*Journal of Survey Statistics and Methodology (2024) **00**, 1–26*

THE EFFICACY OF PROPENSITY SCORE MATCHING FOR SEPARATING SELECTION AND MEASUREMENT EFFECTS ACROSS DIFFERENT SURVEY MODES

ELIUD KIBUCHI *
PATRICK STURGIS
GABRIELE B. DURRANT
OLGA MASLOVSKAYA

Effective evaluation of data quality between data collected in different modes is complicated by the confounding of selection and measurement effects. This study evaluates the utility of propensity score matching (PSM) as a method that has been proposed as a means of removing selection effects across surveys conducted in different modes. Our results show large differences in estimates for the same variables between parallel face-to-face and online surveys, even after matching on standard demographic variables. Moreover, discrepancies in estimates are still present after matching between surveys conducted in the same (online) mode, where differences in measurement properties can be ruled out a priori. Our findings suggest that PSM has substantial limitations as a method for separating measurement and selection differences across modes and should be used only with caution.

KEY WORDS: Face-to-face interviews; Mixed mode; Mode effects; Online surveys; Propensity score matching.

ELIUD KIBUCHI is a Research Associate at the MRC/CSO Social and Public Sciences Unit, School of Health and Wellbeing, University of Glasgow, Glasgow, United Kingdom. PATRICK STURGIS is Professor of Quantitative Social Science at the Department of Methodology, The London School of Economics and Political Science, London, United Kingdom. GABRIELE B. DURRANT is Professor of Social Statistics & Survey Methodology at the School of Social Statistics and Demography, University of Southampton, Southampton, United Kingdom. OLGA MASLOVSKAYA is an Associate Professor of Survey Research and Social Statistics at the School of Social Statistics and Demography, University of Southampton, Southampton, United Kindgom.
This study design and analysis were not preregistered.
This work was supported by the UK Economic and Social Research Council (ESRC) as part of Work Package 1 of the ESRC National Centre for Research Methods (2014–2019) (grant number ES/L008351/1), the ESRC National Centre for Research Methods (2020–2025) (grant number ES/T000066/1), and the Economic and Social Research Council PhD Studentship grant number ES/J500161/1. Eliud Kibuchi is also funded by the Medical Research Council (MC UU_00022/2) and the Scottish Government Chief Scientist Office (SPHSU17).
*Address correspondence to Eliud Kibuchi, MRC/CSO Social and Public Health Sciences Unit, School of Health and Wellbeing, University of Glasgow, Clarice Pears Building, 90 Byres Road, Glasgow G12 8TB, United Kingdom; E-mail: eliud.kibuchi@glasgow.ac.uk.

https://doi.org/10.1093/jssam/smae017

Statement of Significance

Face-to-face interview surveys are widely viewed as providing the highest quality data of all survey modes. However, they continue to experience increasing costs and declining response rates and face growing challenges of maintaining field forces following the coronavirus 2019 pandemic. As a result of these problematic trends, the use of mixed-mode survey designs has increased, despite their susceptibility to mode effects and the difficulty this poses for comparability. Here, we focus on a commonly used methodological approach—propensity score matching (PSM)—as a means of removing or reducing sample compositional differences caused by differential selection effects between surveys conducted in online and face-to-face modes. We find that nontrivial mode differences between face-to-face and online surveys remain after matching. This provides a note of caution on the use of PSM as a means of isolating the observational and non-observational error components of mode effects in surveys.

1. INTRODUCTION

For many years, face-to-face interviews have been considered the gold standard method of data collection in survey research (de Leeuw 1992; Dillman et al. 2009b). The positive features of face-to-face interviews can mostly be attributed to the presence of interviewers, who can locate and persuade sample members to take part, resulting in higher response rates and potentially more representative samples. Interviewers are also able to motivate respondents to complete complex or cognitively demanding questions, to provide explanations or clarifications for ambiguous wordings, and to use show cards and other supporting materials, all of which are expected to improve measurement quality (Campanelli et al. 1997). However, the substantial costs of face-to-face interviewing, along with increasing rates of nonresponse, have precipitated substantial growth in the use of online data collection and mixed-mode designs in a range of social and economic surveys (Dillman et al. 2009b; Williams and Brick 2018; Bethlehem and Biffignandi 2021; Maslovskaya et al. 2022).

