Estimating mode Effects Using Propensity Score Methods in Community Life Survey

Eliud Kibuchi, *University of Southampton*

Patrick Sturgis, *University of Southampton*

Gabi Durrant, *University of Southampton*

Olga Maslovskaya, *University of Southampton*

Joel Williams, *Kantar Public*

Abstract

Social sciences are currently relying on mixed mode for carrying out surveys with aim of improving response rates, increasing coverage, and reducing costs. However, systematic differences may be present in such data resulting to biased estimates of mode effects. The propensity score, defined as the conditional probability of treatment assignment, given observed the covariates, can be used to balance the covariates in survey modes, and therefore reduce bias. In this study, propensity score methods are used to separate mode effects from selection effects in United Kingdom's Community Life Survey collected using face-to-face interviews and online questionnaires.

Keywords: propensity score methods, face-to-face, online, mode effects,

Introduction

The issue of maintaining and improving survey quality while keeping the financial costs in check has received considerable critical attention in recent years. In response to this, survey researchers are becoming more innovative by adapting and modifying survey operations on continuous basis. Recently, the use of the online surveys for data collection as a way of addressing the issue of survey quality while maintaining costs at desired levels has received increasing attention (de Leeuw, 2005). Traditionally, many surveys relied on face-to-face , mail and telephone surveys for data collection that are costly and time inefficient (Dillman, Phelps, et al., 2009; Kreuter, 2013; Lyberg & Kasprzyk, 2011). Nowadays, most surveys combine different modes with an aim of increasing response rates, saving costs, and improving measurement quality (de Leeuw, 2005). This may be achieved by following up non-respondents in a different mode, starting survey with a cheaper mode, and using self-completion questionnaires for sensitive questions.

Despite the growth of mixed-mode designs, there is increasing concern of the presence of mode effects caused by varying channels of communication, degree of interviewer involvement, and degree of contact with the respondent (Couper, 2011; de Leeuw, 2005; Dillman & Christian, 2005). Mode effects may affect the quality of the mixed-mode data because it simultaneously create selection effects and measurement effects. Selection effects occur when different types of respondents choose different modes to complete the survey (Dillman, Phelps, et al., 2009; Voogt & Saris, 2005). On the other hand, measurement effects refer to the influence of a survey mode on the answers respondents give (Voogt & Saris, 2005). That is, same respondent will give different answers based on the mode used. Measurement effects results due to measurement errors caused by primacy and recency effects, recall bias, acquiescence, interviewer effects and social desirability (de Leeuw, 2005; Dillman, Smyth, & Christian, 2009; O'Muircheartaigh, 1997).

The key challenge among survey methodologist is to investigate unbiased measurement effects in mixed-mode surveys. This is because sample and mode effects are confounded which rises a key research question of the most effective ways of the separating them. One of the methods proposed that can be used to separate mode and sample effects is propensity score matching (Lugtig, Lensvelt-Mulders, Frerichs, & Greven, 2011). Propensity score matching involves generating a new sample based on the propensity scores (Austin, 2011; P R Rosenbaum & Rubin, 1983). The propensity score is defined as the probability of treatment selection given observed baseline covariates and was introduced by Rosenbaum and Rubin (1983). Lugtig et al. (2011) showed that propensity score matching is effective in explaining differences caused by the sample selections effects in Computer Assisted Telephone Interview (CATI) and Web Assisted Personal Interview

(WAPI). It also important to understand whether propensity score matching is similarly effective under different survey settings especially in complex designs. For example, Lenis et al. (2017) found that incorporation of survey weights in propensity scores does not have an impact estimation of treatment effects. In addition, they recommend that survey weights of the matched control units should inherit the survey weights of the treated units to which they were matched in the outcome analysis. However, recent study by Austin, Jembere, and Chiu (2018) had an inconclusive findings of whether survey weights should be included in estimating propensity scores or not. Austin, Jembere, and Chiu (2018) also recommends that matched control units should retain their natural weights, rather than interitiung the survey weights of the treated unit to which they are matched. Informed by Lenis et al. (2017 and Austin, Jembere, and Chiu (2018) findings it becomes necessary to address the isssue of incorporating survey weights in propensity score matching for mode effects studies. Secondly, whether or not survey weights have an effect in outcome analysis of mode effects especially the decision to use natural or inherited weights on the control units .

Efficient implementation of propensity score matching between two survey samples obtained using different modes will reduce sample effects for a successful estimation of mode effects. The mode effects are caused by the way different survey modes vary in the extent they provide access to survey populations and costs involved. For example, face-to-face are viewed to be nationally representative and also costly (Holbrook, Krosnick, & Pfent, 2007; Roberts, 2007). On the other hand, online surveys are cost effective relative to face-to-face interviews although they are not viewed as being nationally representative (Roberts, 2007). The existing literature on survey modes shows mixed evidence on their effect on survey quality (de Leeuw, 2005; Dillman, Phelps, et al., 2009). For example, online surveys generate data with of lower quality compared to face-to-face surveys (Heerwegh, 2009; Heerwegh & Loosveldt, 2008; Lozar Manfreda & Vehovar, 2002). In addition, face-to-face interviews are susceptible to social desirability bias in for surveys with sensitive questions compared to online surveys (Lee & Sargeant, 2011; Szolnoki & Hoffmann, 2013). However, Dodou and De Winter (2014) noted that some of the social desirability bias may be attributed to sampling errors rather than mode effects. Recently, Williams (2017) found that mode effects are twice the sample effects in a survey conducted using face-to-face and online/postal questionnaires. Lugtig et al. (2011) found that propensity score matching is good at explaining differences that can be attributed to by sample selection on two online samples. What remains unclear, however, is whether similar conclusions may be obtained in the UK context.

The main objective of this paper is to add to the body of evidence that seeks to address two main research questions. Firstly, it aims to evaluate the relative performance of different methods of implementing and formulating propensity score models in mixed-mode studies under complex survey setting. Secondly, it aims to use propensity score methods to disentangle selection and measurement effects in face-to-face and online surveys. In order to answer these questions the focus involves obtaining the change in distribution of attitudinal and behavioural survey outcomes taken from Community Life Survey (CLS). The CLS addresses the latest trends in areas of volunteering, well-being, charitable giving, local actions and networks in United Kingdom. For every survey variable, Absolute Percentage Differences (APD) based on the approach proposed by Sturgis et al. (2017) are computed. The APD estimates obtained after propensity score matching represent mode effects. In addition, the difference in APD estimates before and after matching accounts for selection effects.

The first section of this paper gives literature reviews on mode effects, selection effects and propensity score matching. Section 2 describes the data, followed by methodology employed for the analysis in section 3. The key findings from analyses are presented in section 4, and section 5 discusses implications of the results for survey practice

Background

According to Jäckle, Roberts, and Lynn (2010) the choice of the mode of data collection has potential to influence the way respondents answer questions. Currently, there is a vast wealth of literature investigating the size and the cause of these mode effects (de Leeuw, 2005; Dillman, 2000; Jäckle et al., 2010). Overall, previous research has established that face-to-face surveys have key strengths compared to other modes in terms of complexity and quality of the data collected (de Leeuw, 2005; Szolnoki & Hoffmann, 2013). This is because face-to-face interviews are mainly well structured, flexible and adaptable due to personal interaction between interviewers and respondents (Dillman, 2007). However, this comes with significantly higher costs as well as additional potential sources of response and interviewer bias caused by norms of social interaction (de Leeuw, 2005). Largely due to the higher costs in face-to-face interviews, the use of online surveys has been on the rise over recent years. The online surveys are cost effective, enable fast data processing, and are flexible in terms of providing more complex displays and designs (Beebe, Mika, Harrison, Anderson, & Fulkerson, 1997). However, online surveys may be unrepresentative of the population due to selection and coverage errors (Blasius & Brandt, 2010). Selection bias arises because initiative to either participate or not in online surveys is voluntarily among targeted respondents and also internet access is not 100 percent among member of the population (Couper, 2000; Hoogendoorn & Daalmans, 2009). In addition, it is difficult to verify the identity of the surveyed person affecting the relevance of the collected information.

Over the last 10 years, there has been an increasing amount of studies comparing face-to-face interviews and online surveys and often generating complex and, at times, contradictory results. The mode effects are often examined in terms of social desirability and data quality when using interviewer and self-administered questionnaires. Social desirability bias is type of measurement error that occurs when a respondent provides an answer that is more socially acceptable. For example, respondents usually over report socially desirable behaviours such as voting (Holbrook & Krosnick, 2010; Silver, Anderson, & Abramson, 1986) and donating to charitable organisations (Bekkers & Wiepking, 2011; Lee & Sargeant, 2011). However, there is under-reporting when it comes to answering socially undesirable behaviours such as drug use or other stigmatised behaviour (Newman et al., 2002). For example, Newman et al. (2002) found a positive effect of abstinence on sensitive questions and other stigmatised behaviour in face-to-face interviews compared to computer-assisted self-interviewing (CASI) . Dillman et al. (2009) found that respondents on telephone interviews tend to give more extreme responses on positive ends of the scale on satisfaction and dissatisfaction questions compared to web surveys. However, Szolnoki and Hoffmann (2013) did not find any social desirability bias on questions about wine consumption frequency and preferences in face-to-face, telephone and online quota surveys. A meta-analysis on 51 studies by Dodou and De Winter (2014) found that there is no difference in social desirability between paper-and-pencil and online surveys and any previous large effects may be due to sampling errors. Although online surveys are viewed as preferred modes when asking questions on socially undesirable behaviour, many respondents are now becoming more concerned about the information they provide because of the insecure nature of online environment (Corritore, Kracher, & Wiedenbeck, 2003)

Heerwegh and Loosveldt (2008) find that data quality is lower in online surveys due to item nonresponse compared to face-to-face surveys. This is because online surveys are mainly completed in less controlled environment making it likely to have higher incidences of item nonresponse (Lozar Manfreda & Vehovar, 2002). In addition, internet use is associated with multi-tasking that may distract some respondents making them skip some questions (Lozar Manfreda & Vehovar, 2002). Online respondents are also more likely to use less mental energy leading to a potentially higher degree of satisficing compared to face-to-face where an interviewer is present for questioning and guidance (Heerwegh & Loosveldt, 2008; Krosnick, 1991). Heerwegh and Loosveldt (2008) and (Heerwegh, 2009) found that online surveys are more likely to have “don’t know” responses compared to face-to-face. Although, many studies have found differences in data quality between face-to-face and online surveys, there is need for further research to investigate the extent of measurement effects under different population and survey topics. In addition, internet access though mobile phones and other portable devices has improved over time leading to better representative samples.

