Separating sample selection from measurement effects in surveys conducted in different modes

Eliud Kibuchi, Gabriele B. Durrant, Patrick Sturgis, Olga Maslovskaya,

University of Southampton

Abstract

The appeal of using mixed-mode surveys has increased in an effort to reduce nonresponse, non-coverage and survey costs. However, different modes may differ due to selection and measurement effects that are confounded. This raises the issue of understanding the effects of mixing modes because it is a challenge to separate mode selection and measurement effects. One of the statistical strategies for adjusting mode effects is propensity score matching that separates selection and measurement effects by rendering the different groups comparable statistically. Therefore, this study examines measurement effects in addressed based online surveying (ABOS) and face-to-face while adjusting for selection effects using propensity score matching. In addition, we evaluate different methods of handing survey weights in propensity score matching analyses of mixed-mode surveys under complex design. The results show mode differences in ABOS and face-to-face before and after matching. Therefore, we cannot assume that ABOS is better or similar than face-to-face. The results also show that propensity score matching had only a minimal effect on the magnitude of mode effects for the surveys considered. This was true even when both surveys were conducted online, which suggests that propensity score matching cannot reliably remove selection differences between surveys.

Keywords: mixed-mode surveys, propensity score matching, selection effects, measurement effects

Introduction

Surveys are increasingly using mixed-mode data collection because of their potential to improve coverage and responses at affordable survey costs (de Leeuw, 2005). However, mixed-mode data collection designs often exhibit differences on the same variables which are difficult to interpret because they comprise a mix of both selection and measurement effects (Tourangeau & Plewes, 2013). In general, interviewer-mediated modes are more representative and generate data of higher quality compared to self-administered modes (Heerwegh, 2009; Roberts, 2007). However, interviewer-mediated modes are costly and many survey organisations are switching to cost effective self-administered modes especially online despite their low response rates and quality (de Leeuw, 2005; Vannieuwenhuyze, Loosveldt, & Molenberghs, 2017). It is against this backdrop that Community Life Survey established a mode experiment to explore whether addressed based online surveying (ABOS) produces data of similar or higher quality than face-to-face. Using data from this experiment Williams (2017b), concluded that ABOS with low response rates produce data with lower net error than a higher response rate face-to-face survey. This is a surprising finding and this paper aims to assess whether it is reasonable and robust conclusion, given the available evidence. It is crucial to note that conclusion made by Williams (2017b) on ABOS being a better mode compared to face-to-face was based on the estimates from online (follow up) and face-to-face. Williams (2017b) makes the two online surveys comparable by applying attrition weights obtained from ABOS sample to the online (follow up) to control for differences in the sampling design effects. However, these attrition weights may have led to underestimation of the mode effects because some of the variables used for their computation were confounded with the choice of the mode making the conclusions not fully warranted. Therefore, this paper also develops on what Williams (2017b) did by using propensity score matching rather than attrition weighting to make comparisons between three samples (i.e. face-to-face, online (follow up), and ABOS). The motivation to use propensity score matching to control for selection effects is informed by the Lugtig et al.(2011). They showed that propensity score matching to be effective at removing selection effects. In addition, we investigate whether controlling for survey weights in propensity score models and outcome analysis affects the estimation of mode effects.

Therefore, this paper contains two main objectives. First, investigates whether ABOS can produce data of better or equivalent quality to face-to-face as suggested by Williams(2017b). Second, it seeks to assess whether propensity score matching is able to remove selection differences as suggested by Lugtig et al.(2011). In addition, this paper evaluates whether different formulations of propensity score models based on surveys weights influence the estimation of propensity scores and mode effects. . The two objectives are addressed using Community Life Survey (CLS) that consists of three experimental samples: face-to-face, online (follow up) and Address Based Online Sampling (ABOS).

Data

Data used in this study are drawn from the Community Life Survey (CLS), which carried out a mixed-mode experiment between July to September 2014 (Williams, 2017b). The survey focuses on issues that are key to encouraging social action and empowering communities such as volunteering, donating, community engagement, civil duty and well-being. The CLS mixed-mode experiment consists of three independent samples that were all administered the same questionnaire: a face-to-face, an online follow up survey and an Addressed Based Online

Surveying (ABOS). The fieldwork for the threes studies was undertake from July to September 2014.

The face-to-face Survey is a multi-stage random sample design was employed for the face-to-face CLS and up to 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.

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 number of respondents for 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 and 161(10.2%) using postal. The postal sub-sample was excluded in this analysis because the focus is on face-to-face and online surveys.

Finally, addressed 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). . Fieldwork for the current study 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 web and 48(5.4%) using postal which were excluded for the final analysis.