One of the key concerns about this shift to online data collection has been that these self-completion designs tend to yield considerably lower response rates than has been the norm for interviewer-administered surveys. Recent research has questioned the true strength of the correlation between response rates and nonresponse bias (Groves and Peytcheva 2008; Sturgis et al. 2017a). Nevertheless, lower response rates in online surveys still increase the risk of nonresponse bias and hinder their comparability with surveys conducted in modes that attain higher levels of cooperation. A key question is, when a repeated cross-sectional survey transitions from face-to-face to online self-

completion, are the observed measurement differences due to true change or due to sample composition arising from differential nonresponse between modes? The same question can, of course, be asked of mixed-mode surveys which are a standard approach to including offline respondents in (otherwise) online survey designs.

In addition to sample composition, changing a survey to a different mode or introducing mixed-mode designs can hinder comparability because different modes are characterized by heterogeneous measurement error properties (de Leeuw 2005). The presence of differential measurement errors can bias prevalence estimates and distort temporal comparisons in both longitudinal and cross-sectional studies. Measurement effects are most likely to occur when a survey changes from an interviewer-administered to a self-administered mode, or vice versa because the presence of an interviewer can affect respondent behavior in a number of ways (Klausch et al. 2017). For example, there are well-known benefits of interviewers when it comes to measurement quality compared to self-completion because participants are less likely to use satisficing response styles and interviewers' capability of probing encourages valid responses (Krosnick 1991; Goldenbeld and De Craen 2013). However, interviewer administration is more prone to social desirability bias compared to self-completion, especially in surveys involving behavioral and attitudinal questions (Kreuter et al. 2008; Heerwegh 2009; Burkill et al. 2016; Berzelak and Vehovar 2018).

The quality of data produced in different modes would ideally be evaluated by comparing them to external benchmarks or carrying out experiments where respondents are randomly assigned to modes (de Leeuw 2005; Voogt and Saris 2005; Dillman et al. 2009a; Tourangeau 2017). However, high-quality benchmarks are rarely available for most survey variables and most mixed-mode surveys do not randomly assign respondents to modes. This means that respondents interviewed in different modes usually differ by baseline characteristics that are related to both the survey measures and mode selection effects. In such situations, measurement effects are confounded with selection effects, and it is only possible to separate them using methods that require strong assumptions (Voogt and Saris 2005; Weisberg 2005b; Vannieuwenhuyze and Loosveldt 2013). A common approach to achieving this separation is to attempt to balance samples collected in different modes across a vector of measured characteristics using adjustment methods such as propensity score matching (PSM) (Rosenbaum and Rubin 1983; Schonlau and Couper 2006; Austin 2011; Suzer-Gurtekin et al. 2018).

PSM seeks to remove sample compositional differences in baseline characteristics between groups by matching on the probability of exposure (in this case, survey mode) to approximate a randomized assignment design (Rosenbaum and Rubin 1983; Austin 2011). PSM has been used to adjust for confounding in studies exploring mode effects in telephone and web surveys

(Lugtig et al. 2011; Capacci et al. 2018), web and mail (Suzer-Gurtekin et al. 2018), and telephone and mail (Pintor et al. 2015).

The aim of this study is, therefore, to evaluate the performance of PSM as a method of removing or reducing selection effects between samples collected in different modes. We use the specially designed 2014 Community Life Survey (CLS) in England to compare estimates of the same quantities across three studies conducted at the same time using different modes and sample designs with exactly the same questionnaire. We address two primary research questions: (i) how effective is PSM in removing differences in selection effects between modes? and (ii) to what extent do mode measurement effects change after matching based on question characteristics?

The remainder of the paper proceeds as follows. We first provide a concise review of the literature on the effect of survey mode on data quality and the methodological approaches that have been used for conditioning out compositional differences between samples. We then describe the datasets used in the empirical part of the paper and set out our analysis strategy, with the key findings from the analyses presented after that. We conclude with a summary of our main findings, a consideration of the potential limitations of the study, and the implications of our results for survey practice.

1.2 Survey Mode Effects

Differential sample composition between face-to-face and online surveys may arise due to both coverage and self-selection error (Bethlehem 2010). In online surveys, undercoverage is typically found among groups without internet access who tend to be older, less educated, rural-dwelling, and in lower income groups (Bethlehem 2010; Tijdens and Steinmetz 2016). In addition to lower response rates affecting nonresponse error, online surveys are also susceptible to selection bias through within-household selection procedures because there is no interviewer to carry out this procedure (Bethlehem 2010; Khazaal et al. 2014). This may result in differences in sample composition between respondents in online and face-to-face modes, making the estimates obtained noncomparable.