In order to investigate the Total Survey Error (TSE) in mixed-mode survey there is need to separate selection and measurement effects. However, how to separate selection and measurement effects is not straightforward because they are completely confounded. The literature suggests different methods of disentangling mode effects. For example, Jäckle, Roberts, and Lynn (2010) suggest using response matching based on socio-demographic variables that are closely related with variables of the interest. Vannieuwenhuyze, Loosveldt, and Molenberghs (2010) proposes use of proportions and the mean of a multinomial variable by comparing a mixed-mode dataset with a comparable single-mode dataset. However, this method is only applicable if a comparable dataset is available. Lugtig et al. (2011) proposes the use of propensity score matching to disentangle selection and measurement effects. Propensity score matching entails matching treated to control participants based on the estimated propensity scores. The common matching algorithms include greedy matching, genetic matching, and optimal matching (Guo & Fraser, 2014; Leite, 2017; P R Rosenbaum, 2002).

Data

Community Life Survey (CLS) mixed-mode experiment data collected during July–September 2014 by Kantar Public in England were analysed in this study. In CLS survey, the respondents were asked questions on issues that are key to encouraging social action and empowering communities such as volunteering, donating, community engagement, civil duty and well-being.

For the study, three samples considered include Face-to-Face survey, online follow up survey and online Addressed Based Online Surveying (ABOS).

Face to Face Survey

The multi-stage random sample design was employed for the face-to-face CLS. Face-to-face survey involved drawing a stratified random sample of postal sectors found in England. Postal sectors were used as Primary Sampling Units (PSU's) where each PSU was supposed to have a minimum of 500 addresses. PSU's with less than 500 addresses were combined. The sampling error was minimised by sorting the PSU's list based on ethnic mix profile, regions and estimated survey prevalence of individuals based on 2001 census. This ensured that the sample drawn was representative. Then, a systematic random sample of 24 addresses from within each sampled postal sector based on latest edition of the Postcode Address File was drawn. All addresses within each postcode were listed alphanumerically before the sample was drawn to ensure maximum geographical dispersion. Although, all residential addresses in England had an equal probability of selection the total sampling probability of adults (+16 years) varied due to within-address sampling fractions. Interviewers visited the sampled addresses to establish the occupants' residential status and the number of dwelling units at the address. Where the number of dwelling units were greater than one, the interviewer used a random number generator to sample one. In addition, the same random number generator was used to sample one adult for interviewer from those residents at the targeted dwelling unit. To make sure the data quality was good, each interviewer was involved in selecting and verifying the respondent at any sampled address. The total number of respondents interviewed using face-to-face was 666.

Online Follow up Survey

The online follow up survey was conducted to respondents who had previously participated in the main face-to-face interviewer of the Community Life Survey version of 2013-14 and had given consent to be re-contacted. The total number of interviews for main CLS 2013-14 were 5,105, and out of these 4,219 (83%) were invited to participate for online follow up survey. In total 1,576 (37%) of those invited did so with 1,415 (89.8%) using web and 161(10.2%) using postal who were excluded in the final analysis. This was because the main interest is to evaluate mode effects between face-to-face and online surveys.

Online Survey based on Addressed based online Surveying (ABOS)

Addressed based online Surveying (ABOS) design involves drawing a stratified random sample of addresses from the Royal Mail's Residential Postcode Address File (PAF) which includes more than 99% of all residential addresses in the UK (Williams, 2017a). Each address was sampled with equal probability and, at each address, up to a maximum of four individuals were considered. The intention of allowing more than one individual from the same household to participate in a survey was introduced in an attempt 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 a one respondent in 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. Kantar Public has experimented with ABOS design with an aim of improving key design features that will lead to better response rates and sample representativeness (Grant & Williams, 2017) . After drawing the sampled addresses, invitation letters containing username(s), password(s) plus the survey website url are sent to occupant(s) inviting the resident adults (s) to complete the survey online. The ABOS design also has paper option for the population not covered by the internet. Generally, the response rate for the ABOS design ranges between 7% to 25% and conditional incentives are offered with aim of increasing response rates. Williams (2017) found ABOS to have similar profile to dual-frame to Random-Digit Dialling (RDD) design and a less accurate profile compared to face-to-face. ABOS online response rates were found to be lower in deprived areas and areas with multi-household addresses. However, the ABOS design does not have control over which household in multi-household addresses will be samples. The number of completed interviewer in this online (ABOS) interview was 834 with 789 (94.6%) using web and 48(5.4%) using postal which were excluded for the final analysis.

Each survey contained socio-demographic and area characteristics of the respondents. These variables are used as baseline covariates for the propensity score matching since they have direct effect on the probability of treatment assignment and are related to the outcome (Brookhart et al., 2007). Response outcome for the propensity score model are the survey modes. From each survey, all non-demographic variables that were asked for all respondents were used for comparisons between different modes. Each question was first transformed into a set of binary categorical variables and the absolute percentage difference (APD) between the proportions in each category between two modes were calculated (Sturgis et al., 2017). Next we proceed to methodology section detailing the modelling approaches used in this paper.

Methodology

In this study, propensity score matching that entails matching treated and control observations based on a set of baseline respondent covariates is used (Imbens, 2004; P R Rosenbaum & Rubin, 1983). The aim of propensity score matching is to generate a matched sample such that for every respondent in one survey mode there is at least one matched respondent from the other survey mode with similar resemblance. The propensity score summarises the conditional probability to be a respondent in the face-to-face sample, online (follow up) sample or online (ABOS) sample. One of the methodological issues around the use of propensity scores is the use of survey weights (Austin et al., 2018; Dugoff, Schuler, & Stuart, 2008; Lenis & Stuart, 2017; Ridgeway, Kovalchik, Griffin, & Kabeto, 2015). Ridgeway et al. (2015) recommends that sampling weights should be incorporated in propensity score models as survey weights since they improve covariate balance of the matched sample. However, Lenis and Stuart (2017) found that whether or not survey weights are incorporated in the propensity score models does not have an impact on the covariate balance and estimation of treatment effects. Austin, Jembere, and Chiu (2018) compared three different formulations of propensity score models on whether or not the sampling weights are incorporated in propensity scores as survey weights or covariate and were inconclusive with respect to which method should be used. Therefore, there is no clear way of incorporating sampling weights in propensity score models with complex surveys. Regarding estimation of the propensity score in this study, all the three alternatives used by Lenis and Stuart (2017) and Austin, Jembere, and Chiu

(2018) are considered for comparative purposes. The three propensity score models considered are: (1) incorporate socio-demographic variables as covariate without weights (Unweighted model), (2) incorporate survey weights in a weighted estimation (weighted model), and (3) incorporate the survey weights as a covariate in the estimation of the propensity score model (unweighted model with weight as covariate). In addition, this paper will evaluate three different ways of specifying sampling weights for the matched control units when estimating treatment effects. It is important to note that in propensity score matching two types of matched control groups may be created depending on the sampling weights used in the outcome analysis: (1) no weights on the outcome analysis, (2) the population of control units that resemble the treated units, and (3) the population of treated units. For example, when matched control units retain their sampling weights the population of control units resemble that of the treated units. However, Lenis and Stuart (2017) recommends that matched control units should inherit the weights of the treated units to which they were matched to because it improves the performance of estimators under certain non-response mechanisms. On the other hand, Austin, Jembere, and Chiu (2018) recommends that the matched control units should retain their natural weights since they lead to a decreased bias in outcome analysis. Therefore, this study considered three different propensity scores methods specification and three different analytic approaches for outcome analysis.

The propensity scores are estimated using logistic regression (Agresti, 2013). The logistic regression for estimating propensity scores takes the following form. Let y_i denote the binary outcome (i.e. survey modes assigned to survey participants) for respondent i $(i = 1, \dots, n)$ defined as

$$y_i = \begin{cases} 1 & \text{Mode A} \\ 0 & \text{Mode B} \end{cases} \qquad (1)$$

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 standard logistic regression model that accounts for interviewer effects takes the form

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j = \mathrm{B}^T X_j \qquad (2)$$

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 respondent level. The choice of socio-demographic and area variables as covariates in the propensity score model is informed by the existing literature (Brookhart et al., 2007; Cuong, 2013). The final propensity score model contains variables that are significant in univariate analysis based on 5% significant level as proposed by Hirano and Imbens (2001). The true propensity score model is a balancing score and its adequacy is assessed by checking the area of common support (Austin, 2011; Leite, 2017). This is the area of the distribution of the propensity scores where values exists for both treated and control units and is evaluated using histograms and boxplots.

Once the adequacy of the area of common support is satisfied, propensity score matching is implemented using optimal matching (P R Rosenbaum, 2002). Optimal matching was proposed by Rosenbaum (1989) and uses network flow optimisation to produce matches with minimum global distance. In this study optimal matching is implemented using MatchIt package in R (Austin, 2011; Ho, Imai, King, & Stuart, 2009). The matched sample is obtained using one-to-one optimal matching without replacement which guarantees the minimum distances between matched pairs (Gu & Rosenbaum, 1993). The next step after propensity score matching involves assessing the quality of the matched samples using covariate balance. This shows the similarity of the distributions of all covariates in the matched treated and control units. The covariate balance is defined in terms of absolute standardised mean differences (SMD) (Rubin, 2001). According to Rubin (2001) , SMD of less than 0.25 indicate that adequate covariate balance has been achieved. However, this paper is going to adopt a stricter criterion of assessing covariate balance in which SMD should be less than 0.10 as proposed by Austin (2011. In addition, chi-square test and p-values that incorporate information of the sample size will be used as a measure of the balance as recommended by Hansen (2008). Once the matched sample satisfies adequate covariate balance, the next step involves estimating selection and measurement effects.

