359–377.

Blom, A. G., Gathmann, C., & Krieger, U. (2015). Setting Up an Online Panel Representative of the General Population: The German Internet Panel. *Field Methods*, *27*(4), 391–408. https://doi.org/10.1177/1525822X15574494

Blom, A. G., Jäckle, A., & Lynn, P. (2010). The Use of Contact Data in Understanding Cross-National Differences in Unit Nonresponse. *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, 333–354. https://doi.org/10.1002/9780470609927.ch18

Blom, A. G., & Korbmacher, J. M. (2013). Measuring Interviewer Characteristics Pertinent to Social Surveys: A Conceptual Framework. *Survey Methods: Insights from the Field*, 1–16. https://doi.org/10.13094/SMIF-2013-00001

Boreham, A. R., & Constantine, R. (2008). Understanding Society Innovation Panel Wave 1 Technical Report, (7083), 1–14.

Boyd Jr., H. W., & Westfall, R. L. (1955). Interviewers As a Source of Error in Surveys. *Journal of Marketing*, *19*(4), 311–324. https://doi.org/10.2307/1247046

Brick, M. J., & Montaquila, J. M. (2009). Nonresponse and Weighting (pp. 163–185). https://doi.org/10.1016/S0169-7161(08)00008-4

Brick, M. J., Montaquila, J. M., Hagedorn, M. C., Roth, S. B., & Chapman, C. (2005). Implications for RDD design from an incentive experiment. *Journal of Official Statistics*, *21*(4), 571–589.

Brick, M. J., & Williams, D. K. (2013). Explaining Rising Nonresponse Rates in Cross-Sectional Surveys. *The ANNALS of the American Academy of Political and Social Science*, *645*(1), 36–59. https://doi.org/10.1177/0002716212456834

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2007). Variable Selection for propensity score models. *American Journal of Epidemiology*, *163*(12), 1149–1156.

Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. CRC press. https://doi.org/10.1201/b10905

Browne, W. J., Draper, D., & David, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, *1*(1), 473–550.

Browne, W. J., Kelly, M., Charlton, C. M. J., & Pillinger, R. (2016). *MCMC estimation in MLwiN: Version 2.36*.

Bryson, A., Dorsett, R., & Purdon, S. (2002). The use of propensity score matching in the evaluation of active labour market policies. *Policy Studies Institute and National Centre for Social Research*, (4), 57.

Buck, N., & McFall, S. (2012). Understanding Society: design overview. *Longitudinal and Life Course Studies*, *3*(1), 5–17. https://doi.org/10.14301/llcs.v3i1.159

Burkill, S., Copas, A., Couper, M. P., Clifton, S., Prah, P., Datta, J., … Erens, B. (2016). Using the web to collect data on sensitive behaviours: A study looking at mode effects on the British national survey of sexual attitudes and lifestyles. *PLoS ONE*, *11*(2), 1–12. https://doi.org/10.1371/journal.pone.0147983

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*(1), 31–72. https://doi.org/10.1111/j.1467-6419.2007.00527.x

Callegaro, M. (2013). Paradata in Web Surveys. In *Improving Surveys with Paradata* (pp. 259–279). Hoboken, New Jersey: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118596869.ch11

Callegaro, M., Lozar Manfreda, K., & Vehovar, V. (2015). *Web Survey Methodology (Research Methods for Social Scientists)*. Thousand Oaks,CA: Sage.

Campanelli, P. (1997). Testing Survey Questions: New Directions in Cognitive Interviewing. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *55*(1), 5–17. https://doi.org/10.1177/075910639705500103

Campanelli, P., & O'Muircheartaigh, C. (1999). Interviewers , Interviewer Continuity , and Panel Survey Nonresponse 1. *Quality & Quantity*, *33*, 59–76.

Campanelli, P., Sturgis, P., & Purdon, S. (1997). Can you hear me knocking? and investigation into the impact of interviewers on survey response rates. *National Centre for Social Research*.

Cannell, C. F., Miller, P. V., & Oksenberg, L. (1981). Research on Interviewing Techniques. *Sociological Methodology*, *12*, 389. https://doi.org/10.2307/270748

Cantor, D., O'Hare, B. C., & O'Connor, K. S. (2008). The Use of Monetary Incentives to Reduce Nonresponse in Random Digit Dial Telephone Surveys. In J. M. Lepkowski, C. Tucker, J. M. Brick, E. D. de Leeuw, L. Jape, P. J. Lavrakas, … R. L. Sangster (Eds.), *Advances in Telephone Survey Methodology* (pp. 471–498). Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9780470173404.ch22

Carlson, B. L., & Williams, S. (2001). A comparison of two methods to adjust weights for non-response: propensity modeling and weighting class adjustments. *Proceedings of the Annual Meeting of the American Statistical Association, August 5-9, 2001*.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2016). Stan : A Probabilistic Programming Language. *Journal of Statistical Software*, *VV*(Ii).

Charlton, C. M. J., Michaelides, D. T., Parker, R. M. A., Cameron, B., Szmaragd, C., Yang, H., … W.J., B. (2013). Stat-JR version 1.0. *Centre for Multilevel Modelling, University of Bristol & Electronics and Computer Science, University of Southampton.*

Church, A. H. (1993). Estimating the Effect of Incentives on Mail Survey Response Rates : A Meta-Analysis. *Public Opinion Quarterly*, *57*(1), 62–79.

Cialdini, R. B. (1984). *Influence: The New Psychology of Modern Persuasion*. New York: Quill.

Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). New York: wil.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New Jersey: Hillsdale. https://doi.org/10.1016/C2013-0-10517-X

Converse, J., & Presser, S. (1986). *Survey Questions: Handcrafting the Standardized Questionnaire (Quantitative Applications in the Social Sciences)*. Thousand Oaks, CA US: Sage Publications.

Couper, M. P. (1998). Measuring Survey Quality in a CASIC Environment. Retrieved from http://www.amstat.org/sections/srms/proceedings/papers/1998_006.pdf

Couper, M. P. (2000). *Handbook of Web Surveys*. *Public Opinion Quarterly* (Vol. 64). https://doi.org/10.1086/318641

Couper, M. P. (2011). The future of modes of data collection. *Public Opinion Quarterly*, *75*(5 SPEC. ISSUE), 889–908. https://doi.org/10.1093/poq/nfr046

Couper, M. P., & Groves, R. M. (1996). Social environmental impacts on survey cooperation. *Quality and Quantity*, *30*(2), 173–188. https://doi.org/10.1007/BF00153986

Couper, M. P., & Schlegel, J. (1998). Evaluating the NHIS CAPI instrument using trace files. In *In Annual Meeting of the American Association for Public Opinion Research*. St. Louis. Retrieved from https://ww2.amstat.org/sections/srms/Proceedings/papers/1997_142.pdf

Cuong, N. V. (2013). Which covariates should be controlled in propensity score matching ? Evidence from a simulation study. *Statistica Neerlandica*, *67*(2), 169–180. https://doi.org/10.1111/stan.12000

Currivan, D. (2005). The Impact of Providing Incentives to Initial Telephone Survey Refusers on Sample Composition and Data Quality.

Dalenius, T. (1974). *Ends and Means of Total Survey Design. Report from Research project Errors in Surveys*.

Damlen, P., Wakefield, J., & Walker, S. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(2), 331–344. https://doi.org/10.1111/1467-9868.00179

David, M., Little, R. J. A., Samuhel, M. E., & Triest, R. K. (1983). Nonrandom nonresponse models based on the propensity to respond. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 168–173.

de Bruijne, M., & Wijnant, A. (2013). Comparing Survey Results Obtained via Mobile Devices and Computers: An Experiment With a Mobile Web Survey on a Heterogeneous Group of Mobile Devices Versus a Computer-Assisted Web Survey. *Social Science Computer Review*, *31*(4), 482–504. https://doi.org/10.1177/0894439313483976

de Leeuw, E. D. (1992). *Data Quality in Mail, Telephone and Face to Face surveys*. Vrije Universiteit te Amsterdam.

de Leeuw, E. D. (2005). To Mix or Not to Mix Data Collection Modes in Surveys. *Journal of Official Statistics*, *21*(2), 233–255. https://doi.org/10.4324/9780203843123

de Leeuw, E. D. (2009). Types of Mixed Mode Designs. Retrieved from https://edithl.home.xs4all.nl/Past/MiOA09.pdf

de Leeuw, E. D. (2018). Mixed-Mode: Past , Present , and Future. *Survey Research Methods*, *12*(2), 75–89.

de Leeuw, E. D., & de Heer, W. (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.) (p. 41–54 in Survey Nonresponse). New York: Wiley.

de Leeuw, E. D., & Hox, J. J. (2011). Internet surveys as part of a mixed-mode design. (M. Das, P. Ester, & L. Kaczmirek, Eds.), *Social and Behavioral Research and the Internet*. New York.

de Leeuw, E. D., Hox, J. J., & Huisman, M. (2003). Prevention and treatment of item nonresponse.pdf. *Journal of Official Statistics*, *19*(2), 153–176.

de Leeuw, E. D., Hox, J. J., & Kef, S. (2003). Computer-Assisted Self-Interviewing Tailored for Special Populations and Topics. *Field Methods*, *15*(3), 223–251. https://doi.org/10.1177/1525822X03254714

de Leeuw, E. D., Hox, J. J., Snijkers, G., & de Heer, W. (1998). Interviewer opinions, attitudes and strategies regarding survey participation and their effect on response. In A. Koch & R. Porst (Eds.), *Nonresponse in Survey Research.* (Zuma Nachr, pp. 239–248). Mannheim: FRG: ZUMA.