Regarding measurement effects, it has long been known that interviewers can induce both random and systematic errors in face-to-face surveys (Campanelli et al. 1997). For example, the presence of interviewers can make respondents more likely to provide answers that are socially acceptable due to a desire to be perceived positively according to social norms (Berzelak and Vehovar 2018). Online surveys, on the other hand, provide more confidentiality compared to face-to-face surveys, making them less susceptible to social desirability bias (Kreuter et al. 2008; Berzelak and Vehovar 2018). In face-to-face surveys, interviewers can motivate respondents and help them to understand ambiguous or complex questions which is mostly not possible in online

surveys (Roberts 2007; Revilla and Saris 2013). The presence of an interviewer may also reduce the likelihood of respondents adopting satisficing response strategies (Krosnick 1991; Roberts 2007) but may also lead to measurement differences due to interviewer effects (Weisberg 2005a).

As might be expected from this body of research, the survey methodological literature contains many examples of sometimes quite large differences between estimates produced in different modes. For example, Burkill et al. (2016) and Villar and Fitzgerald (2017) found differences in response distributions across the two modes where the same respondents provided responses to attitudinal and behavioral questions, first in a face-to-face survey and then in an online survey. Both studies also reported higher levels of agreement on behavioral compared to attitudinal items in face-to-face and online surveys. Similarly, Heerwegh (2009) found higher rates of "don't knows" and item nonresponse in an online sample compared to the corresponding face-to-face sample. As is typical in these kinds of observational designs, these differences may have been due, at least in part, to selection effects since compositional differences were not controlled.

As we noted earlier, a widely used method for addressing the "measurement/selection" confound is PSM (Rosenbaum and Rubin 1983; Austin 2011). While PSM has mostly been used in medicine and epidemiology contexts (Austin 2008; Granger et al. 2020; Medaglio et al. 2022), it has also been applied in mixed-mode survey contexts (Lugtig et al. 2011; Pintor et al. 2015; Capacci et al. 2018; Suzer-Gurtekin et al. 2018). Notably, Lugtig et al. (2011) applied PSM to separate mode effects from two parallel probability-based surveys conducted using computer-assisted telephone interview (CATI) and online modes of data collection. After matching, they found that large differences between telephone and online surveys remained but there were fewer and smaller differences between a probability and a nonprobability online survey.

Suzer-Gurtekin et al. (2018) were able to remove selection differences caused by mode on a measure of health status in the World Values Survey between two self-administered (web and mail) surveys after controlling for selection effects using PSM. Similarly, Pintor et al. (2015) found PSM to be effective in reducing the magnitude of mode effects in healthcare discrimination estimates among children and adults in the US Minnesota public healthcare program. They found that responses of unfair treatment were 13 percentage points higher among phone respondents compared to those who responded by mail. Capacci et al. (2018), on the other hand, applied PSM to explore mode effects between computer-assisted web interviewing (CAWI) and CATI in a proprietary panel survey in the United Kingdom and Italy. They found large mode differences between modes remained after matching, where CAWI respondents were less likely to support healthy eating policies compared to their CATI counterparts in both countries.

This study aims to build on and extend this existing body of work by applying PSM to a specially designed study in which the same questionnaire was administered in both face-to-face and online survey contexts. We estimate mode effects before and after matching across 133 outcome variables, covering a broad range of topic areas. We present our results separately for attitudinal and behavioral questions as the former are likely to have greater susceptibility to social desirability pressures in face-to-face compared to online surveys (Tourangeau and Yan 2007; Kreuter et al. 2008; Revilla and Saris 2013). We may thus expect larger mode differences after matching for attitudinal compared to behavioral questions. We also assess our results according to whether the question used single or multiple response options, with the expectation that multiple response options will be more cognitively demanding and, hence, more prone to satisficing effects both in the face-to-face and online surveys (Krosnick 1991). However, satisficing effects are expected to be lower in face-to-face than in online surveys due to interviewers' ability to facilitate the response process, which in turn reduces the cognitive demands placed on respondents (Heerwegh and Loosveldt 2008).