Estimation of selection and measurement effects

The selection and measurement effects are evaluated using Absolute Percentage Differences (APD)). This analysis procedure is adopted from Sturgis et al. (2017). The APD estimates are preferred to assess selection and measurement effect because they are naturally interpretable compared to other measures such as the differences in the means and standardised differences (Sturgis et al., 2017). APD estimates treat all percentage differences as equivalent irrespective of the size of the discrepancy relative to the sample considered. For example, an APD estimate obtained by differencing proportions of a categorical level between two survey modes at 10% and 5% has the same APD as the difference between 75% in face to face and 70% , despite the first discrepancy being larger in relative terms than the second. The APD are estimated from attitudinal and behavioural variables contained in the three samples of CLS before and after matching. Let assume we have two survey modes A and B. Then APD are computed by differencing the percentage proportions in each category for survey mode A and the proportion survey mode B. 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. Only those categorical levels with proportions between 5% and 95% to the total sample of the selected variables were considered for the final computation of the APD. This aims to remove those categorical levels that may have undue influence on the selection and measurement effects. The APD are computed as follows: Let $\hat{\pi}_{ijA}$ and

$\hat{\pi}_{ijB}$ denote the estimated percentage proportions for question j and categorical level i for survey modes A and B respectively. Then, an APD estimate denoted as y_{ij} is defined $y_{ij} = \left|\hat{\pi}_{ijA} - \hat{\pi}_{ijB}\right|$. The APD estimates obtained for different questions before and after matching are presented graphically with an aim of assessing their changes based on different propensity score model specifications. The difference of APD estimates before APD and after matching based on different propensity model specifications represent selection effects. The APD estimates obtained after matching represent measurement effects.

The next step involves analysis of the APD estimate before and after matching in a multilevel framework. Generally, measurement effects are caused by the differences in how items are presented to the respondent, primacy and recency effects, interviewer effects and social desirability bias (de Leeuw, 2005; Dillman, Smyth, et al., 2009; O'Muircheartaigh, 1997). Therefore, it becomes necessary to investigate whether any association exists between the APD estimates and questions characteristics. In addition, APD estimates are clustered across questions and it is important to adjust for the dependency induced by the estimates from the same question to understand how much of unexplained variation may be attributed to the questions. The choice of APD estimates to analyse in multilevel modelling for matched sample are those with the highest difference in selection bias (i.e. before – after matching). The first step for multilevel analysis of APD estimates involves normalising them by taking the natural log to deal with any skewness. The multilevel model is specified as a two level model where the log transformed APD estimates are defined at level 1 and within question variable at level 2. Let the response variable y_{ij} be defined as the logged APD $y_{ij} = log\left|\hat{\pi}_{ijA} - \hat{\pi}_{ijB}\right|$. Then the multilevel model accounting for question level takes the form:

$$Y_{ij} = \beta_0 + x'_{ij}\boldsymbol{\alpha} + v_{0j} + e_{ij}$$

where x'_{ij} is a vector of question level characteristics with coefficient vector $\boldsymbol{\alpha}$, v_{0j} is a random intercept and e_{ij} is the error term. The random intercept and error term variances are assumed to follow a normal distribution with zero mean and constant variances: $v_{0j} \sim N(0, \sigma_v^2)$, and $e_{ij} \sim N\left(0, \sigma_{ij}^2\right)$.

Results

Face to Face and Online (Follow up)

Table 1 presents the standardised mean differences (SMD) for matched samples of face-to-face and online (follow up) constructed using optimal matching. The sample size of the face-to-face before and after matching was constant at n=666. The online (follow-up) matched sample size was n=666 (47.1%) indicating that 749 (52.9%) of respondents were lost after matching since the original sample size was n=1, 415. In general, all the three propensity score methods resulted in good balance on the observed baseline covariates since the distance is less 0.10. The method incorporating the survey weights as a covariate in the estimation of propensity score mode resulted in noticeably the best balance than other methods. The SMD for all covariates in the matched samples is less than 0.10 indicating adequate covariate balance.

Table 1: Balance in baseline covariates based on for the three different approaches to propensity score matching models in face-to-face and online (follow up) samples

		Specification of propensity score model		
Variable {Ref}	Categories	Unweighted	weight as covariate	Weighted
Distance		0.025	0.009	0.024
Age	16 to 34 years	0.004	0.025	0.035
	35 to 49 years	0.015	0.007	0.011
	50 to 64 years	0.007	0.014	0.021
	65 to 74 years	0.025	0.004	0.049
	Over 75 years	0.013	0.000	0.004
Race {Others }	White	0.000	0.004	0.004
Number of adults in household {1}	2	0.018	0.024	0.066
	3	0.010	0.043	0.019
	4 or more	0.020	0.047	0.013
Income	0 to < £15K	0.041	0.026	0.005
	£15K to <£40K	0.004	0.030	0.009
	>£40K	0.033	0.009	0.042
Tenure {Private rent}	Mortgaged	0.003	0.010	0.000
	Outright ownership	0.016	0.029	0.010
	Social rent	0.012	0.020	0.000
Education {No qualification }	Other qualification	0.021	0.000	0.003
	Degree or above	0.046	0.031	0.050
GOR {London)	East Midlands	0.028	0.000	0.044
	East of England	0.005	0.023	0.037
	North East	0.045	0.026	0.045
	North West	0.075	0.075	0.089
	South East	0.053	0.018	0.085
	South West	0.043	0.000	0.000
	West Midlands	0.022	0.017	0.022
	Yorkshire and Humberside	0.032	0.005	0.032
Sampling weights		-	0.020	-

The area of the distribution of the propensity scores represented by histograms and boxplots and the distribution of baseline covariates, chi-square test and p-values are presented in the Appendix. Histograms and boxplots show that show there is adequate common support for estimation of measurement effects for face-to-face and online (follow up) samples. In addition, the distribution of baseline covariates, chi-square test and p-values after matching indicates good balance.

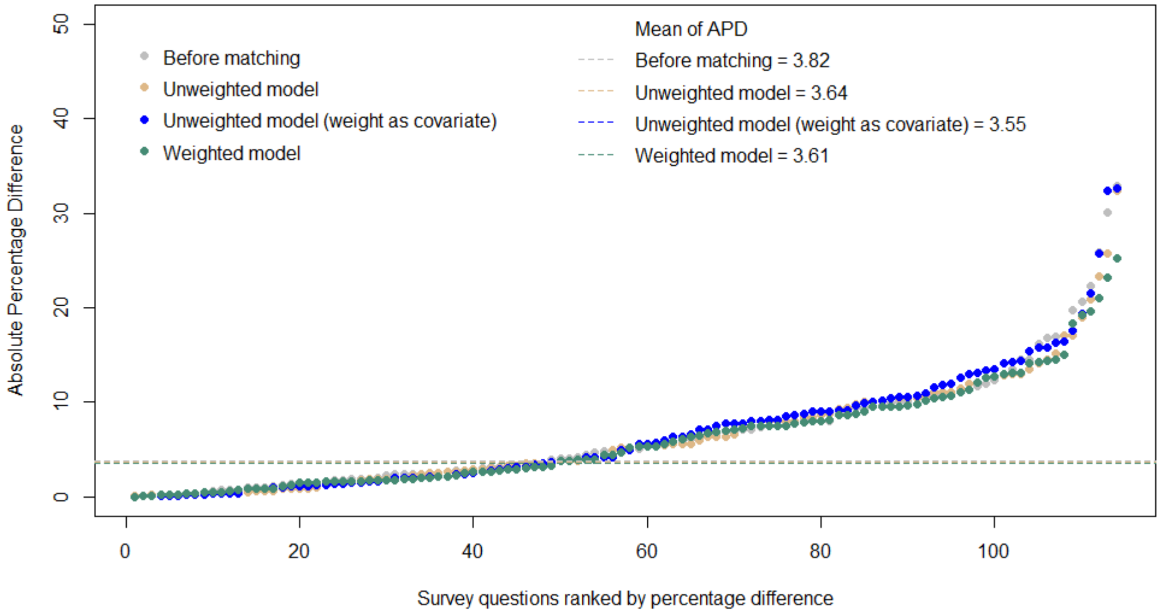

Figure 1: Estimated mode effects (Absolute Percentage Difference) by Question before and after matching for face-to-face and online (Follow up)

Figure 1 summarises the results of APD estimates obtained in four samples of face-to-face and online (follow up). These samples are before matching sample and three matched samples based on three different specifications of propensity score models. The horizontal lines in Figure 1 represent the geometric mean for each of the four samples. The APD patterns observed in Figure 2 are consistent across the four samples. In the context of this analysis, geometric mean has an advantage over arithmetic mean because it is less affected by very small and very large values in skewed data. In addition, geometric mean is recommended when dealing with numbers that are percentages and ratios because it tends to give the correct figure for the final value. The APD estimates obtained based on inherited or natural weights of the matched samples are similar to those obtained without controlling for survey weights as presented in Figure 4. This implies that using either natural or inherited survey weights in the computation of APD estimates does not lead to any changes in APD in the three different propensity score models.

The mean APD estimate before matching is 3.8 and reduces by an average of 0.2 percentage points on the matched samples that represents selection effects. The highest reduction in mean APD estimates of 0.3 percentage points is observed on the matched samples estimated using propensity score model that incorporate survey weights as covariates which also had the best covariate balance. On average the estimated measurement effects for face-to-face and online (follow up) matched samples is 3.6 percentage points which is close to 3.8 percentage points obtained by

Williams (2017b). Table 2 summarises the aggregated mean differences of the four samples for face-to-face and online (follow up). It can be observed that the great reduction in mean APD of about 1.2 percentage points before and after matching is in the 2.6-5.0% classification. This indicate that selection effects are highly pronounced in questions with APD estimates contained in this classification.

Table 2: Aggregate APD estimates for face-to-face vs online (follow up) obtained from four samples.