Dehejia, R. H., & Wahba, S. (2002). Propensity Score-Matchin Mathods for Nonexperimental Causal Studies. *The Review of Economics and Statistics*, *84 (1)*(1), 151–161. https://doi.org/10.1162/003465302317331982

DeMaio, T. J. (1980). Refusals: Who, Where and Why. *Public Opinion Quarterly*, *44*(2), 223. https://doi.org/10.1086/268586

Deming, E. (1944). On Errors in Surveys. *American Sociological Review*, *9*(359–369).

Diamond, A., & Sekhon, J. S. (2013). Genetic Matching for Estimating Causal Effects. *Review of Economics and Statistics*, *95*(3), 932–945. https://doi.org/10.1162/REST_a_00318

Diehr, P., Chen, L., Patrick, D., Feng, Z., & Yasui, Y. (2005). Reliability , effect size , and responsiveness of health status measures in the design of randomized and clusterrandomized trials, *26*, 45–58. https://doi.org/10.1016/j.cct.2004.11.014

Dijkstra, W., & Smit, J. H. (2002). Persuading reluctant recipients in telephone surveys. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey Nonresponse* (pp. 121–134). New York: Wiley.

Dillman, D. A. (2002). *Mail and Internet Surveys: The Tailored Design Method* (2nd ed.). New York: Wiley.

Dillman, D. A. (2007). *Mail and Internet Surveys*.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method* (3rd ed.). New York, NY: John Wiley & Sons.

Dillman, D. A., & Tarnai, J. (1989). Administrative issues in mixed mode surveys. In R. M. Groves, P. P. Biemer, L. E. Lyberg, L. Massry, I. Nicholls, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 509–528). New York.

Dixon, J. (2002). The Effects of Item and Unit Nonresponse on Estimates of Labor Force Participation. *Joint Statistical Meetings, NY, NY*, 803–806. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:The+Effects+of+Item+and+Unit+Nonresponse+on+Estimates+of+Labor+Force+Participation.#2

Drago, E. (2015). The Effect of Technology on Face-to-Face Communication. *Elon Journal of Undergraduate Research in Communications*, *6*(1), 13–19.

Duan, Y. (2005). *A Modified Bayesian Power Prior Approach with Applications in Water Quality Evaluation*. Faculty of Virginia Polytechnic Institute and State University. Retrieved from https://vtechworks.lib.vt.edu/bitstream/handle/10919/29976/dissertation.pdf?sequence=1&isAllowed=y

DuGoff, E. H., Schuler, M., & Stuart, E. A. (2014). Generalizing Observational Study Results: Applying Propensity Score Methods to Complex Surveys. *Health Services Research*, *49*(1), 284–303. https://doi.org/10.1111/1475-6773.12090

Dunn, E. C., Richmond, T. K., Milliren, C. E., & Subramanian, S. V. (2015). Health & Place Using cross-classified multilevel models to disentangle school and neighborhood effects : An example focusing on smoking behaviors among adolescents in the United States. *Health & Place*, *31*, 224–232.

Durrant, G. B., D'Arrigo, J., & Müller, G. (2013). Modeling Call Record Data: Examples from Cross-Sectional and Longitudinal Surveys, 281–308.

Durrant, G. B., D'Arrigo, J., & Steele, F. (2011). Using paradata to predict best times of contact, conditioning on household and interviewer influences. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *174*(4), 1029–1049. https://doi.org/10.1111/j.1467-985X.2011.00715.x

Durrant, G. B., D'Arrigo, J., & Steele, F. (2013). Analysing Interviewer Call Record Data by Using a Multilevel Discrete-Time Event History Modelling Approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176*(1), 251–269.

Durrant, G. B., Groves, R. M., Steele, F., Staetsky, L., Blom, A. G., de Leeuw, E. D., & Hox, J. J. (2010). Effects of Interviewer Attitudes and Behaviors on Refusal in Household Surveys. *Public Opinion Quarterly*, *74*(1), 1–36. https://doi.org/10.1093/poq/nfp098

Durrant, G. B., & J D'Arrigo, J. (2014). Doorstep Interactions and Interviewer Effects on the Process Leading to Cooperation or Refusal. *Sociological Methods & Research*, *43*(3), 490–518. https://doi.org/10.1177/0049124114521148

Durrant, G. B., & Kreuter, F. (2013). Editorial: The use of paradata in social survey research. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176*(1), 1–3. https://doi.org/10.1111/j.1467-985X.2012.01082.x

Durrant, G. B., Maslovskaya, O., & Smith, P. W. F. (2015). Modeling Final Outcome and Length of Call Sequence to Improve Efficiency in Interviewer Call Scheduling. *Journal of Survey Statistics and Methodology*, *3*(3), 397–424. https://doi.org/10.1093/jssam/smv008

Durrant, G. B., Maslovskaya, O., & Smith, P. W. F. (2017). Using Prior Wave Information and Paradata: Can They Help to Predict Response Outcomes and Call Sequence Length in a Longitudinal Study? *Journal of Official Statistics*, *33*(3), 801–833. https://doi.org/10.1515/jos-2017-0037

Durrant, G. B., & Steele, F. (2007). Multilevel modelling of refusal and noncontact nonresponse in household surveys: evidence from six UK government surveys, 1–36. Retrieved from http://eprints.soton.ac.uk/46013/

Durrant, G. B., & Steele, F. (2009). Multilevel modelling of refusal and non-contact in household surveys: Evidence from six UK Government surveys. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *172*(2), 361–381. https://doi.org/10.1111/j.1467-985X.2008.00565.x

Dykema, J., Lepkowski, J. M., & Blixt, S. (2012). The Effect of Interviewer and Respondent Behavior on Data Quality: Analysis of Interaction Coding in a Validation Study. In L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey Measurement and Process Quality* (pp. 287–310). Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118490013.ch12

Edwards, B., Maitland, A., & O'Connor, K. S. (2017). Measurement error in survey operations management: detection, quantification, visualization, and reduction. In P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. Lyberg, … B. T. West (Eds.), *Total Survey Error in Practice*. Hoboken,NJ: John Wiley & Sons.

Ellen Hansen, S., Benson, G., Bowers, A., Pennell, B.-E., Lin, Y., Duffey, B., … Cibelli Hibben, K. (2016). *Cross-Cultural Survey Guidelines Survey Quality*. Retrieved from https://ccsg.isr.umich.edu/images/PDFs/CCSG_Survey_Quality.pdf

Elze, M. C., Gregson, J., Baber, U., Williamson, E., Sartori, S., Mehran, R., … Pocock, S. J. (2017). Comparison of Propensity Score Methods and Covariate Adjustment: Evaluation in 4 Cardiovascular Studies. *Journal of the American College of Cardiology*, *69*(3), 345–357. https://doi.org/10.1016/j.jacc.2016.10.060

Ertefaie, A., & Stephens, D. A. (2010). Comparing Approaches to Causal Inference for Longitudinal Data: Inverse Probability Weighting versus Propensity Scores. *The International Journal of Biostatistics*, *6*(2). https://doi.org/10.2202/1557-4679.1198

Evans, M., Jang, G. H., & Jan, M. E. (2011). Weak Informativity and the Information in One Prior Relative to Another, *26*(3), 423–439. https://doi.org/10.1214/11-STS357

Fan, W., & Yan, Z. (2010). Factors affecting response rates of the web survey : A systematic review. *Computers in Human Behavior*, *26*(2), 132–139. https://doi.org/10.1016/j.chb.2009.10.015

Fearn, T., Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1996). Bayesian Data Analysis. *Biometrics*, *52*(3), 1160. https://doi.org/10.2307/2533081

Fearn, T., Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B., & Fearn, T. (2004). Bayesian Data Analysis. *Biometrics*, *52*(3), 696. https://doi.org/10.1007/s13398-014-0173-7.2

Ferkingstad, E., & Rue, H. (2015). Improving the INLA approach for approximate Bayesian inference for latent Gaussian models. *Electronic Journal of Statistics*, *9*(2), 2706–2731. https://doi.org/10.1214/15-EJS1092

Fong, Y., Rue, H., & Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics*, *11*(3), 397–412. https://doi.org/10.1093/biostatistics/kxp053

Fowler, F. J., & Mangione, T. W. (1990). *Standardized Survey Interviewing: Minimizing Interviewer- Related Error*. Newbury Park, Calif: SAGE Publications, Inc.