2. DATA AND METHODS

The data for this study come from the CLS which is an annual general population survey of adults in England conducted by Kantar Public on behalf of the Department for Digital, Culture, Media and Sport since 2012 (Williams 2017; Kantar Public 2022). The survey covers topics relating to empowering communities such as identity and social networks, community engagement, civic engagement, volunteering, social action, subjective wellbeing, and loneliness. The inferential population is all adults in England aged 16+ and living in private residence (Williams 2017; Kantar Public 2022). In 2014, administration of the CLS questionnaire was implemented using three independent samples as part of a research and development initiative exploring the feasibility of moving the survey to online self-completion. These three samples were a face-to-face survey, an online follow-up of the previous year's face-to-face survey, and a standalone push-to-web online survey—a fresh sample, which was collected online, and where respondents were invited through an invitation letter sent to their home address. All three samples were administered the same questionnaire. Fieldwork for all three survey samples took place between July and September 2014. The study design is summarized in figure 1. The response rates are calculated using the RR1 formula of the American Association of Public Opinion Research (AAPOR) (AAPOR 2016).

The face-to-face sample was a clustered, stratified random sample of addresses drawn from the Postcode Address File (PAF) with a single adult randomly selected by the interviewer at each household (where an address contained more than one household, a single household was randomly selected).

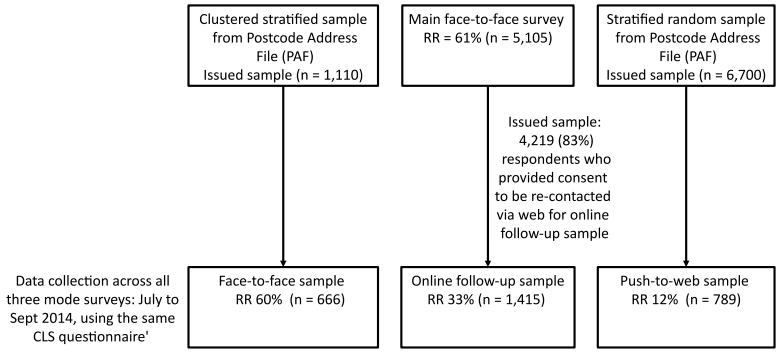

Figure 1. Graphical Illustration of the Study Design for the Three Different Samples Collected in Different Modes: Face-to-Face, Online Follow-Up, and Push-to-Web [RR = Response Rate based on AAPOR (RR1)].

Interviews were conducted using computer-assisted personal interviewing. The issued sample size was 1,110, and 666 respondents were successfully interviewed, resulting in a response rate of 60 percent.

The online follow-up survey was drawn from respondents who had participated in the previous face-to-face round of the CLS in April 2013/14 ($n = 5{,}105$) and who had given consent to be recontacted, which was the case for 83 percent ($n = 4{,}219$) of cases. The net response for the online follow-up survey was 37 percent ($n = 1{,}576$) among invited cases, with 33 percent ($n = 1{,}415$) responding online and 3 percent ($n = 161$) completing a paper questionnaire. The paper questionnaire respondents are excluded from the analysis here because our focus is on the comparison of face-to-face and online self-completion surveys.

The push-to-web sample was a stratified random sample of addresses drawn from the PAF in a single stage and with addresses sampled with equal probability of selection. Letters containing username(s), password(s), and the survey website link were mailed to 6,700 sampled addresses inviting one resident adult to complete the survey online. Where there was more than one eligible adult (16+ years) at an address, the adult who had the last birthday was asked to complete the survey. The achieved sample size was 834, with 789 using online completion, representing an overall response rate of 12 percent. [The 45 respondents (5 percent) who completed the survey using a paper questionnaire are excluded from the analysis for the same reason as above.]

Across the three surveys, there were large differences in response rates, with the standalone face-to-face survey approach reaching by far the highest response rate (60 percent), the online survey following from a previous face-to-face interview achieving 33 percent, and the push-to-web cross-sectional sample achieving the lowest response rate of 12 percent. These differences in

response rates across modes should lead us to expect, a priori, to observe differences in estimates for across survey variables resulting from differential sample composition.

We focus on 133 questions that were administered in all three surveys with full wordings provided in table S1 in the supplementary data online. All questions were recoded to binary variables to enable computation of the absolute percentage difference (APD) because this is more intuitively interpretable than other statistics such as means or medians which depend arbitrarily on the metric of the response scale (see the methodology section 2.1 for further detail).