APD Classifications	Frequency (%)	Mean APD			
		Before Matching	Unweighted	weight as covariate	Weighted
0 -2.5%	34 (29.82)	0.810	1.006	0.972	1.017
2.6-5.0%	22 (19.30)	3.449	2.296	2.064	2.380
6.0-10.0%	32 (28.07)	7.134	6.810	6.609	6.437
11.0-15.0%	16 (14.04)	11.652	11.154	11.848	11.143
16.0-20.0%	5 (4.39)	17.347	13.759	14.516	13.737
>20.0%	5 (4.39)	25.972	23.371	25.271	21.550
Overall	114 (100.0)	3.820	3.641	3.552	3.607

Table 3: Estimated coefficients for Logged Absolute Percentage Difference of face-to-face and online (follow-up) samples (weight as covariate model)

	Model 1		Model 2		Model 3		Model 4	
	β\{exp(β)\}	SE	β\{exp(β)\}	SE	β\{exp(β)\}	SE	β\{exp(β)\}	SE
Intercept	1.308*\{3.683\}	0.090	1.004*\{2.728\}	0.139	1.751*(5.758\}	0.203	1.787*\{5.971\}	0.212
Question Type (Ref :Attitudinal)								
Behavioural					0.271 \{1.312\}	0.283	0.272 \{1.312\}	0.284
Response Categories (Ref: 2 or more)								
Yes/no					-1.449*\{0.235\}	0.289	-1.449*\{0.235\}	0.289
Matching (Ref: Before Matching)								
After Matching							-0.073\{0.930\}	0.124
Random Effects								
Question			1.171	1.082	0.713	0.845	0.711	0.843
ϵ_{ij}	1.839	1.356	0.860	0.927	0.876	0.936	0.879	0.938
AIC	791.974		729.644		711.672		715.663	
ICC			57.652		44.893		44.713	

β =coefficient; exp (β)=ratios of estimated absolute difference; SE=standard error; $*p < 0.05$

Table 3 presents the coefficient estimates, exponents of coefficient estimates representing ratios of APD estimates, and corresponding standard errors for the multilevel analysis of the logged APD estimates for obtained before and after matching (based on PS model 2). Table 2 shows that the average mode effects for the face-to-face and online (follow up) samples across the questions is 3.7 percentage points. The random effects estimates in model 2 that contains no covariates show that question accounts for 58 percent of the total variability in the APD estimates for mode effects. The coefficient estimate for fixed effect of response category variable is significant as shown in model 3. The results show that questions with two response category levels "yes or no" have 80% less mode effects compared to those questions with "2 or more" levels. This suggests that questions with 2 or more categorical levels create difficulty among respondents when choosing the answers, In addition, such questions are likely to be affected by primacy and recency effects, recall bias, and acquiescence . The question type variable (i.e. attitudinal or behavioural) is not significant. Controlling for both question type and response categories reduces the total variability explained by question to 45% representing 13% reduction. Lastly, model 4 shows that APD obtained before and after matching are not statistically significant. This indicates that selection effects present in face-to-face and online (follow up) samples may not be problem in the overall estimation of measurement effects. However, in the evaluation of total survey framework the APD estimates obtained in the matched sample are unbiased estimates of measurement effects.

Face to Face and Online (ABOS)

Table 4: Balance in baseline covariates based on for the three different approaches to propensity score matching models in face-to-face and online (ABOS) samples

		Specification of propensity score model		
Variable {Ref}	Categories	Unweighted	weight as covariate	Weighted
Distance		0.505	0.506	0.496
Age	16 to 34 years	0.035	0.035	0.032
	35 to 49 years	0.070	0.062	0.070
	50 to 64 years	0.045	0.052	0.041
	65 to 74 years	0.065	0.065	0.065
	Over 75 years	0.164	0.164	0.164
Race {Others}	White	0.098	0.116	0.107
Number of adults in household {1}	2	0.231	0.237	0.243
	3	0.097	0.092	0.072
	4 or more	0.074	0.067	0.081
Income	0 to < £15K	0.077	0.077	0.072
	£15K to <£40K	0.064	0.064	0.064
	>£40K	0.121	0.121	0.127
Education {No qualification }	Other qualification	0.030	0.039	0.036
	Degree or above	0.187	0.183	0.183
GOR {London}	East Midlands	0.017	0.006	0.006
	East of England	0.046	0.046	0.028
	North East	0.064	0.064	0.070
	North West	0.049	0.062	0.035
	South East	0.142	0.160	0.165
	South West	0.059	0.032	0.043
	West Midlands	0.122	0.122	0.113
	Yorkshire and Humberside	0.000	0.005	0.005
Number of children {0}	1	0.009	0.005	0.014
	2	0.005	0.019	0.005
	3 or more	0.120	0.120	0.120
Sampling weights		-	0.266	-

Table 4 presents SMD for the socio-demographic and geographical office region variables under the three different propensity score models for face-to-face and online (ABOS). Table 4 shows that covariate balance on matched samples based on the three specifications of propensity score models is not adequate. The SMD obtained for global distance of the covariates used in the three different propensity models is larger at 0.50 than the threshold value of 0.10. This suggests that the matching procedure is not effective in balancing the baseline covariates between the face-to-face and online (ABOS) samples. The area of the distribution of the propensity scores represented by histograms and boxplots and the distribution of baseline covariates, chi-square test and p-values are presented in the Appendix. These measures show that common support between face-face and online (ABOS) samples to be inadequate for one-to-one optimal matching. The distribution of baseline covariates, chi-square test and p-values after matching using the propensity scores for the model with weights as covariates indicates that balance is not achieved. Therefore, the measurement effects based on matched sample for face-to-face and online (ABOS) samples are not estimated due to lack of common support. It is also important to note that lack of common support between the face-to-face and online (ABOS) samples may be a sign of representativeness issues(Bryson, Dorsett, & Purdon, 2002).

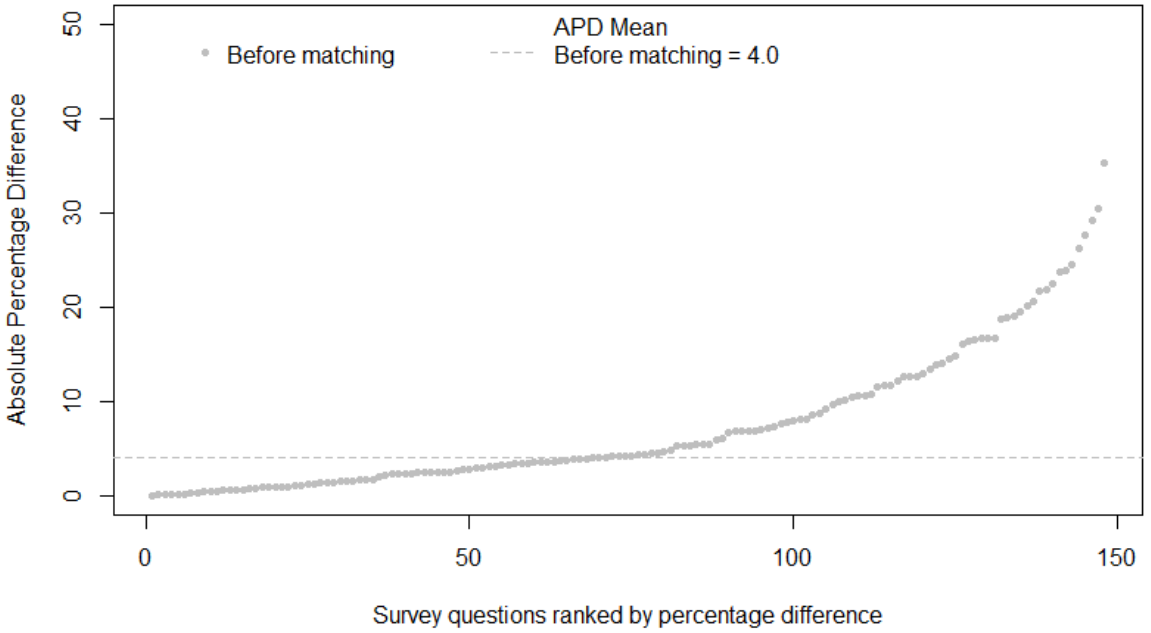

Figure 2: Estimated mode effects (Absolute Percentage Difference) by Question face-to-face and for online (ABOS) samples before matching.

Figure 2 presents the APD estimates results between face-to-face and online (ABOS) before matching. Figure 2 shows that the average the mode effects for face-to-face and online (ABOS) is 4.0 percentage points that is close to 3.8 percentage points obtained for face-to-face and online (follow up) samples. The small difference in APD of only 0.2 percentage points between the face-to-face and online (ABOS) and face-to-face vs online (follow up) indicates that online (ABOS) may be useful when the knowledge of basic population is known. In addition, this may imply that online (ABOS) sample is suited to be used in studies where representativeness is not required. Considering the differences observed between face-to-face vs online (follow up) samples and face-to-face vs online (ABOS) samples it is important to investigate whether any differences exists between the two online samples.