Freese, J., & Long, J. S. (2006). *Regression models for categorical dependent variables using stata*. *Stata Pres, College Station* (3rd ed.). Texas: STATA Press.

Fricker, S., & Tourangeau, R. (2010). Examining the relationship between nonresponse propensity and data quality in two national household surveys. *Public Opinion Quarterly*, *74*(5), 934–955. https://doi.org/10.1093/poq/nfq064

Gelman, A. (2002). Prior distribution. *Encyclopedia of Environmetrics*, *3*, 1634–1637. https://doi.org/10.1002/9780470057339.vap039

Gelman, A. (2006). Prior distribution for variance parameters in hierarchical models. *Bayesian Analysis*, *1*(3), 515–533. https://doi.org/10.1214/06-BA117A

Gelman, A., & Hill, J. L. (2007). Data analysis using regression and multilevel/hierarchical models. *Policy Analysis*, 1–651. https://doi.org/10.2277/0521867061

Gelman, A., Hwang, J., & Vehtari, A. (2013). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 1–20. https://doi.org/10.1007/s11222-013-9416-2

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, *2*(4), 1360–1383. https://doi.org/10.1214/08-AOAS191

Gelman, A., Stevens, M., & Chan, V. (2002). Regression modeling and meta-analysis for decision making: A cost-benefit analysis of incentives in telephone surveys. *Journal of Business & Economic Statistics*, *21*(2), 213–225.

Ghosh, M. (2011). Objective Priors: An Introduction for Frequentists. *Statistical Science*, *26*(2), 187–202. https://doi.org/10.1214/10-STS338

Gill, J. (2014). *Bayesian methods: A social and behavioral sciences approach* (20th ed.). CRC press.

Gill, J., & Walker, L. (2005). Elicited Priors for Bayesian Model Specification in Political Science Research. *Journal of Politics*, *67*(3), 841–872. https://doi.org/10.1111/j.1468-2508.2005.00342.x

Gjonça, E., & Calderwood, L. (2004). 2. Socio-demographic characteristics. Retrieved from http://www.elsa-project.ac.uk/uploads/elsa/report03/ch2.pdf

Goldberg, M., Chastang, J. F., Leclerc, A., Zins, M., Bonenfant, S., Kaniewski, N., … Imbernon, E. (2001). Socioeconomic , Demographic , Occupational , and Health Factors Associated with Participation in a Long-term Epidemiologic Survey : A Prospective Study of the French GAZEL Cohort and Its Target Population. *American Journal of Epidemiology*, *154*(4).

Goldenberg, K. L., Statistics, L., Levin, K., Hagerty, T., Shen, T., Cantor, D., … Dc, W. (1997). Procedures for Reducing Measurement Error in Establishment Surveys. In *American Association for Public Opinion Research* (pp. 1–7). Norfolk, Virginia. Retrieved from https://www.bls.gov/osmr/pdf/st970090.pdf

Goldstein, H. (2011). *Multilevel statistical models* (Vol. 922). John Wiley & Sons.

Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning Variation in Multilevel Models. *Understanding Statistics*, *1*(4), 223–231. https://doi.org/10.1207/S15328031US0104_02

Goyder, J. (1987). *The Silent Minority: Nonrespondents on Sample Surveys*. Cambridge: Polity Press.

Goyder, J., Lock, J., & McNair, T. (1992). Urbanization effects on survey nonresponse: a test within and across cities. *Quality and Quantity*, *26*(1). https://doi.org/10.1007/BF00177996

Grendár, M. (2012). Is the p-value a good measure of evidence? Asymptotic consistency criteria. *Statistics and Probability Letters*, *82*(6), 1116–1119. https://doi.org/10.1016/j.spl.2012.02.018

Grilli, L., Metelli, S., & Rampichini, C. (2014). Bayesian estimation with integrated nested Laplace approximation for binary logit mixed models. *Journal of Statistical Computation and Simulation*, *85*(13), 2718–2726. https://doi.org/10.1080/00949655.2014.935377

Groves, R. M. (1989). *Survey Errors and Survey Costs*. New York: John Wiley & Sons, Inc.

Groves, R. M. (2006). Nonresponse rates and nonrespons bias in household surveys. *Public Opinion Quarterly*, *70*(5), 646–675. https://doi.org/10.1093/poq/nfl033

Groves, R. M., Cialdini, R. B., & Couper, M. P. (1992). UNDERSTANDING THE DECISION TO PARTICIPATE IN A SURVEY, *56*, 475–495.

Groves, R. M., & Couper, M. P. (1996). contact level influences on cooperation in Face to Face surveys.pdf. *Journal of Official Statistics*, (12), 63–83.

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in Household Interview Surveys*. New York: John Wiley and Sons.

Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. A. (2002). *Survey Nonresponse*. New York: Wiley.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011). *Survey Methodology* (2nd ed.). Hoboken,NJ: John Wiley & Sons.

Groves, R. M., & Heeringa, S. G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *169*(3), 439–457. https://doi.org/10.1111/j.1467-985X.2006.00423.x

Groves, R. M., Jr Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology* (2nd ed.). Hoboken, New Jersey: Wiley.

Groves, R. M., & Lyberg, L. E. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, *74*(5), 849–879. https://doi.org/10.1093/poq/nfq065

Groves, R. M., & Mcgonagle, K. (2001). A Theory-Guided Interviewer Training Protocol Regarding Survey Participation. *Journal of Official Statistics*, *17*, 249–265.

Groves, R. M., O'Hare, B. C., Gould-Smith, D., Benkí, J. R., & Maher, P. (2007). Telephone Interviewer Voice Characteristics and the Survey Participation Decision. In J. M. Lepkowski, N. C. Tucker, J. M. Brick, E. de Leeuw, L. Japec, P. J. Lavrakas, … R. L. Sangster (Eds.), *Advances in telephone survey methodologyTelephone Survey Methodology* (pp. 385–400). John Wiley & Sons. https://doi.org/10.1002/9780470173404.ch18

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly*, *72*(2), 167–189. https://doi.org/10.1093/poq/nfn011

Groves, R. M., Singer, E., & Corning, A. D. (2000). Leverage-Saliency Theory of Survey Participation Description and an Illustration. *The Public Opinion Quarterly*, *64*(3), 299–308.

Guikema, S. D. (2007). Formulating informative, data-based priors for failure probability estimation in reliability analysis. *Reliability Engineering and System Safety*, *92*(4), 490–502. https://doi.org/10.1016/j.ress.2006.01.002

Guikema, S. D., & Pate-Cornell, M. E. (2004). Bayesian Analysis of Launch Vehicle Success Rates. *J Spacecraft Rockets*, *41*(1), 93–102. https://doi.org/10.2514/1.9268

Guo, S., Barth, R. P., & Gibbons, C. (2006). Propensity score matching strategies for evaluating substance abuse services for child welfare clients. *Children and Youth Services Review*, *28*(4), 357–383. https://doi.org/10.1016/j.childyouth.2005.04.012

Guo, S., & Fraser, M. W. (2014). *Propensity Score Analysis: Statistical Methods and Applications (Advanced Quantitative Techniques in the Social Sciences)* (2nd Editio). SAGE Publications.

Hahn, J. (1998). On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica*, *66*(2), 315. https://doi.org/10.2307/2998560

Hansen, B. B. (2008). The essential role of balance tests in propensity-matched observational studies: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin,Statistics in Medicine. *Statistics in Medicine*, *27*(12), 2050–2054. https://doi.org/10.1002/sim.3208

Hansen, K. M. (2006). The Effects of Incentives , Interview Length , and Interviewer Characteristics on Response Rates in a CATI-Study. *International Journal of Public Opinion Research*, (19), 112–121. https://doi.org/10.1093/ijpor/edl022

Hansen, M. H., Hurwitz, W. N., & Bershad, M. A. (1961). MEASUREMENT ERRORS IN CENSUSES AND SURVEYS. *Bulletin of the International Statistical Institute*, *38*, 359–374.