Methodology

Propensity Score Matching (PSM) is used to remove sample selection differences between the three independent samples. 0nly the variables that are associated with respondents and may result in non-random differences across different modes are considered as covariates for the propensity score model (Brookhart et al., 2007; Guo & Fraser, 2014). The final propensity score model consist 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 score model is assessed by checking the area of common support that is evaluated using histograms and boxplots. In this analysis, adequate covariate balance is achieved if SMD of less than 0.10 is obtained for each covariate as proposed by Austin (2011). Respondents for two different modes are matched on the logit of the propensity score using greedy nearest neighbour Matching (NNM) without replacement. The choice of greedy NNM is informed by its superior performance in terms of reduced bias for the estimated treatment effects compared to optimal algorithms (Austin, 2012). A matching caliper of width equal to 0.2 of the standard deviation of the logit of the propensity score has been used as a way of improving the quality of matching because it tends to have a modest bias (Austin, 2009). Greedy NNM is implemented using MatchIt package in R (Austin, 2011; Ho, Imai, King, & Stuart, 2009). Three different formulations of propensity score models are fitted: (1) no survey weights in the model (unweighted model), (2) survey weights are incorporated as a covariate in the estimation (weight as covariate), (3) survey weights incorporated as a weighted estimation (weighted model). The quality of the matched samples which is defined in terms of absolute standardised mean differences (SMD) less than 0.10 is then assessed (Rubin, 2001).
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 to. Therefore, nine different methods for estimating the measurement effects: three different methods for estimating the propensity score combined with three different analytic strategies within each matched sample.

Estimation of selection and measurement effects

The mode effects are evaluated using Absolute Percentage Differences (APD). The APD estimates are preferred because they are more intuitively interpretable compared to other measures such as relative absolute differences (Schouten, van den Brakel, Buelens, van der Laan, & Klausch, 2013). The APD is calculated by taking the un-signed difference in the proportion for each survey outcome across independent samples. For categorical variables, APD estimates 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. In order 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. The APD estimates are compared before and after matching and presented graphically based on different formulations of propensity scores . The median is preferred as a measure of central tendency due to outliers and skewness in the distribution of APD estimates.

Results

The SMD results for the three different formulations depending on whether or not weighted regression was used to estimate propensity scores all indicate good balance since all produced SMD lower than 0.10. This indicates that propensity score matching has effectively balanced the face-to-face and online (follow up), face-to-face and ABOS, and ABOS and online (follow up) based on the observed covariates.

The three approaches of outcome analysis for mode effects namely: (1) no survey weights; (2) respondents from matched control mode retain their natural survey weights; and (3) matched control respondents inherit the weights of the other mode respondents to which they are matched have very similar APD estimates across the three matched samples. In this paper, only the weighted estimates using approach (2) are presented here.

Figure 1 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. The pattern of the plots in Figure 1 re similar across the three analysis samples before matching. There is variability in APD estimates after matching across the questions. The median APD for the face-to-face and online (follow up) is at 5 percentage points before matching and increases to 5.5 percentage points after matching. An increase of mode effects after matching indicates that measurement effects across face-to-face and online (follow up) are additive in nature (Schouten et al., 2013; Tourangeau, 2017). This result may indicate that the bulk of the mode effects are due to measurement effects.

For the face-to-face and ABOS, the median APD decreases by 0.2 percentage points after matching from 4.2 to 4.0 percentage points. Although the face-to-face and ABOS sample indicates a small reduction in median APD after matching the overall interpretation of the results is similar to that of face-to-face and online (follow up) Lastly, we observe that the ABOS and online (follow up) has a median APD of 2.6 and 1.9 percentage points before and after

matching represent a 0.7 percentage reduction. The fact that the median of the two online samples is lower after matching suggests that a large part of the of the APD differences in face-to-face and online samples is likely due to measurement effect. In addition, the results of the two online samples goes against the general expectation that their median APD estimate to be close to zero after matching since the mode is same. However, this is not the case and shows that propensity score matching only removes around 30% of the total mode difference. This shows that propensity score matching cannot be assumed to remove all selection effects as concluded by Lugtig et al.(2011). It is also important to note that the APD differences in the two online samples could be because of unobserved background characteristics not controlled for in the propensity score model. It could also partly be due to measurement differences due to the different devices used across the two online samples (de Bruijne & Wijnant, 2013).