Online (ABOS) and Online (Follow up)

Table 5: Balance in baseline covariates based on for the three different approaches to propensity score matching models in online (ABOS) and online (follow up) samples

		Specification of propensity score model		
Variable {Ref}	Categories	Unweighted	weight as covariate	Weighted
Distance		0.024	0.044	0.034
Age	16 to 34 years	0.015	0.025	0.025
	35 to 49 years	0.009	0.012	0.009
	50 to 64 years	0.003	0.006	0.000
	65 to 74 years	0.017	0.043	0.030
	Over 75 years	0.010	0.005	0.020
Number of adults in household {1}	2	0.016	0.117	0.029
	3	0.029	0.085	0.022
	4 or more	0.016	0.120	0.004
Number of children (0)	1	0.056	0.020	0.016
	2	0.000	0.017	0.012
	3 or more	0.010	0.031	0.031
GOR {London)	East Midlands	0.051	0.000	0.102
	East of England	0.009	0.039	0.043
	North East	0.040	0.053	0.040
	North West	0.041	0.021	0.021
	South East	0.003	0.041	0.006
	South West	0.016	0.041	0.024
	West Midlands	0.078	0.046	0.064
	Yorkshire and Humberside	0.000	0.054	0.004
Sampling weights		-	0.097	-

Table 5 presents the SMD for matched samples of online (ABOS) and online (follow up). The matched samples contains 781 respondents. The SMD in the matched samples based on three different propensity score models are less than 0.10 indicating that online (ABOS) and online (follow up) respondents are adequately balanced. The unweighted propensity score model had the minimum global distance at 0.024among the three different models indicating the best balance. The histograms and boxplots presented in the Appendix show that there is an adequate common support of propensity scores obtained for online (ABOS) and online (follow up) for effective matching. The distribution of baseline covariates as presented in the Appendix indicate that covariate balance is attained after matching for online (ABOS) and online (follow up)

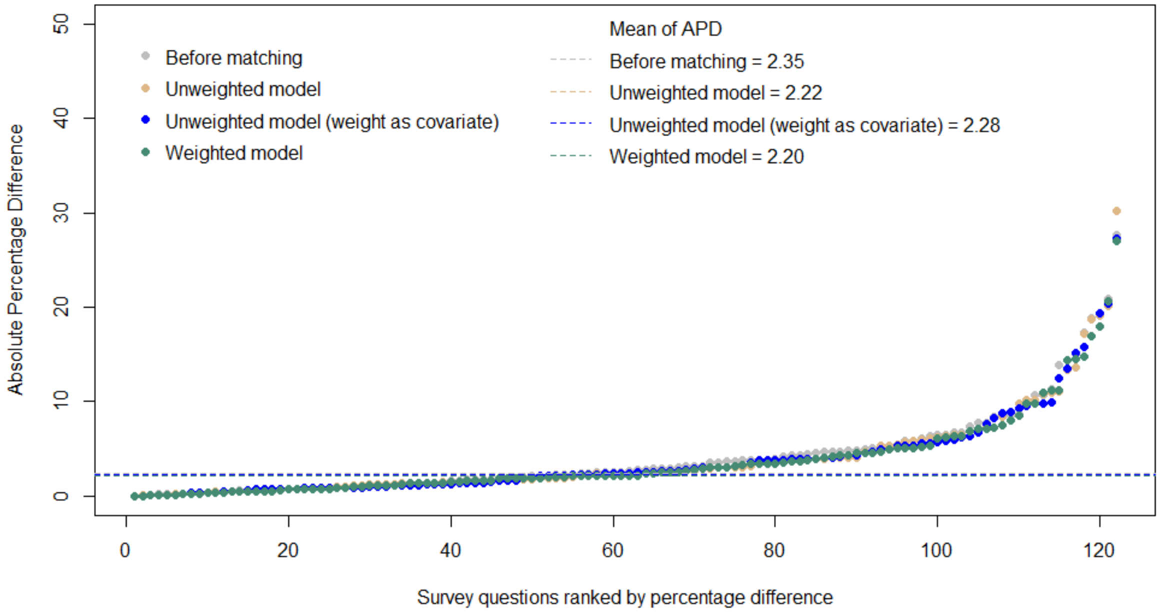

Figure 4: Estimated mode effects (Absolute Percentage Difference) by Question before and after matching for online (ABOS) and online (Follow up).

Figure 4 summarises the results of the APD estimates obtained in four different samples (i.e. before matching sample and three different matched samples based on the propensity score model specifications. The mean APD estimate before matching is 2.4 and reduces by an average of 0.1 percentage point after matching. The matched sample based on the weighted propensity score model has the highest reduction in mean APD estimates at 0.2 percentage points representing selection effects. The mean APD estimate of 2.2 after matching based on weighted model is supposed to be measurement effect. Taking into account that both online (ABOS) and online (follow up) samples are based on same mode, then these measurement effects may be attributed to d devices used by respondents in online surveying. In addition, the use of natural or inherited weights on the control respondents resulted in APD estimates similar to those presented in Figure 4 based on the three different propensity score model specifications.

Table 6 presents the aggregated mean differences classified into 6 groups. It can be observed that almost 75% of the questions had APD estimate less than 5 percentage points. This shows that measurement effects for online (ABOS) and online (follow up) are modest compared to face-to-face vs online samples because the influence of the mode on the answers respondents give is low.

Table 6: Aggregate APD estimates for online (ABOS) vs online (follow up) obtained from four samples.

APD Classifications	Frequency (%)	Mean APD			
		Before Matching	Unweighted	weight as covariate	Weighted
0 -2.5%	58 (48.74)	0.901	1.084	1.174	1.164
2.6-5.0%	30 (25.20)	3.701	2.728	2.602	2.568
6.0-10.0%	19 (15.97)	6.645	4.676	4.597	4.022
11.0-15.0%	6 (14.04)	11.774	9.034	9.393	8.633
16.0-20.0%	4 (5.04)	17.543	17.027	16.702	16.997
>20.0%	2 (1.38)	24.077	24.629	23.035	21.407
Overall	119 (100.0)	2.352	2.222	2.278	2.201

Table 3

Table 7: Estimated coefficients for Logged Absolute Percentage Difference of online (ABOS) and online (follow-up) samples (weight as covariate model)

	Model 1		Model 2		Model 3		Model 4	
	$\beta\{\exp(\beta)\}$	SE	$\beta\{\exp(\beta)\}$	SE	$\beta\{\exp(\beta)\}$	SE	$\beta\{\exp(\beta)\}$	SE
Intercept	0.797*{2.218}	0.087	0.577*{1.781}	0.220	1.148*{3.152}	0.169	1.198*{3.311}	0.193
Question Type (Ref :Attitudinal)								
Behavioural					-0.007 {0.992}	0.242	-0.008 {0.992}	0.260
Response Categories (Ref: 2 or more)								
Yes/no					-0.993*{0.383}	0.244	-0.959*{0.383}	0.265
Matching (Ref: Before Matching)								
After Matching							-0.099 {0.905}	0.105
Random Effects								
Question			0.704	0.839	0.478	0.691	0.478	0.691
$\boldsymbol{\epsilon_{ij}}$	1.483	1.218	0.803	0.895	0.809	0.900	0.811	0.900
AIC	794.794		736.886		725.064		728.810	
ICC			46.731		37.149		37.093	

β =coefficient; exp (β)=ratios of estimated absolute difference; SE=standard error; *$p < 0.05$

Table 7 presents the multilevel analysis for the APD estimates obtained before and after matching based on the weighted propensity score model for online (ABOS) and online (follow up). Table 7 shows that the average APD of the online (ABOS) and online (follow up) is 2.2 percentage points. This difference is significant and represents the estimated measurement effects that exist between the two online samples. Turning to the random effects estimates, model 2 containing no covariates shows that question contributes 47 percent of the total variability in the APD estimates. The results for fixed effects in models 3 show that questions with two-response categories "yes/no" have 62%

less measurement effects compared to questions with "2 or more response categories". This shows that response categories are significantly associated with measurement effects. It is also important to note that using either online (ABOS) or online (follow up) will lead to a reduction of measurement effects attributed to response category " 2 or more categories" by 18% compared to face-to-face vs online samples that is at 80%. Similar to face-to-face vs online (follow up), the question type (i.e. behavioural or attitudinal) is not significant. The total variability explained by question reduces by 10% after controlling for fixed effects. The matching indicator (i.e. before and after matching) in model 4 is not significant indicating occurrence of selection effects between the online (ABOS) and online (follow up) may not be problem in estimation of measurement effects.

Discussion

In this paper we explore different ways of formulating propensity score models to disentangle mode effects (i.e. selection effect and measurement effects) within a mixed-mode survey context. The first issue considered was the formulation of the propensity score models based on three different ways of specifying the survey weights. The three different models formulated were: (1) no weights were used at all in the propensity score model, (2) the weights were incorporated as a covariate in the estimation of the propensity score model, and (3) the survey weights were incorporated as weights in a weighted regression analysis. The performance of different methods of using propensity scores with survey weights was assessed based on covariate balance of baseline covariates. We found that none of the three different propensity models resulted in better balance of baseline covariates than other specifications of the propensity score models. These results are consistent with the ones found by Austin, Jembere, and Chiu (2018) and Lenis et al. (2017).

This study also evaluated the use of survey weights in the outcome analysis based on the motivation by contradictory findings by Lenis et al. (2017) and Austin, Jembere, and Chiu (2018). Lenis et al. (2017) recommend use of inherited weights on the matched control units because they help improving the performance of estimators where nonresponse depends on the baseline covariates and treatment assignment. On the other hand, Austin, Jembere, and Chiu (2018) recommend that natural weights should be retained on the matched control units because they lead to decreased bias on treatment effects. The results in this paper found that incorporating either natural or inherited survey weights on the outcome analysis (i.e. estimation of APD estimates) had no effect on the overall APD estimates obtained. The similar APD estimates across the three analytical methods of outcome analyses indicate the robustness of APD in quantifying selection and measurement effects.

The other purpose of this paper is to use propensity score methods to disentangle selection and measurement effects in face-to-face and online surveys. The findings by Lugtig et al. (2011) that propensity score matching could be helpful in disentangling selection and measurement effects motivated this study. The mean APD estimates for the face-to-face and online (follow up) samples after matching indicates that measurement effects are present in this two samples. This results are consistent with findings by Williams (2017b).The existence of measurement effects in between face-to-face and online (follow up) samples may attributed to the sensitive nature of some questions especially in donating and volunteering which my lead to social desirability bias (Bekkers & Wiepking, 2011). It is also important to note that difference in APD estimates before and after matching which represents selection effects was not significant. This implies that the greater part of the differences in face-to-face and online (follow up) is due to measurement effects. This results are also consistent with findings by Williams (2017b) who found that data collection is responsible for the bulk of the mismatch in mixed-mode samples.