Hanson, T. E., Sullivan, S., & Mcgowan, C. (2015). *National Survey for Wales 2014-15 Technical Report*. Cardiff.

Hanson, T. E., Sullivan, S., & Mcgowan, C. (2016). *National Survey for Wales Field Test Technical Report*. Cardiff. Retrieved from http://gov.wales/docs/caecd/research/2016/160315-national-survey-field-test-technical-report-en.pdf

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Haunberger, S. (2010). The effects of interviewer , respondent and area characteristics on cooperation in panel surveys : a multilevel approach. *Quality and Quantity*, *44*(5), 957–969. https://doi.org/10.1007/s11135-009-9248-5

Heerwegh, D. (2009). Mode Differences Between Face-to-Face and Web Surveys: An Experimental Investigation of Data Quality and Social Desirability Effects. *International Journal of Public Opinion Research*, *21*(1), 111–121.

Heerwegh, D., & Loosveldt, G. (2008). Face-to-face versus web surveying in a high-internet-coverage population: Differences in response quality. *Public Opinion Quarterly*, *72*(5), 836–846. https://doi.org/10.1093/poq/nfn045

Heinze, G., & Jüni, P. (2011). An overview of the objectives of and the approaches to propensity score analyses. *European Heart Journal*, *32*(14), 1704–1708. https://doi.org/10.1093/eurheartj/ehr031

Held, L., Schrodle, B., & Rue, H. (2010). Posterior and cross-validatory predictive checks: A comparison of MCMC and INLA. *Statistical Modelling and Regression Structures: Festschrift in Honour of Ludwig Fahrmeir*, (91), 91–110. https://doi.org/10.1007/978-3-7908-2413-1_6

Hernan, M. (2004). A definition of causal effect for epidemiological research. *Journal of Epidemiology and Community Health*, *58*(4), 265–271. https://doi.org/10.1136/jech.2002.006361

Heyneman, S. P. (2006). Global issues in higher education. *EJournal USA*, *11*(1), 52–55.

Hirano, K., & Imbens, G. W. (2001). Estimation of Causal Effects using Propensity Score Weighting : An Application to Data on Right Heart Catheterization. *Health Services & Outcomes Research Methodology*, 259–278.

Hirano, K., Imbens, G. W., & Geert, R. (2003). Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica*, *71*(4), 1161–1189. https://doi.org/10.1111/1468-0262.00442

Ho, D., Imai, K., King, G., & Stuart, E. A. (2009). Package 'MatchIt': Nonparametric Preprocessing for Parametric Casual Inference. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.409.3968&rep=rep1&type=pdf

Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone versus Face-to-Face Interviewing of National Probability Samples with Long Questionnaires. *Public Opinion Quarterly*, *67*(1), 79–125. https://doi.org/10.1086/346010

Holbrook, A. L., Krosnick, J. A., & Pfent, A. (2007). The Causes and Consequences of Response Rates in Surveys by the News Media and Government Contractor Survey Research Firms 1. *Advances in Telephone Survey Methodology*, *60607*, 499–458.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (2nd ed.). New York: Wiley.

Hox, J. J., & de Leeuw, E. D. (2002). The influence of interviwers' attitude and behavior on household survey nonresponse: An international comparison. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 103–120). New York: Wiley.

Hox, J. J., de Leeuw, E. D., & Klausch, T. (2017). Mixed-Mode Research: Issues in Design and Analysis. In P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. Lyberg, … B. T. West (Eds.), *Total Survey Error in Practice* (pp. 511–530). Hoboken,NJ.

Hu, B., Shao, J., & Palta, M. (2006). Pseudo-R2 in Logistic Regression Model. *Statistica Sinica*, *16*, 847–860. Retrieved from http://www3.stat.sinica.edu.tw/statistica/oldpdf/a16n39.pdf

Huddy, L., Billig, J., Bracciodieta, J., Moynihan, P. J., & Pugliani, P. (1997). The Effect of Interviewer Gender on the Survey Response. *Political Behavior*, *19*(3), 197–220.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, *86*, 4–29.

Jäckle, A., Lynn, P., & Burton, J. (2015). Going Online with a Face-to-Face Household Panel: Effects of a Mixed Mode Design on Item and Unit Non-Response. *Survey Research Methods*, *9*(1), 57–70. https://doi.org/10.18148/srm/2015.v9i1.5475

Jäckle, A., Lynn, P., Sinibaldi, J., & Tipping, S. (2011). The effect of interviewer personality , skills and attitudes on respondent co-operation with face-to-face surveys.

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

Jäckle, A., Roberts, C., Lynn, P., Robert, C. P., & Lynn, P. (2010). Assessing the Effect of Data Collection Mode on Measurement. *International Statistical Review*, *78*(1), 3–20. https://doi.org/10.1111/j.1751-5823.2010.00102.x

Jannink, J.-L. (2003). *Likelihood of Bayesian, and MCMC Methods in Quantitative Genetics. Crop Science* (Vol. 43). https://doi.org/10.2135/cropsci2003.1574

Janson, C. G. (2003). Factorial Social Ecology: An Attempt at Summary and Evaluation. *Annual Review of Sociology*, *6*(1), 433–456. https://doi.org/10.1146/annurev.so.06.080180.002245

Jobe, J. B., & Mingay, D. J. (1991). Cognition and survey measurement: History and overview. *Applied Cognitive Psychology*, *5*(3), 175–192. https://doi.org/10.1002/acp.2350050303

Joe, H. (2008). Accuracy of Laplace approximation for discrete response mixed models.

Computational Statistics and Data Analysis, *52*, 5066–5074. https://doi.org/10.1016/j.csda.2008.05.002

Kalton, G., & Flores-cervantes, I. (2003). Weighting Methods, *19*(2), 81–97.

Kaminska, O., & Foulsham, T. (2013). Understanding Sources of Social Desirability Bias in Different Modes : Evidence from Eye-tracking. *Institute for Social & Economic Research*, 1–11.

Kazmi, W. H., Obrador, G. T., Khan, S. S., Pereira, B. J. G., & Kausz, A. T. (2004). Late nephrology referral and mortality among patients with end-stage renal disease: A propensity score analysis. *Nephrology Dialysis Transplantation*, *19*(7), 1808–1814. https://doi.org/10.1093/ndt/gfg573

Kim, J. K., & Kim, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canadian Journal of Statistics*, *35*(4), 501–514. https://doi.org/10.1002/cjs.5550350403

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

Kish, L. (1965). *Survey Sampling*. New York: Wiley.

Klausch, T., & Schouten, B. (2015). Selection Error in Single- and Mixed-Mode Surveys of the Dutch General Population. *Journal of the Royal Statistical Society: Series A*, *178*(4), 945–961.

Knies, G. (2014). Understanding Society-The UK Household Longitudinal Study Waves 1-5 User Manual. *Innovation*, (October), 1–6. https://doi.org/10.2307/3348243

Kreuter, F. (2013). *Improving surveys with paradata: Introduction." Improving Surveys with Paradata*. https://doi.org/10.1017/CBO9781107415324.004

Kreuter, F., Couper, M. P., & Lyberg, L. E. (2010). The use of paradata to monitor and manage survey data collection. *Measurement*, 282–296.

Kreuter, F., & Kohler, U. (2009). Analysing contact sequences in call records data.Potenetial and limitations of sequence indicators for nonresponse adjustments in the European Social Survey. *Journal of Official Statistics*, *25*(2), 200–226.

Kreuter, F., & Olson, K. (2011). Multiple Auxiliary Variables in Nonresponse Adjustment. *Sociological Methods & Research*, *40*(2), 311–332. https://doi.org/10.1177/0049124111400042

Kreuter, F., & Olson, K. (2013). Paradata for Nonresponse Error Investigation Paradata for Nonresponse Error Investigation. *Improving Surveys with Paradata: Analytic Uses of Process Information*, 13–42.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T. M., Casas-cordero, C., … Raghunathan, T. (2010). Using proxy measures and other correlates of survey outcomes to adjust for non-response: examples from multiple surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *173*(2), 389–407. https://doi.org/10.1111/j.1467-985X.2009.00621.x

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, *72*(5), 847–865. https://doi.org/10.1093/poq/nfn063

Krosnick, J. A. (1991). Response stratergies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213–236.

Krosnick, J. A. (1999). Survey Research. *Annu. Rev. Psychol*, *50*, 537–567.

Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. *New Directions for Evaluation*, *1996*(70), 29–44. https://doi.org/10.1002/ev.1033

Kruschke, J. K. (2011). *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. Burlington, MA: Academic Press.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., … Can, T. and. (2016). caret: Classification and Regression Training. R package version 6.0-68. Retrieved from https://cran.r-project.org/package=caret

Lam, P. (2008). MCMC Methods : Gibbs Sampling and the Metropolis-Hastings Algorithm.

Laurie, H., & Lynn, P. (2009). The Use of Respondent Incentives on Longitudinal Surveys. In P. Lynn (Ed.), *Methodology of Longitudinal Surveys* (pp. 205–233). Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470743874.ch12

Lavrakas, P. J. (2008). *Encyclopedia of Survey Research Methods*. Thousand Oaks, Calif: SAGE Publications.

Lavrakas, P. J., Consultant, I., Dennis, J. M., Peugh, J., Shand-lubbers, J., Lee, E., & Charlebois, O. (2012). MAPOR 2012 Incentives paper – Lavrakas et al., (November), 1–11.

Lavrakas, P. J., McPhee, C., & Jackson, M. (2016). Conceptual Background on Response Propensity Modeling (RPM) for Allocating Differential Survey Incentives: JSM 2016 -

Survey Research Methods Section 3280 Purpose, Rationale, and Operationalization. In *Presented at the 71st annual conference of the American Association for Public Opinion Research*. Austin Texas.

Lebanon, G. (2006). Metropolis-Hastings and Gibbs Sampling. Retrieved from http://theanalysisofdata.com/notes/metropolis.pdf

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics*, *22*(2), 329–349.

Leite, W. (2017). *Practical Propensity Score Methods Using R*. Carlifornia: SAGE Publications.

Lemay, M., & Durand, C. (2002). The Effects of Interviewer Attitude on Survey Cooperation. *Bulletin de Methodologie Sociologique*, *76*, 27–44.

Lenis, D., Nguyen, T. Q., Dong, N., & Stuart, E. A. (2017). It's all about balance: propensity score matching in the context of complex survey data. *Biostatistics*, (February). https://doi.org/10.1093/biostatistics/kxx063

Lesser, V. M., Newton, L. A., & Yang, D. (2012). Comparing Item Nonresponse across Different Delivery Modes in General Population Surveys. *Survey Practice*, *5*(2), 1–4. https://doi.org/10.29115/SP-2012-0009

Lessler, J., & Kalsbeek, W. (1992). *Nonsampling Error in Surveys*. New York: Wiley.

Letki, N. (2006). Investigating the Roots of Civic Morality : Trust , Social Capital , and Institutional Performance. *Political Behavior*, *28*(4), 305–325. https://doi.org/10.1007/s11109-006-9013-6

Levy, P. S., Lemeshow, S., Groves, R. M., Kalton, G., & Rao, J. N. K. (2008). *Survey Errors and Survey Costs*.

Liebetrau, A. M. (1983). *Measures of Association*. Newbury Park, CA: Sage Publications.

Linden, A. (2015). Graphical displays for assessing covariate balance in matching studies. *Journal of Evaluation in Clinical Practice*, *21*(2), 242–247. https://doi.org/10.1111/jep.12297

Lindley, D. (1972). *Bayesian Statistics: A Review*. Philadelphia: Society for Industrial and Applied Mathematics.

Little, R. (1986). Survey nonresponse adjustments. *International Statistical Review*, *54*, 139–157. https://doi.org/10.2307/1403140

Lorah, J. (2018). Effect size measures for multilevel models : definition , interpretation , and TIMSS example. *Large-Scale Assessments in Education*, *6*(8). https://doi.org/10.1186/s40536-018-0061-2

Lozar Manfreda, K., & Vehovar, V. (2002). Do Mail and Web Survey Provide Same Results? *Metodološki Zvezki*, *18*, 149–169.

Lugtig, P., Lensvelt-Mulders, G. J. L. M., Frerichs, R., & Greven, A. (2011). Estimating nonresponse bias and mode effects in a mixed-mode survey. *International Journal of Market Research*, *53*(5), 669–686.

Lugtig, P., & Toepoel, V. (2016). The Use of PCs, Smartphones, and Tablets in a Probability-Based Panel Survey: Effects on Survey Measurement Error. *Social Science Computer Review*, *34*(1), 78–94. https://doi.org/10.1177/0894439315574248

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects : a comparative study. *Statistics in Medicine*, *23*, 2937–2960. https://doi.org/10.1002/sim.1903

Lunn, D. J., Spiegelhalter, D. J., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, *28*(25), 3049–3067. https://doi.org/10.1002/sim.3680

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. J. (2000). WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*(4), 325–337. https://doi.org/10.1023/A:1008929526011

Lyberg, L. E. (2012). Survey quality. *Survey Methodology*, *38*(2), 107–130.

Lyberg, L. E., & Kasprzyk, D. (2011). Data Collection Methods and Measurement Error: An Overview. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Seymour (Eds.), *Measurement Errors in Surveys* (pp. 237–257). New York: Wiley.

Lynch, S. M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. New York, NY: Springer.

Lynn, P. (2001). The impact of incentives on response rates to personal interview surveys: Role and perceptions of interviewers. *International Journal of Public Opinion Research*, *13*(3), 326–336.

Lynn, P. (2009). Sample design for Understanding Society. *Understanding Society Working Paper Series*, *2009-01*. Retrieved from http://research.understandingsociety.org.uk/publications/working-paper/2009-01

Mangione, T. W., Jr Fowler, F. J., & Thomas A., L. (1992). Question Characteristics and Interviewer Effects. *Journal of Official Statistics*, *8*(3), 293–307.

Mark, S., Telmo, N., Cord, H., Jonathon, M., Javier, S., Ron, T., … Ryan, K. (2016). epiR: Tools for the Analysis of Epidemiological Data. R package version 0.9-74. Retrieved from https://cran.r-project.org/package=epiR

Mayer, T. S., & Brien, E. O. (2001). Interviewer Refusal Aversion Training to Increase Survey Participation. *Proceedings of the Annual Meeting of the American Statistical Association*, August 5–9, 2001.

Mcgrath, D. E. (2005). An Incentives Experiment in the U . S . Consumer Expenditure Quarterly Survey. *ASA Section on Survey Research Methods*, 3411–3418.

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *The Journal of Mathematical Sociology*, *4*(1), 103–120. https://doi.org/10.1080/0022250X.1975.9989847

Mercer, A., Caporaso, A., Cantor, D., & Townsend, R. (2015). How much gets you how much? Monetary incentives and response rates in household surveys. *Public Opinion Quarterly*, *79*(1), 105–129. https://doi.org/10.1093/poq/nfu059

Merkle, D. M., & Edelman, M. (2002). Nonresponse in Exit Polla: A comprehensive Analysis. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 243–258). New York: Wiley.

Meterko, M., Restuccia, J. D., Stolzmann, K., Mohr, D., Brennan, C., Glasgow, J., & Kaboli, P. (2015). Response rates, nonresponse bias, and data quality: Results from a national survey of senior healthcare leaders. *Public Opinion Quarterly*, *79*(1), 130–144. https://doi.org/10.1093/poq/nfu052

Meyers, J. L., & Beretvas, S. N. (2006). The Impact of Inappropriate Modeling of Cross-Classified Data Structures The Impact of Inappropriate Modeling of Cross-Classified Data Structures. *Multivariate Behavioral Research*, *41*(4), 473–497. https://doi.org/10.1207/s15327906mbr4104

Mizes, J. S., Fleece, E. L., & Roos, C. (1984). Incentives for Increasing Return Rates: Magnitude Levels, Response Bias, amd Format. *Public Opinion Quarterly*, *48*(4), 794. https://doi.org/10.1086/268885

Moore, J. C., Durrant, G. B., & Smith, P. W. F. (2018). Data set representativeness during data collection in three UK social surveys: generalizability and the effects of auxiliary covariate choice. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *181*(1), 229–248. https://doi.org/10.1111/rssa.12256

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

Moss, G. M. (1981). Factors Affecting Response Rate and Response Speed in a Mail Survey of Part-Time University Students, *XI*(3).

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*(3), 691–692. https://doi.org/10.1093/biomet/78.3.691

Newman, J. C., Des Jarlais, D. C., Turner, C. F., Gribble, J., Cooley, P., & Paone, D. (2002). The differential effects of face-to-face and computer interview modes. *American Journal of Public Health*, *92*(2), 294–297. https://doi.org/10.2105/AJPH.92.2.294

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

Nguyen, T. L., Collins, G. S., Spence, J., Daurès, J. P., Devereaux, P. J., Landais, P., & Le Manach, Y. (2017). Double-adjustment in propensity score matching analysis: choosing a threshold for considering residual imbalance. *BMC Medical Research Methodology*, *17*(1), 1–8. https://doi.org/10.1186/s12874-017-0338-0

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., … Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. New York: Wiley.