Any observed differences between the face-to-face and push-to-web survey could be due to a mix of both selection and measurement effects. However, differences in estimates between the push-to-web and online follow-up surveys must be due to differential sample composition only, because the questionnaires were identical and completed in the same mode. The rate of missing values (i.e., item nonresponse) across the variables considered was low, ranging from 0.4 percent for online follow-up to 2.0 percent for push-to-web so we undertook a complete case analysis (Schafer 1999).

2.1 Methodology

We used PSM to mitigate selection differences between the three samples by matching respondents on a set of common observed baseline covariates (Rosenbaum and Rubin 1983; Imbens 2004). The aim was to generate a matched sample, such that for every respondent there was at least one respondent from the comparison sample with similar characteristics on the vector of matching variables. As is standard practice, we followed the following four steps: (i) estimation of propensity scores, (ii) matching of cases, (iii) evaluation of the matching quality, and (iv) estimation of mode differences after matching.

2.1.1 Estimation of propensity scores.

The propensity scores were estimated using a logistic regression model where y_i denotes the mode assigned to individual i as presented in equation (1):

$$y_i = \begin{cases} 1 & \text{mode A} \\ 0 & \text{mode B} \end{cases}, \tag{1}$$

for each individual $i = 1, \ \ldots \ n$, with assignment probabilities for y_i, denoted $\pi_i = Pr(y_i = 1)$, and $(1 - \pi_i) = Pr(y_i = 0)$. The logistic regression model is presented in equation (2) and takes the form:

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

where $\mathrm{B} = (\beta_0, \beta_1, \ldots, \beta_j)$ is a vector of regression coefficients and X_j is a vector of covariates. We use the binary logistic model instead of multinomial modeling, since the latter relies on an additional assumption of independence of irrelevant alternatives which is not always realistic (Vijverberg 2011). In any event, the fitted probabilities from a set of binary logistic regressions will be identical to those obtained in multinomial logistic regressions.

We used the following variables for the estimation of the propensity scores: sex, age, marital status, number of children, paid work, income, ethnic group, number of adults in household, education, tenure, main language, and government office region (GOR). We included only sociodemographic and area-level variables in X_i because they likely influence survey outcomes (i.e., responses to behavioral and attitudinal questions) and the choice of the survey mode (Brookhart et al. 2007; Chen et al. 2022). Austin et al. (2007) showed that a propensity score model that includes confounders (i.e., covariates that are associated with exposure and the outcome) or includes all variables associated with an outcome leads to a reduction in selection bias. Additionally, inclusion of attitudinal and behavioral variables in the propensity model may result in biased estimates because the matching variables are themselves likely to be influenced by survey mode (Brookhart et al. 2007; Cuong 2013). The propensity score models are adjusted using weights which account for unequal selection probabilities in the sample designs and calibration to population totals. We assessed the predictive performance of the propensity score models using pseudo-R squared and area under the curve (AUC)–receiver operating characteristics curve (Plewis et al. 2012).

2.1.2 Matching cases.

The adequacy of the propensity scores for matching was evaluated using a measure of the "area of common support," as is standard in PSM (Austin 2011; Leite 2017). This measures the degree of overlap in the distribution of propensity scores of respondents in the different samples and is assessed using histograms (Austin 2011; Leite 2017). We then proceed to match respondents on the propensity scores using nearest neighbor matching with replacement, such that control cases that have been matched with treated cases are available for consideration as potential matches for other treated cases. The choice of the survey mode to be in the treatment group rather than the control group is informed by the obtained response rate—the mode with the higher response rate is chosen as the treatment group—as these influences both the internal and external validity of the matched samples (Brigham et al. 2009; Malay and Chung 2012). Therefore, the face-to-face mode, given its higher response rate (60 percent) compared to online follow-up (33 percent) and push-to-web (12 percent), is used as the treatment group in respective comparative samples,

while the online follow-up is used as the treatment group in the comparison sample with push-to-web. As a sensitivity analysis, we have also reestimated all models with the treatment and control groups reversed and the results are substantively the same, which would be expected because matching is based on similar propensity scores between the two groups. (The results of these analyses are available from the corresponding author upon request.)