The propensity scores for face-to-face and online (ABOS) samples lacked an adequate common support for one-to-one optimal matching. As such, it was not possible to separate selection and measurement effects. However, is should be noted that APD estimates for face-to-face and online (ABOS) before matching was slightly higher than that face-to-face and online (follow up). This promoted comparison of two online samples (ABOS and follow up). When the two online samples are compared, the matched respondents from both sample are similar in terms of baseline covariates but not in the APD estimates. This indicates that measurement effects are present between the two online samples and may be attributed to device effects. This contradicts findings by Lugtig et al. (2011) that found that propensity score matching may explain any differences in matched online samples. However, it is also important to note that almost all respondents in Lugtig et al. (2011) may have completed online survey using personal computers (PC's). However, over the last 5 years respondents are also using smartphones and tablets in completing surveys (Callegaro, 2013). Considering that Lugtig and Toepoel (2016) found that measurement errors are large in tablets and smartphones compared to PC's it is possible the differences in the two online samples (i.e. ABOS and follow up) is due self-selection of respondents to a particular device. Our findings also show that the questions in the survey accounts for about half of the total variability in the APD estimates. The response category variable was significant predictor for APD estimates indicating that measurement effects may originate from the differences in the number of categories presented to the respondent. The type of question whether attitudinal or behavioural was not significant.

The findings from this study have implications for survey practice. The approach used here to disentangle selection and measurement effects could be used as a way to explain differences that occur in mixed-mode design. Propensity score matching offers a cheap and easy alternative of estimating measurement effects in mixed-mode surveys compared to experimental designs that are always sometimes impractical in survey practice. Secondly, we have showed that difference in data collection is responsible for most of measurement effects between face-to-face and online samples. Therefore, it would be important for survey methodologists to consider the hidden price of data quality when switching between modes in the surveys.

While the methodological approach and findings provide an understanding of the use of propensity score matching in disentangling selection and measurement effects in mixed-mode surveys, this study is not without limitation. First, although APD estimates of the matched samples represent measurement effects it is not possible to distinguish whether they are due to interviewer effects and social desirability bias, primacy and recency effects, recall bias, or acquiescence. Although the results presented were estimated based on three different propensity scores it would have be advisable to consider matching approaches such as greedy matching.

References

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

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

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

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.

Bekkers, R., & Wiepking, P. (2011). Accuracy of self-reports on donations to charitable organizations. *Quality and Quantity*, *45*(6), 1369–1383. http://doi.org/10.1007/s11135-010-9341-9

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

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.

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.

Callegaro, M. (2013). Do you know which device your respondent has used to take your online survey. *Survey Practice*, *3*(6), 1–12. Retrieved from http://www.surveypractice.org/index.php/SurveyPractice/article/view/250/pdf

Corritore, C. L., Kracher, B., & Wiedenbeck, S. (2003). On-line trust: Concepts, evolving themes, a model. *International Journal of Human Computer Studies*, *58*(6), 737–758. http://doi.org/10.1016/S1071-5819(03)00041-7

Couper, M. P. (2000). *Handbook of Web Surveys*. *Public Opinion Quarterly* (Vol. 64). http://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. http://doi.org/10.1093/poq/nfr046

Cuong, N. V. (2013). Which covariates should be controlled in propensity score matching ? Evidence from a simulation study. *Statistica Neerlandica*, *67*(2), 169–180. http://doi.org/10.1111/stan.12000 de Leeuw, E. (2005). To Mix or Not to Mix Data Collection Modes in Surveys. *Journal of Official Statistics*, *21*(2), 233–255. http://doi.org/10.4324/9780203843123

Dillman, D. A. (2000). *Mail and Internet Surveys: The Tailored Design Method*. New York: Wiley.

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

Dillman, D. A., & Christian, L. M. (2005). Survey Mode as a Source of Instability in Responses across Surveys. *Field Methods*, *17*(1), 30–52. http://doi.org/10.1177/1525822X04269550

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., & Messer, B. L. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. *Social Science Research*, *38*(1), 1–18. http://doi.org/10.1016/j.ssresearch.2008.03.007

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.

Dodou, D., & De Winter, J. C. F. (2014). Social desirability is the same in offline, online, and paper surveys: A meta-analysis. *Computers in Human Behavior*, *36*, 487–495. http://doi.org/10.1016/j.chb.2014.04.005

Dugoff, E. H., Schuler, M., & Stuart, E. A. (2008). Generalizing Observational Study Results : Applying Propensity Score Methods to Complex Surveys, 284–303. http://doi.org/10.1111/1475-6773.12090

Grant, C., & Williams, J. (2017). *The FCA ' s Financial Lives Survey 2017 Technical Report*.

Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms. *Journal of Computational and Graphical Statistics*, *2*, 405–420.

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

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. http://doi.org/10.1002/sim.3208

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. http://doi.org/10.1093/poq/nfn045

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.

Ho, D., Imai, K., King, G., & Stuart, E. (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., & Krosnick, J. A. (2010). Social desirability bias in voter turnout reports. *Public Opinion Quarterly*, *74*(1), 37–67. http://doi.org/10.1093/poq/nfp065

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.

Hoogendoorn, A., & Daalmans, J. (2009). Nonresponse in the Recruitment of an Internet Panel Based on Probability Sampling. *Survey Research Methods*, *3*(2), 59–72. Retrieved from http://w4.ub.uni-konstanz.de/srm/article/viewArticle/1551

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., Roberts, C., & Lynn, P. (2010). Assessing the Effect of Data Collection Mode on Measurement. *International Statistical Review*, *78*(1), 3–20.

Kreuter, F. (2013). Facing the Nonresponse Challenge. *Annals of the American Academy of Political and Social Science*, *645*(1), 23–35. http://doi.org/10.1177/0002716212456815

Krosnick, J. A. (1991). Response stratergies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213–236.

Lee, Z., & Sargeant, A. (2011). Dealing with social desirability bias: an application to charitable giving. *European Journal of Marketing*, *45*(5), 703–719. http://doi.org/10.1108/03090561111119994

Leite, W. (2017). *Practical Propensity Score Methods Using R*. Carlifornia: SAGE Publications.

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). http://doi.org/10.1093/biostatistics/kxx063

Lenis, D., & Stuart, E. A. (2017). IT ’ S ALL ABOUT BALANCE : PROPENSITY SCORE MATCHING IN THE CONTEXT OF COMPLEX SURVEY DATA, (February).

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. http://doi.org/10.1177/0894439315574248

Lyberg, L., & 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.

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. http://doi.org/10.2105/AJPH.92.2.294

O'Muircheartaigh, C. (1997). Measurement error in surveys: A historical perspective. In L. Lyberg, P. Beimer, M. Collins, E. de Leeuw, C. S. Dippo, S. N, & T. D (Eds.), *In Survey Measurement and Process Quality* (pp. 1–25). New York: John Wiley & Sons.

Ridgeway, G., Kovalchik, S. A., Griffin, B. A., & Kabeto, M. U. (2015). Propensity Score Analysis with Survey Weighted Data, *3*(2), 237–249. http://doi.org/10.1515/jci-2014-0039

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/

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. http://doi.org/10.1093/biomet/70.1.41

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology*, *2*, 169–188. http://doi.org/10.1017/CBO9780511810725.030

Silver, B. D., Anderson, B. A., & Abramson, P. R. (1986). Who Overreports Voting ? *American Political Science Association*, *80*(2), 613–624.

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. http://doi.org/10.1093/poq/nfw055

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. http://doi.org/10.1016/j.wep.2013.10.001

Vannieuwenhuyze, J., Loosveldt, G., & Molenberghs, G. (2010). A method for evaluating mode effects in mixed-mode surveys. *Public Opinion Quarterly*, *74*(5), 1027–1045. http://doi.org/10.1093/poq/nfq059

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. http://doi.org/10.1093/poq/nfn045

Williams, J. (2017a). *Address Based Online Surveying (ABOS) What is ABOS ?* Retrieved from ADDRESS BASED ONLINE SURVEYING (ABOS) What is ABOS ?

Williams, J. (2017b). *Community Life Survey Disentangling sample and mode effects*. Retrieved from Community Life Survey Disentangling sample and mode effects

Appendix

Face to Face and Online (Follow up)

Table A1: SMD for baseline covariates for face-to-face and online (follow up) samples before and after matching (weight as covariates model)