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

Olson, K. (2006). Survey participation , nonresponse bias , measurement error bias , and total bias. *Public Opinion Quarterly*, *70*(5), 737–758. https://doi.org/10.1093/poq/nfl038

Olson, K., & Groves, R. M. (2012). An examination of within-person variation in response propensity over the data collection field period. *Journal of Official Statistics*, *28*(1), 29–51.

Olson, K., Smyth, J. D., & Wood, H. M. (2012). Does Giving People Their Preferred Survey Mode Actually Increase Survey Participation Rates? An Experimental Examination. *Public Opinion Quarterly*, *76*(4), 611–635. https://doi.org/10.1093/poq/nfs024

Oravecz, Z., Huentelman, M., & Vandekerckhove, J. (2015). Sequential Bayesian updating for Big Data (in press). *Big Data in Cognitive Science: From Methods to Insights*, 100–150.

Patrick, M. E., Singer, E., Boyd, C. J., Cranford, J. A., & Mccabe, S. E. (2013). Incentives for College Student Participation in Web-Based Substance Use Surveys. *Journal of Addict Behaviour*, *38*(3), 1710–1714. https://doi.org/10.1016/j.addbeh.2012.08.007.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford: Oxford University Press.

Peterson, G., Griffin, J., LaFrance, J., & Li, J. (2017). Smartphone Participation in Web Surveys. In P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. Lyberg, … B. T. West (Eds.), *Total Survey Error in Practice* (pp. 203–233). Hoboken,NJ. https://doi.org/10.1002/9781119041702.ch10

Petrolia, D. R., & Bhattacharjee., S. (2009). Revisiting Incentive Effects Evidence from a Random-Sample Mail Survey on Consumer Preferences for Fuel EthanoL. *Public Opinion Quarterly*, *73*(3), 537–550. https://doi.org/10.1093/poq/nfp038

Peytcheva, E., & Groves, R. M. (2009). Using variation in response rates of demographic subgroups as evidence of nonresponse bias in survey estimates. *Journal of Official Statistics*, *25*(2), 193.

Pfeffermann, D., & Rao, C. R. (2009). *Handbook of Statistics_29A: Sample Surveys: Design, Methods and Applications. Vol. 29A*. Elsevier.

Pforr, K., Blohm, M., Blom, A. G., Erdel, B., Felderer, B., Fräßdorf, M., … Rammstedt, B. (2015). Are incentive effects on response rates and nonresponse bias in large-scale, face-to-face surveys generalizable to Germany? Evidence from ten experiments. *Public Opinion Quarterly*, *79*(3), 740–768. https://doi.org/10.1093/poq/nfv014

Pickery, J., & Loosveldt, G. (2002). A Multilevel Multinomial Analysis of Interviewer Effects on Various Components of Unit Nonresponse. *Quality & Quantity*, (36), 427–437.

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

Pischke, S. (2007). Lecture Notes on Measurement Error. https://doi.org/http://dx.doi.org/10.1055/s-0029-1246197

Plewis, I., Ketende, S., & Calderwood, L. (2012). Assessing the accuracy of response propensity models in longitudinal studies. *Survey Methodology*, *38*(2), 167–171.

Purdon, S., Campanelli, P., & Sturgis, P. (1999). Interviewers ' Calling Strategies on Face-to-Face Interview Surveys. *Journal of Offcial Statistics*, *15*(2), 199–216.

Putnam, R. D. (1995a). Bowling Alone : America ' s Declining Social Capital. *Journal of Democracy*, *6*(1), 65–78.

Putnam, R. D. (1995b). Social Capital: Measurement and Consequences. Retrieved from http://www.oecd.org/innovation/research/1825848.pdf

R Core Team. (2015). A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.r-project.org/

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

Rasbash, J., Steele, F., Browne, W. J., Goldstein, H., & Charlton, C. M. J. (2012). A User ' s Guide to MLwiN, v2.26. *Centre for Multilevel Modelling, University of Bristol*.

Rees, P. H. (1971). Factorial ecology: An extended definition, survey and critique of the field. *Economic Geography*, *47*, 220–233. https://doi.org/10.2753/JEI0021-3624440403

Revilla, M. A. (2010). Quality in Unimode and Mixed-Mode designs: A Multitrait-Multimethod approach. *Survey Research Methods*, *4*(3), 151–164.

Revilla, M. A., & Saris, W. E. (2013). A Comparison of the Quality of Questions in a Face-to-face and a Web Survey. *International Journal of Public Opinion Research*, *25*(2), 242–253.

Ridgeway, G., Kovalchik, S. A., Griffin, B. A., & Kabeto, M. U. (2015). Propensity Score Analysis with Survey Weighted Data. *Journal of Causal Inference.*, *3*(2), 237–249. https://doi.org/10.1515/jci-2013-0007.Targeted

Rindfuss, R. R., Choe, M. K., Tsuya, N. O., Bumpass, L. L., & Tamaki, E. (2015). Do low survey response rates bias results? Evidence from Japan. *Demographic Research*, *32*, 797–828. https://doi.org/10.4054/DemRes.2015.32.26

Roberts, C. (2007). Mixing modes of data collection in surveys : A methodological review ESRC National Centre for Research Methods. *NCRM Methods Review Paper (Unpublished)*, (March), 1–26. Retrieved from http://eprints.ncrm.ac.uk/418/

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77. https://doi.org/10.1186/1471-2105-12-77

Roose, H., Waege, H., & Agneessens, F. (2003). Respondent related correlates of response behaviour in audience research. *Quality and Quantity*, *37*(4), 411–434. https://doi.org/10.1023/A:1027379211819

Rosenbaum, P. R. (1987). Model-Based Direct Adjustment. *Journal of the American Statistical Association*, *82*(398), 387. https://doi.org/10.2307/2289440

Rosenbaum, P. R. (1989). Optimal Matching for Observational Studies. *American Statistical Association*, *84*(408), 1024–1032.

Rosenbaum, P. R. (2002). *Observational Studies*. New York: Springer.

Rosenbaum, P. R., & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score P. *The American Statistician*, *39*(1), 33–38.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, (1), 34–58.

Rubin, D. B., & Rosenbaum, P. R. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*, 516–524.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations.pdf. *Journal of the Royal Statistical Society*, *Series B*(71), 319–392.

Rue, H., Martino, S., & Chopin, N. (2010). Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society*, *71*(2), 319–392.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., & Lindgren, F. K. (2016). Bayesian Computing with INLA: A Review, 1–26. Retrieved from http://arxiv.org/abs/1604.00860

Sahlqvist, S., Song, Y., Bull, F., Adams, E., Preston, J., & Ogilvie, D. (2011). Effect of questionnaire length, personalisation and reminder type on response rate to a complex postal survey: Randomised controlled trial. *BMC Medical Research Methodology*, *11*(1), 62. https://doi.org/10.1186/1471-2288-11-62

Sakshaug, J. W., & Eckman, S. (2017). Following Up with Nonrespondents via Mode Switch and Shortened Questionnaire in an Economic Survey: Evaluating Nonresponse Bias, Measurement Error Bias, and Total Bias. *Journal of Survey Statistics and Methodology*, *5*(4), 454–479. https://doi.org/10.1093/jssam/smw039

Särndal, C.-E., & Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. John Wiley & Sons, Ltd. https://doi.org/10.1002/0470011351

Särndal, C.-E., & Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review/Revue Internationale de Statistique*, 279–294.

Schaeffer, N. C., Dykema, J., & Maynard, D. W. (2010). Interviewers and Interviewing. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research* (pp. 437–470). Binley, UK: Emerald.

Schejbal, J. A., & Lavrakas, P. J. (1995). Panel attrition in a dual-frame local area telephone surve. In *Proceedings of the American Statistical Association* (pp. 1035–1039). Section on Survey Research Methods.

Schnell, R., & Trappmann, M. (2006). *The Effect of the Refusal Avoidance Training Experiment on Final Disposition Codes in the German ESS-2* (No. 3). Retrieved from https://www.uni-due.de/~hq0215/documents/2006/2006_EffectRefusalAvoidanceTraining.pdf

Schober, M. F. (2018). The future of face-to-face interviewing. *Quality Assurance in Education*, *26*(2), 290–302. https://doi.org/10.1108/QAE-06-2017-0033

Schouten, B., Calinescu, M., & Luiten, A. (2012). Optimizing quality of response through adaptive survey designs. *Survey Methodology*, *39*(1), 29–58.