In nearest neighbor matching, each treated case is matched to the nearest control case that lies within a specified range, or "caliper." Nearest neighbor methods are often preferred to other matching algorithms due to its ability to include many control cases in the matched sample, where exact matching approaches can result in many discarded cases due to nonmatching (Gu and Rosenbaum 1993; Austin 2009b). Here, we use three caliper specifications of width 0.005, 0.01, and 0.02. The choice of these caliper widths is based on most used calipers in the PS literature (e.g., Austin 2009b). Using a range of caliper widths enables us to assess the sensitivity of our results to the strictness of the matching criteria and, hence, the number of matched cases. We used a matching ratio of one to many (1:M) where each treated case was matched to many control cases (Gu and Rosenbaum 1993; Austin 2009b). Matching with replacement increases the overall quality of matching due to the sufficiently large size of the matched sample because of few discarded cases (Smith and Todd 2005). This tends to keep the bias low at the expense of larger variance caused by reused cases between the two samples (Smith and Todd 2005). The matching was implemented using the MatchIt package in the R statistical software (version 4.0.2) with code available in https://github.com/Kibuchi-eliud/Mode-effects_PSM.git in Github (Ho et al. 2018).

2.1.3 Evaluation of matching quality.

Evaluation of matching quality provides information about the likely effectiveness of the propensity scores to adjust for confounding (Wang and Donnan 2001; Weitzen et al. 2005). The quality of the matching is assessed using standardized mean differences (SMDs). This allows the comparison of differences in means or proportions of different types of covariates without being influenced by the units of measurement, since it standardizes the differences based on the variance of the samples (Stuart 2010; Linden 2015; Leite 2017). For dichotomous variables, SMDs are calculated as presented in equation (3):

$$\text{SMD} = \frac{(\hat{p}_A - \hat{p}_B)}{\sqrt{(\hat{p}_A(1-\hat{p}_A) + \hat{p}_B(1-\hat{p}_B))/2}}, \tag{3}$$

where $\hat{p}_A$ and $\hat{p}_B$ are proportions of dichotomous variables in modes A and B, respectively. The SMD is a robust approach for evaluating the covariate balance before and after matching because it is not affected by differential sample sizes across comparison groups (Stuart 2010; Austin 2011). Adequate covariate balance for matched samples is achieved if the values of the SMDs are

below 0.1 for all the covariates included in the propensity model (Austin 2011; Lenis et al. 2017; Nguyen et al. 2017). However, a covariate with a value greater than 0.1 in matched samples represents an imbalance between matched comparison groups. An important caveat is that the 0.1 threshold is somewhat arbitrary, and other equally reasonable values could have been used instead. Therefore, in some instances, moderate imbalance SMDs (<0.2) are acceptable, particularly in small samples because they are expected to occasionally occur, even after correctly specifying the propensity score model (Austin 2009a).

2.1.4 Estimation of mode differences.

Mode differences are estimated using APDs between the same attitudinal and behavioral variables measured in different sample modes. The APD was chosen because it produces a more intuitively interpretable quantity compared to other measures such as standardized scores or relative absolute differences (Sturgis et al. 2017). The APD was calculated by taking the unsigned difference in the proportion for each survey outcome across two comparison modes. For categorical variables with K response levels, $(K-1)$ APDs are derived from modal categories, where the omitted categories are those with the lowest frequencies.

Since we cannot rule out residual selection effects on unobserved characteristics between matched face-to-face and online samples, we use the matched online follow-up and push-to-web samples to explore mode effects that we know are unrelated to the measurement differences. Since the questionnaires and survey mode are the same in this comparison, differences in APD between matched online follow-up and push-to-web must be due to measurement effects alone. We also consider median APD estimates by whether the question is behavioral or attitudinal in nature and by whether the question is binary or multicategory to assess whether measurement effects differ by question type. Our expectations here are that the residual differences in APD after matching will be greater for behavioral than attitudinal and for multiple than single response options.