Variable {Ref}	Categories	Before Matching			After Matching (weight as Covariate model)		
		Face-to-face	Online follow up	Tests	Face-to- Face	Online follow up	Tests
		Freq (%)	Freq (%)	P-value (SMD)	Freq (%)	Freq (%)	P-value (SMD)
Age	16 to 34 years	156 (23.4)	226 (16.0)	**0.001** (0.273)	156 (23.4)	157 (23.6)	0.990 (0.030)
	35 to 49 years	142 (21.3)	387 (27.4)		142 (21.3)	146 (21.9)	
	50 to 64 years	168 (25.2)	415 (29.4)		168 (25.2)	166 (24.9)	
	65 to 74 years	107 (16.1)	255 (18.1)		107 (16.1)	101 (15.2)	
	Over 75 years	93 (14.0)	127 (9.0)		93 (14.0)	96 (14.4)	
Race {Others}	White	579 (86.9)	1297 (92.0)	**0.001** (0.165)	287 (43.1)	280 (42.0)	0.740 (0.021)
Number of adults in household	1	228 (34.2)	349 (24.8)	**0.001** (0.218)	228 (34.2)	227 (34.1)	0.975 (0.026)
	2	331 (49.7)	817 (57.9)		331 (49.7)	337 (50.6)	
	3	72 (10.8)	149 (10.6)		72 (10.8)	70 (10.5)	
	4 or more	35 (5.3)	95 (6.7)		35 (5.3)	32 (4.8)	
Income	0 to < £15K	302 (45.3)	596 (42.3)	**0.001** (0.158)	302 (45.3)	291 (43.7)	0.866 (0.047)
	£15K to <£40K	206 (30.9)	529 (37.5)		206 (30.9)	210 (31.5)	
	>£40K	62 (9.3)	163 (11.6)		62 (9.3)	70 (10.5)	
	No data	96 (14.4)	122 (8.7)		96 (14.4)	95 (14.3)	
Education	No Qualifications	255 (38.3)	380 (27.0)	**0.001** (0.269)	255 (38.3)	250 (37.5)	0.712 (0.045)
	Other Qualifications	284 (42.6)	645 (45.7)		284 (42.6)	277 (41.6)	
	Degree or above	127 (19.1)	385 (27.3)		127 (19.1)	139 (20.9)	
GOR	London	90 (13.5)	142 (10.1)	**0.007** (0.218)	90 (13.5)	100 (15.0)	0.754 (0.123)
	East Midlands	53 (8.0)	103 (7.3)		53 (8.0)	58 (8.7)	
	East of England	81 (12.2)	165 (11.7)		81 (12.2)	82 (12.3)	
	North East	39 (5.9)	76 (5.4)		39 (5.9)	32 (4.8)	
	North West	88 (13.2)	197 (14.0)		88 (13.2)	71 (10.7)	
	South East	87 (13.1)	266 (18.9)		87 (13.1)	99 (14.9)	
	South West	56 (8.4)	154 (10.9)		56 (8.4)	64 (9.6)	
	West Midlands	92 (13.8)	154 (10.9)		92 (13.8)	87 (13.1)	
	Yorkshire and Humberside	80 (12.0)	153 (10.9)		80 (12.0)	73 (11.0)	
Number of children	0	491 (73.7)	1014 (71.9)	0.755 (0.052)	491 (73.7)	488 (73.3)	0.733 (0.062)
	1	76 (11.4)	184 (13.0)		76 (11.4)	84 (12.6)	
	2	71 (10.7)	151 (10.7)		71 (10.7)	62 (9.3)	
	3 or more	28 (4.2)	61 (4.3)		28 (4.2)	32 (4.8)	
Paid work {No}	Yes	339 (50.9)	781 (55.4)	0.062 (0.090)	339 (50.9)	350 (52.6)	0.583 (0.033)
Tenure	private rent	150 (22.5)	243 (17.2)	**0.001** (0.267)	150 (22.5)	149 (22.4)	0.992 (0.017)
	Mortgaged	172 (25.8)	462 (32.8)		172 (25.8)	171 (25.7)	
	Outright ownership	226 (33.9)	551 (39.1)		226 (33.9)	231 (34.7)	
	Social rent	118 (17.7)	154 (10.9)		118 (17.7)	115 (17.3)	
Language {Other}	English	627 (94.1)	1358 (96.3)	**0.033** (0.102)	627 (94.1)	627 (94.1)	1.000 (0.001)
Gender {Female}	Male	287 (43.1)	615 (43.6)	0.859 (0.011)	287 (43.1)	280 (42.0)	0.740 (0.021)
Marital Status {married}	Single	197 (29.6)	308 (21.8)	**0.001** (0.178)	197 (29.6)	179 (26.9)	0.301 (0.060)

Before matching: face-to-face = 666 and Online (follow up) =1,410 respondnets

After matchin: face-to-face = 666 and Online (follow up) = 666 respondnets

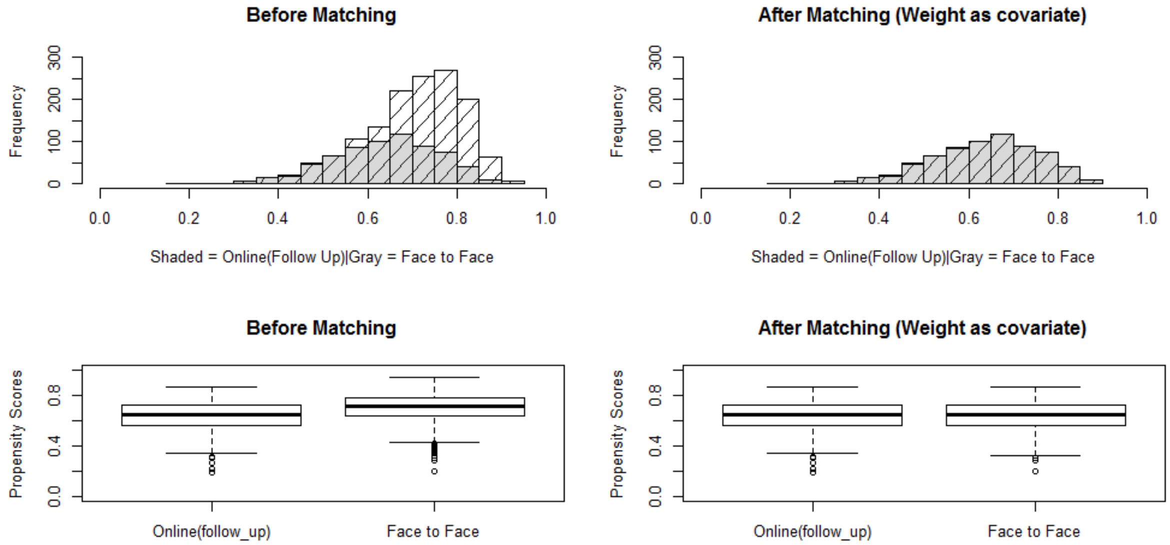

Figure A1: Histograms (top panel) and boxplots (lower panel) of the distributions of propensity scores of face-to-face and online (follow up) samples before and after matching.

Face to Face and Online (ABOS)

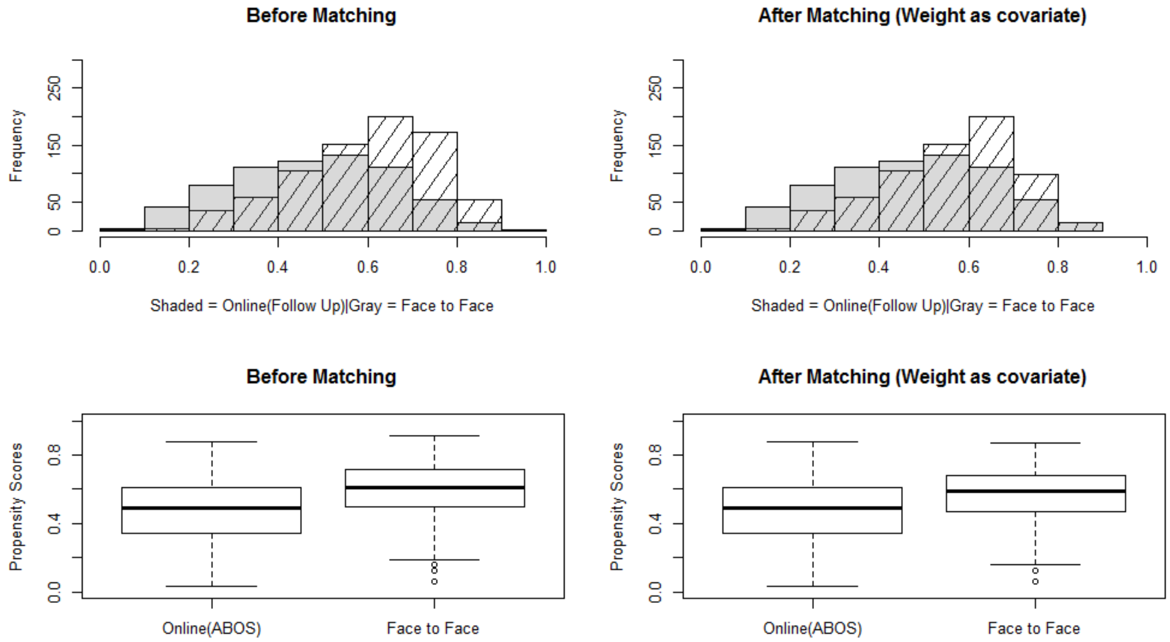

Figure A2: Histograms (top panel) and boxplots (lower panel) of the distributions of propensity scores of face-to-face and online (ABOS) samples before and after matching.

Table A2: SMD for baseline covariates for face-to-face and online (ABOS) samples before and after matching (weight as covariates model)

Variable {Ref}	Categories	Before Matching			After Matching (weight as Covariate model)		
		Face-to-face	Online (ABOS)	Tests	Face-to- Face	Online (ABOS)	Tests
		Freq (%)	Freq (%)	P-value (SMD)	Freq (%)	Freq (%)	P-value (SMD)
Age	16 to 34 years	156 (23.4)	175 (22.4)	**0.001** (0.243)	156 (23.4)	146 (21.9)	**0.012** (0.197)
	35 to 49 years	142 (21.3)	203 (26.0)		142 (21.3)	159 (23.9)	
	50 to 64 years	168 (25.2)	206 (26.4)		168 (25.2)	183 (27.5)	
	65 to 74 years	107 (16.1)	142 (18.2)		107 (16.1)	123 (18.5)	
	Over 75 years	93 (14.0)	127 (9.0)		93 (14.0)	55 (8.3)	
Race {Others}	White	579 (86.9)	712 (91.2)	**0.012** (0.136)	579 (86.9)	605 (90.8)	**0.029** (0.124)
Number of adults in household	1	228 (34.2)	120 (15.4)	**0.001** (0.477)	228 (34.2)	120 (18.0)	**0.001** (0.376)
	2	331 (49.7)	461 (59.0)		331 (49.7)	410 (61.6)	
	3	72 (10.8)	110 (14.1)		72 (10.8)	91 (13.7)	
	4 or more	35 (5.3)	90 (11.5)		35 (5.3)	45 (6.8)	
Income	0 to < £15K	302 (45.3)	305 (39.1)	**0.002** (0.205)	302 (45.3)	262 (39.3)	**0.026** (0.167)
	£15K to <£40K	206 (30.9)	308 (39.4)		206 (30.9)	246 (36.9)	
	>£40K	62 (9.3)	83 (10.6)		62 (9.3)	77 (11.6)	
	No data	96 (14.4)	85 (10.9)		96 (14.4)	81 (12.2)	
Education	No Qualifications	255 (38.3)	199 (25.5)	**0.001** (0.332)	255 (38.3)	194 (29.1)	**0.001** (0.222)
	Other Qualifications	284 (42.6)	341 (43.7)		284 (42.6)	297 (44.6)	
	Degree or above	127 (19.1)	241 (30.9)		127 (19.1)	175 (26.3)	
GOR	London	90 (13.5)	85 (10.9)	**0.001** (0.308)	90 (13.5)	82 (12.3)	0.058 (0.214)
	East Midlands	53 (8.0)	65 (8.3)		53 (8.0)	54 (8.1)	
	East of England	81 (12.2)	77 (9.9)		81 (12.2)	71 (10.7)	
	North East	39 (5.9)	30 (3.8)		39 (5.9)	29 (4.4)	
	North West	88 (13.2)	128 (16.4)		88 (13.2)	102 (15.3)	
	South East	87 (13.1)	160 (20.5)		87 (13.1)	123 (18.5)	
	South West	56 (8.4)	87 (11.1)		56 (8.4)	62 (9.3)	
	West Midlands	92 (13.8)	67 (8.6)		92 (13.8)	64 (9.6)	
	Yorkshire and Humberside	80 (12.0)	82 (10.5)		80 (12.0)	79 (11.9)	
Number of children	0	491 (73.7)	594 (76.1)	**0.023** (0.160)	491 (73.7)	504 (75.7)	0.083 (0.142)
	1	76 (11.4)	91 (11.7)		76 (11.4)	75 (11.3)	
	2	71 (10.7)	84 (10.8)		71 (10.7)	75 (11.3)	
	3 or more	28 (4.2)	12 (1.5)		28 (4.2)	12 (1.8)	
Paid work {No}	Yes	339 (50.9)	443 (56.7)	**0.031** (0.117)	339 (50.9)	368 (55.3)	0.124 (0.087)
Tenure	private rent	150 (22.5)	176 (22.5)	**0.001** (0.271)	150 (22.5)	147 (22.1)	**0.001** (0.277)
	Mortgaged	172 (25.8)	238 (30.5)		172 (25.8)	199 (29.9)	
	Outright ownership	226 (33.9)	298 (38.2)		226 (33.9)	262 (39.3)	
	Social rent	118 (17.7)	69 (8.8)		118 (17.7)	58 (8.7)	
Language {Other}	English	627 (94.1)	758 (97.1)	**0.009** (0.142)	627 (94.1)	645 (96.8)	**0.025** (0.131)
Gender {Female}	Male	287 (43.1)	371 (47.5)	0.104 (0.089)	287 (43.1)	320 (48.0)	0.078 (0.100)
Marital Status {married}	Single	197 (29.6)	197 (25.2)	0.073 (0.098)	197 (29.6)	165 (24.8)	0.056 (0.108)