Schouten, B., & Cobben, F. (2007). *R-indexes for the comparison of different fieldword strategies and data collection modes*. Retrieved from http://hummedia.manchester.ac.uk/institutes/cmist/risq/schouten-cobben-2007-a.pdf

Schouten, B., Mushkudiani, N., Shlomo, N., & Durrant, G. B. (2018). A Bayesian analysis of design parameters in survey data collection. *Survey Statistics and Methodology*, 1–34. https://doi.org/10.1093/jssam/smy012

Schouten, B., Shlomo, N., & Skinner, C. (2011). Indicators for monitoring and improving representativeness of response. *Journal of Official Statistics*, *27*(2), 231–253.

Schouten, B., van den Brakel, J., Buelens, B., van der Laan, J., & Klausch, T. (2013). Disentangling mode-specific selection and measurement bias in social surveys. *Social*

Science Research, *42*(6), 1555–1570. https://doi.org/10.1016/j.ssresearch.2013.07.005

Schröder, M., Saßenroth, D., Körtner, J., Kroh, M., & Schupp, J. (2013). Experimental Evidence of the Effect of Monetary Incentives on Cross-Sectional and Longitudinal Response: Experiences from the Socio-Economic Panel (SOEP).

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Belmont CA: Wadsworth, Cengage Learning.

Shettle, C., & Mooney, G. (1999). Monetary Incentives in U . S . Government Surveys. *Journal of Official Statistics Statistics, 15*(2), 231–250.

Simon, J. (2009). *Bayesian Analysis for the Social Sciences*. West Sussex,UK: Wiley.

Simpson, D. P. (1998). Bayesian Computation Using INLA. *Technometrics*, *40*(2), 164. https://doi.org/10.2307/1270667

Singer, E. (2002). The Use of Incentives to Reduce Nonresponse in Household Surveys. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey Nonresponse* (pp. 163–177). New York: Wiley.

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

Singer, E., Groves, R. M., & Corning, A. D. (1999). Differential incentives: Beliefs about practices, perceptions of equity, and effects on survey participation. *Public Opinion Quarterly*, *63*(2), 251–260. https://doi.org/10.1086/297714

Singer, E., Hoewyk, J. Van, Gebler, N., Raghunathan, T., & Mcgonagle, K. (1999). The Effect of Incentives on Response Rates in Interviewer-Mediated Surveys. *Journal of Offcial Statistics*, *15*(2), 217–231.

Singer, E., Hoewyk, J. Van, & Maher, M. P. (2000). Experiments with Incentives in Telephone Surveys. *Public Opinion Quarterly*, *64*(2), 171–188.

Singer, E., & Kohnke-Aguirre, L. (1977). Interviewer Expectation Effects : A Replication and Extension. *Public Opinion Quarterly*, *63*, 245–260.

Singer, E., & Kulka, R. A. (2002). Paying respondents for survey participation. In M. Ver Ploeg, R. A. Moffitt, & C. F. Citro (Eds.), *In Studies of welfare populations: Data collection and research issues* (pp. 105–128). Washington, DC: National Academy Press.

Singer, E., & Maher, M. P. (2000). Experiments with Incentives in Telephone Surveys. *Public Opinion Quarterly*, *64*(2), 171–188.

Singer, E., & Ye, C. (2013). The Use and Effects of Incentives in Surveys. *The ANNALS of the American Academy of Political and Social Science*, *645*(1), 112–141. https://doi.org/10.1177/0002716212458082

Sinibaldi, J., & Eckman, S. (2015). Using call-level interviewer observations to improve response propensity models. *Public Opinion Quarterly*, *79*(4), 976–993. https://doi.org/10.1093/poq/nfv035

Sinibaldi, J., Trappmann, M., & Kreuter, F. (2014). Which is the better investment for nonresponse adjustment: Purchasing commercial auxiliary da ta or collecting interviewer observations? *Public Opinion Quarterly*, *78*(2), 440–473. https://doi.org/10.1093/poq/nfu003

Sirken, M. G., Herrmann, D. J., Schechter, S., Schwarz, N., Tanur, J. M., & Roger, T. (1999). *Cognition and Survey Research*. New York: John Wiley & Sons.

Smith, J. A., & Todd, P. E. (2005). *Does matching overcome LaLonde's critique of nonexperimental estimators? Journal of Econometrics* (Vol. 125). https://doi.org/10.1016/j.jeconom.2004.04.011

Smith, T. W. (1983). The Hidden 25 Percent: An Analysis of Nonresponse on the 1980 General Social Survey. *Public Opinion Quarterly*, *47*(3), 386. https://doi.org/10.1086/268797

Smith, T. W. (2011). The Report of the International Workshop on Using Multi-level Data from Sample Frames, Auxiliary Databases, Paradata and Related Sources to Detect and Adjust for Nonresponse Bias in Surveys*. *International Journal of Public Opinion Research*, *23*(3), 389–402. https://doi.org/10.1093/ijpor/edr035

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel Analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London: Sage Publishers.

Snijkers, G., Hox, J. J., & de Leeuw, E. D. (1999). Interviewers' tactics for fighting survey nonresponse. *Journal of Official Statistics*, *15*(2), 185–198. Retrieved from http://dspace.library.uu.nl/handle/1874/23751%5Cnhttp://dspace.library.uu.nl/bitstream/1874/23751/1/hox_99_interviewers' tactics for fighting .pdf

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian Measures of Model Complexity anf Fit. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, *64*(4), 583–639. https://doi.org/10.1111/1467-9868.00353

Starks, H., Diehr, P., & Curtis, J. R. (2009). The Challenge of Selection Bias and Confounding in Palliative Care Research. *Journal of Palliative Medicine*, *12*(2), 181–187.

https://doi.org/10.1089/jpm.2009.9672

Steel, D. (2007). Bayesian Confirmation Theory and The Likelihood Principle. *Synthese*, *156*(1), 53–77. https://doi.org/10.1007/s11229-005-3492-6

Stoop, I. A. L. (2005). *The hunt for the last respondent*. The Hague, Netherlands.

Stratford, N., Simmonds, N., & Nicolaas, G. (2002). National Travel Survey 2002: Incentives Experiment Report. Retrieved from http://webarchive.nationalarchives.gov.uk/+/http://www2.dft.gov.uk/pgr/statistics/datatablespublications/nts/

Stuart, E. a. (2010). Matching methods for causal inference: A review and a look forward. *Stat*, *25*(1), 1–21. https://doi.org/10.1214/09-STS313.Matching

Sturgis, P., Allum, N., & Brunton-Smith, I. (2009). Attitudes Over Time: The Psychology of Panel Conditioning. In *Methodology of Longitudinal Surveys* (pp. 113–126). Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470743874.ch7

Sturgis, P., Williams, J., Brunton-Smith, I., & Moore, J. (2017). Fieldwork effort, response rate, and the distribution of survey outcomes. *Public Opinion Quarterly*, *81*(2), 523–542. https://doi.org/10.1093/poq/nfw055

Suchman, L., & Jordan, B. (1990). Interactional troubles in face-to-face survey interviews. *Journal of the American Statistical Association*, *85*(409), 232–241. https://doi.org/10.1080/01621459.1990.10475331

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco,CA: Jossey-Bass.

Szolnoki, G., & Hoffmann, D. (2013). Online, face-to-face and telephone surveys - Comparing different sampling methods in wine consumer research. *Wine Economics and Policy*, *2*(2), 57–66. https://doi.org/10.1016/j.wep.2013.10.001

Taylor, B. M., & Diggle, P. J. (2014). INLA or MCMC? A Tutorial and Comparative Evaluation for Spatial Prediction in log-Gaussian Cox Processes. *Journal of Statistical Computation and Simulation*, *84*(10), 2266–2284. https://doi.org/10.1080/00949655.2013.788653

Toepoel, V. (2012). Building Your Own online Panel Via E-mail and Other Digital Media in Handbook of Survey Methodology for the Social Sciences. In G. Lior (Ed.), *Handbook of Survey Methodology for the Social Sciences* (pp. 345–360). New York: Springer.

Tourangeau, R. (2017). Mixing Modes: Tradeoffs Among Coverage, Nonresponse, and

Measurement Error. In P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. Lyberg, … B. T. West (Eds.), *Total Survey Error in Practice* (pp. 115–132). Hoboken,NJ.