3. RESULTS

Table 1 presents SMDs before (i.e., unmatched) and after matching across the three caliper widths (i.e., 0.005, 0.01, and 0.02) for all matching variables in the propensity score models. To preserve space, the full results for propensity score models across three samples (i.e., face-to-face versus online follow-up, face-to-face versus push-to-web, and online follow-up versus push-to-web) including indicators of model fit (pseudo-R squared and AUC values) are given in table S2 in the supplementary data online. The AUC for face-to-face versus push-to-web was 0.71, suggesting a satisfactory fit, while for face-to-

Table 1. Standardized Mean Differences for Variables in the Propensity Score Models across Three Samples (Face-to-Face versus Online Follow-Up, Face-to-Face versus Push-to-Web, and Online Follow-Up versus Push-to-Web) Based on Three Calipers Width Specifications (0.005, 0.01, and 0.02) after Matching

Calipers	Face-to-face and online follow-up				Face-to-face and push-to-web				Online follow-up and push-to-web			
	Unmatched	Caliper = 0.005	Caliper = 0.01	Caliper = 0.02	Unmatched	Caliper = 0.005	Caliper = 0.01	Caliper = 0.02	Unmatched	Caliper = 0.005	Caliper = 0.01	Caliper = 0.02
Sex	0.01	0.03	0.01	0.02	0.10	0.05	0.05	0.06	0.08	0.01	0.03	0.03
Age	0.27	0.02	0.02	0.03	0.26	0.11	0.13	0.16	0.17	0.11	0.12	0.16
Marital status	0.18	0.01	0.01	0.01	0.10	0.01	0.03	0.05	0.08	0.04	0.05	0.07
Number of children	0.05	0.09	0.09	0.09	0.16	0.05	0.09	0.11	0.18	0.05	0.04	0.05
Paid work	0.09	0.05	0.04	0.05	0.13	0.02	0.04	0.05	0.03	0.01	0.01	0.01
Income	0.22	0.08	0.07	0.07	0.22	0.06	0.09	0.11	0.10	0.06	0.07	0.08
Ethnic group	0.17	0.01	0.03	0.05	0.15	0.05	0.06	0.08	0.03	0.10	0.11	0.10
Number of adults in household	0.22	0.11	0.11	0.11	0.48	0.19	0.24	0.30	0.28	0.11	0.11	0.11
Education	0.27	0.06	0.05	0.06	0.35	0.11	0.14	0.18	0.21	0.06	0.07	0.11
Tenure	0.27	0.09	0.09	0.09	0.28	0.14	0.14	0.19	0.14	0.08	0.08	0.08
Main language	0.10	0.08	0.11	0.13	0.20	0.06	0.15	0.18	0.04	0.03	0.04	0.04
Government office region	0.10	0.15	0.15	0.15	0.30	0.15	0.08	0.11	0.15	0.12	0.09	0.09

face versus online follow-up and online follow-up versus push-to-web was 0.65 and 0.62, respectively, indicating a moderate to good fit. The pseudo-*R*-squared values range between 0.08 and 0.17 across three samples indicating a poor predictive power of the model, although we should not expect a high explanatory power from demographic variables alone. The goal of PSM here is to control for confounding factors and not to fully explain the survey mode allocation.

Table 1 shows that for face-to-face and online follow-up, the unmatched sample had 7 of the 12 with SMD values above 0.1. However, after matching (caliper = 0.005) only two variables (i.e., the number of adults in household and GOR) had SMDs above 0.1 but below 0.2; with the variable "main language" also exceeding 0.1 for calipers 0.01 and 0.02. In this matched sample, single-person households were overrepresented in the face-to-face sample, with two-person households and people resident in the Southeast region underrepresented when compared with the online follow-up (table S3 in the supplementary data online). In the unmatched comparison of face-to-face and push-to-web, 10 of the 12 variables had SMDs greater than 0.1 while, after matching, only four variables (i.e., age, number of adults in household, tenure, and GOR) had SMD values above 0.1 but below 0.2 for caliper 0.005; while for calipers 0.01 and 0.02 there were 5 and 8 variables with SMDs above 0.1 respectively and each has one variable with an SMD above 0.2. The single-person households, not educated, and private renters were overrepresented in the face-to-face sample compared to push-to-web; two-person households, having a degree and above, and being resident in the Southeast were underrepresented in the matched samples (table S4 in the supplementary data online). Finally, the online follow-up and push-to-web matched sample had 6 variables with SMDs above 0.1 before matching and only 3 variables exceeded 0.1 after matching across the three different caliper width specifications and were below 0.2. Those aged 16–34, resident in London, and two-person households were overrepresented in the online follow-up sample, with single-person households underrepresented when compared with push-to-web (table S5 in the supplementary data online). The descriptive statistics of the unmatched and matched sample comparisons are presented in tables S3–S5 in the supplementary data online.