Before matching: face-to-face = 666 and Online (ABOS) =781 respondnets

After matchin: face-to-face = 666 and Online (ABOS) = 666 respondnets

Online (ABOS) and Online (follow up)

Table A3: SMD for baseline covariates for online (ABOS) and online (follow up) samples before and after matching (weighted model)

Variable {Ref}	Categories	Before Matching			After Matching (weight as covariate model)		
		Online (ABOS)	Online follow up	Tests	Online (ABOS)	Online follow up	Tests
		Freq (%)	Freq (%)	P-value (SMD)	Freq (%)	Freq (%)	P-value (SMD)
Age	16 to 34 years	175 (22.4)	226 (16.0)	**0.004** (0.174)	175 (22.4)	167 (21.4)	0.959(0.040)
	35 to 49 years	203 (26.0)	387 (27.4)		203 (26.0)	206 (26.4)	
	50 to 64 years	206 (26.4)	415 (29.4)		206 (26.4)	206 (26.4)	
	65 to 74 years	142 (18.2)	255 (18.1)		142 (18.2)	151 (19.3)	
	Over 75 years	55 (7.0)	127 (9.0)		55 (7.0)	51 (6.5)	
Race {Others}	White	712 (91.2)	1297 (92.0)	0.558 (0.030)	712 (91.2)	708 (90.7)	0.792 (0.018)
Number of adults in household	1	120 (15.4)	349 (24.8)	**0.001** (0.284)	120 (15.4)	126 (16.1)	0.931 (0.034)
	2	461 (59.0)	817 (57.9)		461 (59.0)	450 (57.6)	
	3	110 (14.1)	149 (10.6)		110 (14.1)	116 (14.9)	
	4 or more	90 (11.5)	95 (6.7)		90 (11.5)	89 (11.4)	
Income	0 to < £15K	305 (39.1)	596 (42.3)	0.118 (0.097)	305 (39.1)	309 (39.6)	0.367 (0.090)
	£15K to <£40K	308 (39.4)	529 (37.5)		308 (39.4)	306 (39.2)	
	>£40K	83 (10.6)	163 (11.6)		83 (10.6)	98 (12.5)	
	No data	85 (10.9)	122 (8.7)		85 (10.9)	68 (8.7)	
Education	No Qualifications	199 (25.5)	380 (27.0)	0.211(0.078)	199 (25.5)	189 (24.2)	0.841 (0.030)
	Other Qualifications	341 (43.7)	645 (45.7)		341 (43.7)	348 (44.6)	
	Degree or above	241 (30.9)	385 (27.3)		241 (30.9)	244 (31.2)	
GOR	London	85 (10.9)	142 (10.1)	0.231 (0.146)	85 (10.9)	90 (11.5)	0.475 (0.140)
	East Midlands	65 (8.3)	103 (7.3)		65 (8.3)	43 (5.5)	
	East of England	77 (9.9)	165 (11.7)		77 (9.9)	87 (11.1)	
	North East	30 (3.8)	76 (5.4)		30 (3.8)	24 (3.1)	
	North West	128 (16.4)	197 (14.0)		128 (16.4)	122 (15.6)	
	South East	160 (20.5)	266 (18.9)		160 (20.5)	158 (20.2)	
	South West	87 (11.1)	154 (10.9)		87 (11.1)	93 (11.9)	
	West Midlands	67 (8.6)	154 (10.9)		67 (8.6)	81 (10.4)	
	Yorkshire and Humberside	82 (10.5)	153 (10.9)		82 (10.5)	83 (10.6)	
Number of children	0	594 (76.1)	1014 (71.9)	**0.003** (0.175)	594 (76.1)	596 (76.3)	0.903 (0.038)
	1	91 (11.7)	184 (13.0)		91 (11.7)	95 (12.2)	
	2	84 (10.8)	151 (10.7)		84 (10.8)	81 (10.4)	
	3 or more	12 (1.5)	61 (4.3)		12 (1.5)	9 (1.2)	
Paid work {No}	Yes	443 (56.7)	781 (55.4)	0.578 (0.027)	443 (56.7)	458 (58.6)	0.473 (0.039)
Tenure	private rent	176 (22.5)	243 (17.2)	**0.015** (0.143)	176 (22.5)	153 (19.6)	0.308 (0.096)
	Mortgaged	238 (30.5)	462 (32.8)		238 (30.5)	265 (33.9)	
	Outright ownership	298 (38.2)	551 (39.1)		298 (38.2)	287 (36.7)	
	Social rent	69 (8.8)	154 (10.9)		69 (8.8)	76 (9.7)	
Language {Other}	English	758 (97.1)	1358 (96.3)	0.428 (0.072)	758 (97.1)	743 (95.1)	0.067 (0.099)
Gender {Female}	Male	371 (47.5)	615 (43.6)	0.088 (0.078)	371 (47.5)	343 (43.9)	0.170 (0.072)
Marital Status {married}	Single	197 (25.2)	308 (21.8)	0.081 (0.080)	197 (25.2)	197 (25.2)	1.000 (0.001)

Before matching: Online (ABOS)=781 and Online (follow up) =1,410 respondnets

After matchin: Online (ABOS)=781 and Online (follow up) = 781 respondnets

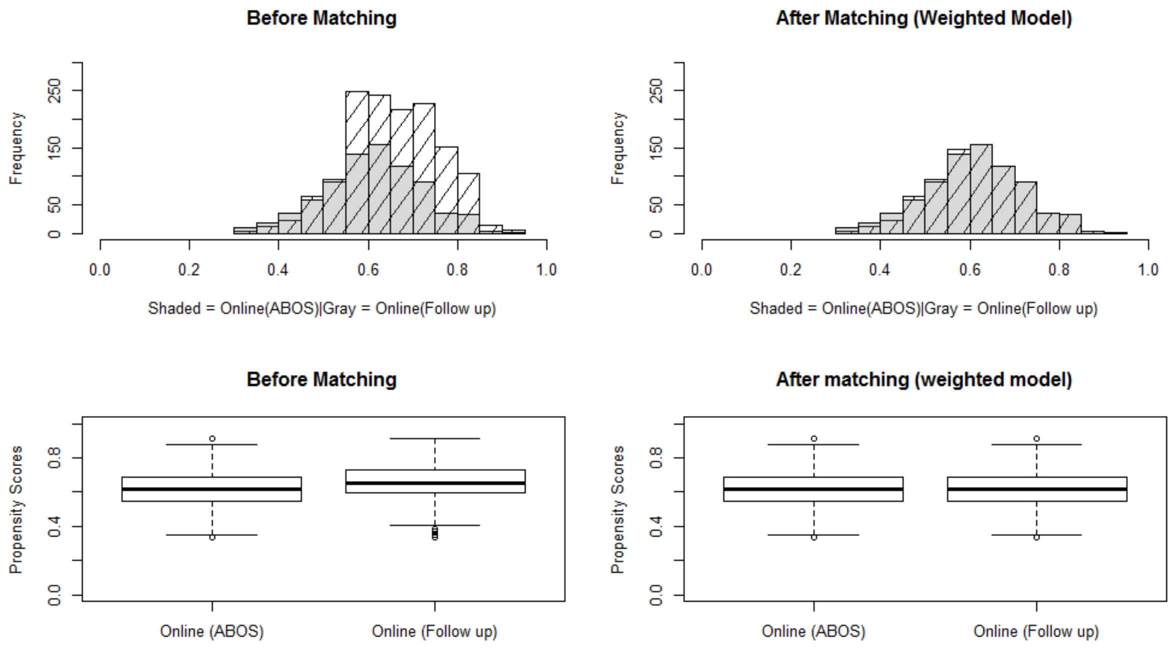

Figure A3: Histograms (top panel) and boxplots (lower panel) of the distributions of propensity scores of online (ABOS) and online (follow up) samples before and after matching.