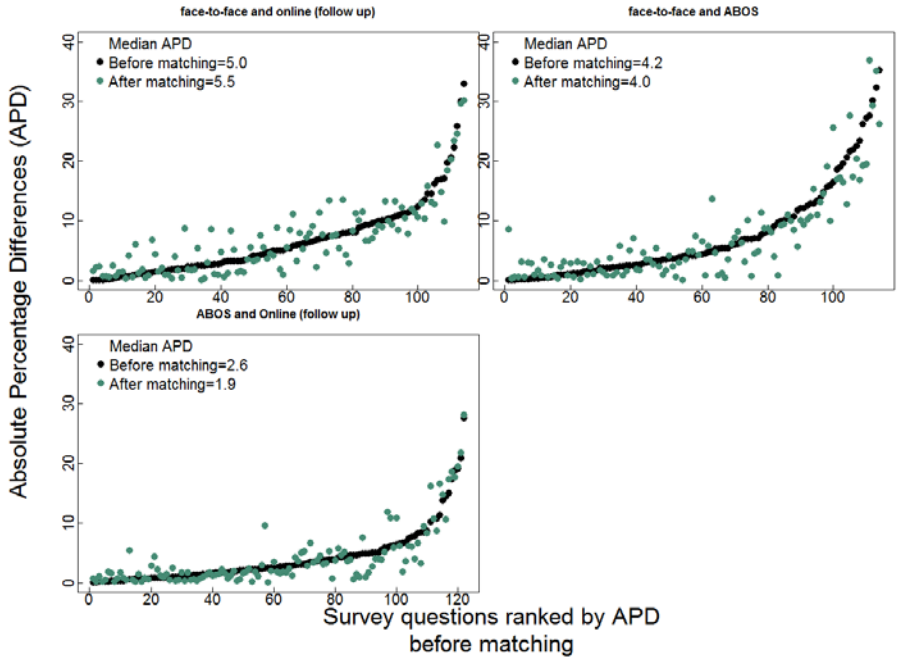

Figure 1: Estimated mode effects by Question before and after matching for face-to-face and online (follow up) (top left panel), face-to-face and ABOS (top right panel), and ABOS and online (follow up) (bottom left panel)

Discussion

The finding show that it does not seem safe to conclude that low response rate ABOS produces better estimates than higher response rate face-to-face. The assertion that ABOS has low net error compared to face-to-face as suggested Williams (2017b) is not really supported by the evidence form our analysis . Therefore, it is important to minimise the overall differences across the two modes before switching from face-to-face to the cost effective ABOS. However, taking a closer look at face-to-face and online samples, we see that in some instances the total mode effects increase after matching an indication that measurement effects across different modes are additive when an overall estimate is used to quantify the differences (Schouten et al., 2013; Tourangeau, 2017). The second finding show that propensity score matching cannot be assumed to effectively remove selection effects as concluded by Lugtig et al.(2011). This is because we still have substantial APD differences after matching the two online samples. Unfortunately, we cannot ascertain the amount of selection effect that remained after matching

using a statistical test. However, the fact that the two online samples have the lowest APD does suggest that a larger part of the difference in APD is due to measurement effects. Alternatively, the substantial bias could have remained due to lack of important confounders that were not observed during data collection. Lastly, we found that it does not make difference whether or not survey weights were incorporated in the estimation of propensity score models and outcome analysis. This is probably because survey weights are computed using socio-demographic variables and therefore provide similar information as other covariates when controlled for in propensity score models.

The results presented in this study have two main substantive implications for survey practice. First, the aim of survey designers is to adopt optimal designs aimed at minimising the substantial difference in APD estimates between ABOS and face-to-face. This may be achieved by exploiting the strengths of both face-to-face and ABOS. Secondly, propensity score matching need further optimisation for effective removal of selection effects in mixed-mode designs. This is possible if variables obtained from sampling frame and administrative data are controlled for in the propensity score model since they unaffected by the choice of the mode. Lastly, the differences between the two online surveys indicate survey designers must put data quality considerations when making device and design decisions for online surveys that can be taken on both personal computers and smartphones.

While we have fully explored the different formulations of propensity score models and estimation of mode effects using absolute percentages this study has some limitations. First, the analysis focuses on only one study that asked attitudinal and behavioural questions from target population of United Kingdom residents. In order to generalise our conclusions into other contexts and countries we need much evidence. In addition, the scope of our study may be limited because of the unobserved characteristics not controlled for in the propensity score model. This study did not consider the direction of measurement errors produced by different modes that may informs whether errors in one mode can offset those in the other leading to the overall reduction of mode effects. Finally yet importantly, this study does not distinguish different mode measurement effects caused by the response styles, interviewer effects, and sensitive questions. These limitations potentially represent other areas of future research.

References

Austin, P. C. (2009). 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. (2011). 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. (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

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.

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. (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

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

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.

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

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.

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/

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. https://doi.org/10.1017/CBO9780511810725.030

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

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., & Plewes, T. J. (2013). *Nonresponse in social science surveys: a research agenda*.

Vannieuwenhuyze, J., Loosveldt, G., & Molenberghs, G. (2017). MIXED-MODE SURVEYS, *74*(5), 1027–1045. https://doi.org/10.1093/poq/nfq059

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