Tourangeau, R., Brick, M. J., Lohr, S., & Li, J. (2016). Adaptive and responsive survey designs: A review and assessment. *Journal of the Royal Statistical Society. Series A: Statistics in Society*. https://doi.org/10.1111/rssa.12186

Tourangeau, R., Conrad, F. G., & Couper, M. P. (2013). *The Science of Web Surveys*. Oxford: Oxford University Press.

Tourangeau, R., & Plewes, T. J. (2013). *Nonresponse in Social Science Surveys*. Washington, D.C.: National Academies Press. https://doi.org/10.17226/18293

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511819322

Tourangeau, R., & Yan, T. (2007). Sensitive Questions in Surveys. *Psychological Bulletin*, *133*(5), 859–883. https://doi.org/10.1037/0033-2909.133.5.859

van de Schoot, R., Broere, J. J., Perryck, K. H., Zondervan-Zwijnenburg, M., & van Loey, N. E. (2015). Analyzing small data sets using Bayesian estimation: the case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *European Journal of Psychotraumatology*, *6*(1), 25216. https://doi.org/10.3402/ejpt.v6.25216

Vannieuwenhuyze, J., & Loosveldt, G. (2013). Evaluating Relative Mode Effects in Mixed-Mode Surveys:: Three Methods to Disentangle Selection and Measurement Effects. *Sociological Methods and Research*, *42*(1), 82–104. https://doi.org/10.1177/0049124112464868

Vannieuwenhuyze, J., Loosveldt, G., & Molenberghs, G. (2010). A method for evaluating mode effects in mixed-mode surveys. *Public Opinion Quarterly*, *74*(5), 1027–1045. https://doi.org/10.1093/poq/nfq059

Vannieuwenhuyze, J., Loosveldt, G., & Molenberghs, G. (2017). MIXED-MODE SURVEYS, *74*(5), 1027–1045. https://doi.org/10.1093/poq/nfq059

Vassallo, R., Durrant, G. B., & Smith, P. W. F. (2016). Separating interviewer and area effects by using a cross-classified multilevel logistic model : simulation findings and implications for survey designs. *Journal of the Royal Statistical Society Series A (Statistics in Society)*, *180*(2), 531–550. https://doi.org/10.1111/rssa.12206

Vassallo, R., Durrant, G. B., Smith, P. W. F., & Goldstein, H. (2015). Interviewer effects on non-response propensity in longitudinal surveys: a multilevel modelling approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *178*(1), 83–99.

https://doi.org/10.1111/rssa.12049

Vehovar, V., Slavec, A., & Berzelak, N. (2012). Costs and Errors in Fixed and Mobile Phone Surveys. In *Handbook of Survey Methodology for the Social Sciences* (pp. 277–295). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-3876-2_16

Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., … Thompson, L. (2014). Use of historical control data for assessing treatment effects in clinical trials. *Pharm Stat.*, *13*(1), 41–54. https://doi.org/10.1002/pst.1589.Use

Villar, A., & Fitzgerald, R. (2017). Using mixed modes in survey data research: Results from six experiments. In M. Breen (Ed.), *Values and Identities in Europe: Evidence from the European Social Survey* (pp. 273–310). Routledge.

Voogt, R. J. J., & Saris, W. E. (2005). Mixed Mode Designs: Finding the Balance Between Nonresponse Bias and Mode Effects. *Journal of Official Statistics*, *21*(3), 367–387. https://doi.org/10.1093/poq/nfn045

Wagner, J. (2016). Using Bayesian Methods to Estimate Response Propensity Models During Data Collection. Retrieved December 22, 2016, from http://hummedia.manchester.ac.uk/institutes/cmist/BADEN/workshop-2016/AAPOR_James.pdf

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond “ *p* &lt; 0.05.” *The American Statistician*, *73*(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Weisberg, H. F. (2005). *The Total Survey Error Approach:a guide to the new science of surveyresearch*. Chicago: University of Chicago Press. https://doi.org/10.7208/chicago/9780226891293.001.0001

Welham, S. J., Gezan, S. A., Clark, S. J., & Mead, A. (2014). *Statistical Methods in Biology: Design and Analysis of Experiments and Regression*. Chapman and Hall/CRC.

West, B. T. (2011). Paradata in Survey Research. *Survey Practice*, *4*(4), 1–8.

West, B. T. (2013). An examination of the quality and utility of interviewer observations in the National Survey of Family Growth. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *176*(1), 211–225. https://doi.org/10.1111/j.1467-985X.2012.01038.x

West, B. T., & Blom, A. G. (2017). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology*, *5*(2), 175–211.

https://doi.org/10.1093/jssam/smw024

West, B. T., & Elliott, M. R. (2014). Frequentist and bayesian approaches for comparing interviewer variance components in two groups of survey interviewers. *Survey Methodology*, *40*(2), 163–188.

West, B. T., & Groves, R. M. (2011). The PAIP Score : A Propensity-Adjusted Interviewer Performance Indicator, 5631–5645.

West, B. T., & Kreuter, F. (2015). A Practical Technique for Improving the Accuracy of Interviewer Observations of Respondent Characteristics. *Field Methods*, *27*(2), 144–162. https://doi.org/10.1177/1525822X14549429

West, B. T., & Sinibaldi, J. (2013). The Quality Of Paradata: A Literature Review. *Improving Surveys with Paradata*, 339–360.

Williams, D., & Brick, M. J. (2018). Trends in U.S. Face-To-Face Household Survey Nonresponse and Level of Effort. *Journal of Survey Statistics and Methodology*, *6*(2), 186–211. https://doi.org/10.1093/jssam/smx019

Williams, J. (2017a). Address Based Online Surveying (ABOS). Retrieved March 18, 2018, from http://the-sra.org.uk/wp-content/uploads/joel-williams-address-based-online-surveying.pdf

Williams, J. (2017b). *Community Life Survey Disentangling sample and mode effects*. Retrieved from Community Life Survey Disentangling sample and mode effects

Williamson, E., Morley, R., Lucas, A., & Carpenter, J. (2012). Propensity scores: From naïve enthusiasm to intuitive understanding. *Statistical Methods in Medical Research*, *21*(3), 273–293. https://doi.org/10.1177/0962280210394483

Winkler, R. L. (1967). The Assessment of Prior Distributions in Bayesian Analysis. *Journal of the American Statistical Association*. https://doi.org/10.2307/2283671

Wolf, C., Joye, D., Smith, T. W., & Fu, Y. (2016). *The SAGE Handbook of Survey Methodology*. 1 Oliver's Yard, 55 City Road London EC1Y 1SP: SAGE Publications Ltd. https://doi.org/10.4135/9781473957893

Wright, G. (2015). An empirical examination of the relationship between nonresponse rate and nonresponse bias. *Statistical Journal of the IAOS*, *31*(2), 305–315. https://doi.org/10.3233/sji-140844

Yan, T., & Curtin, R. (2010). The Relation Between Unit Nonresponse and Item Nonresponse :

A Response Continuum Perspective. *International Journal of Public Opinion Research*, *22*(4), 535–551.

Yu, J., & Cooper, H. (1983). A Quantitative Review of Research Design Effects on Response Rates to Questionnaires. *Journal of Market Research*, *20*(1), 36–44.

Yu, R.-R., Liu, Y.-S., & Yang, M.-L. (2014). Does Interviewer Personality Matter for Survey Outcomes? Evidence from a Face-to-face Panel Study of Taiwan. *調查研究 --- 方法與應用*, *31*, 89–121.

Yu, R., & Abdel-Aty, M. (2013). Investigating different approaches to develop informative priors in hierarchical Bayesian safety performance functions. *Accident Analysis and Prevention*, *56*, 51–58. https://doi.org/10.1016/j.aap.2013.03.023

Zanutto, E. L. (2006). A Comparison of Propensity Score and Linear Regression Analysis of Complex Survey Data. *Journal of Data Science*, *4*, 67–91. https://doi.org/10.6339/JDS.2006.04(1).233

Zhu, M., & Lu, A. Y. (2004). The Counter-intuitive Non-informative Prior for the Bernoulli Family. *Journal of Statistics Education*, *12*, 1–10.

Ziegenfuss, J. Y., Burmeister, K. R., Harris, A., Holubar, S. D., & Beebe, T. J. (2012). Telephone follow-up to a mail survey: when to offer an interview compared to a reminder call. *BMC Medical Research Methodology*, *12*(1), 32. https://doi.org/10.1186/1471-2288-12-32

Zyphur, M. J., & Oswald, F. L. (2015). *Bayesian Estimation and Inference: A User's Guide*. *Journal of Management* (Vol. 41). https://doi.org/10.1177/0149206313501200