Table 2 presents the sample sizes of three matched sample comparisons before and after matching. Overall, the number of matched pairs increased as the width of caliper specifications increased across the three matched sample comparisons, as would be expected. The percentage of discarded cases across matched comparisons is highest for the online follow-up sample and ranges between 45 percent in online follow-up versus push-to-web to 69 percent in face-to-face versus online follow-up. This can be attributed to the higher number of participants in online follow-up compared to face-to-face and push-to-web. The number of reused cases is highest in the face-to-face and push-to-web compared to the other two comparison samples. Common support is

Table 2. Sample Sizes across Three Samples (Face-to-Face versus Online Follow-Up, Face-to-Face versus Push-to-Web, and Online Follow-Up versus Push-to-Web) before and after Matching based on Three Caliper Width Specifications (0.005, 0.01, and 0.02)

	Face-to-face and online follow-up			Face-to-face and push-to-web			Online follow-up and push-to-web		
	Face-to-face	Online follow-up	Reused cases	Face-to-face	Push-to-web	Reused cases	Online follow-up	Push-to-web	Reused cases
Unmatched	666	1,410		663	765		1,410	781	
Matched									
Caliper = 0.005									
Matched	603 (90.5)	431 (30.6)	119	489 (73.8)	549 (71.8)	362	692 (49.1)	477 (61.1)	151
Discarded	63 (9.5)	979 (69.4)		174 (26.2)	216 (28.2)		718 (50.9)	304 (38.9)	
Caliper = 0.01									
Matched	632 (94.9)	445 (31.6)	129	572 (86.3)	613 (80.1)	483	742 (52.6)	499 (63.9)	163
Discarded	34 (5.1)	965 (68.4)		91 (13.7)	152 (19.9)		668 (47.4)	282 (36.1)	
Caliper = 0.02									
Matched	649 (97.4)	452 (32.1)	136	611 (92.2)	624 (81.6)	512	770 (54.6)	504 (64.5)	171
Discarded	17 (2.6)	958 (67.9)		52 (7.8)	141 (18.4)		640 (45.4)	277 (35.5)	

achieved across all three matched sample comparisons according to conventional criteria, as depicted using histograms for calipers with a width of 0.005 before and after matching in figures S1–S3 in the supplementary data online. We have also presented SMD plots for unmatched and matched sample comparisons (calipers of width 0.005) in figures S4–S6 in the supplementary data online. Overall, qualitatively comparable balance in the covariates was attained based on all three different caliper specifications in the matched sample comparisons.

Figure 2 presents the APD estimates for the three sample comparisons before and after matching for all three caliper specifications. The questions on the *X*-axis are ranked by the magnitude of the APD before matching and the *Y*-axis shows the magnitude of APDs. The patterns of the plots in figure 2 are similar across all three sample comparisons before matching, although the magnitudes of the differences are considerably larger for face-to-face and online samples before matching when compared to the difference between the two online samples, as would be expected.

Figure 2 shows that the median APDs for the face-to-face and online follow-up comparison is 5.0 percentage points before matching, *increasing* to between 5.7 and 6.0 percentage points after matching. Similarly, the median APD for the face-to-face and push-to-web comparison *increases* slightly from 4.2 percentage points before matching to between 4.4 and 4.8 after matching. On the face of it, this is a surprising result because the expectation is that the average mode effects should decrease after controlling for selection effects. However, the counterintuitive pattern may be due to selection and measurement effects having opposite signs, such that they counteract rather than reinforce each other (Schouten et al. 2013). Overall, the median APDs when comparing the face-to-face sample and either of the two online samples are approximately two times larger, even after matching, than those found when comparing the online follow-up and push-to-web, indicating the presence of mode effects. If the matching procedure has successfully removed the selection effect component of the mode difference, we conclude that nearly all the mode differences between the face-to-face and online samples are due to measurement effects.

Figure 2 shows that the median APDs for the comparison of the two online surveys before matching is 2.6 percentage points which remains unchanged after matching using caliper width 0.005. However, the median APD increases to 2.7 and 2.9 percentage points for caliper widths 0.01 and 0.02, respectively, after matching. As these two surveys were conducted in the same mode, we can conclude with confidence that measurement effects are approximately zero and that the observed differences can be attributed to the different sampling strategies used for online follow-up and online push-to-web samples. These results demonstrate that even after accounting for selection differences between matched face-to-face and online surveys, the median APD is approximately 2 percentage points or more (i.e., the difference in median APD between matched face-to-face and online samples and matched online samples).

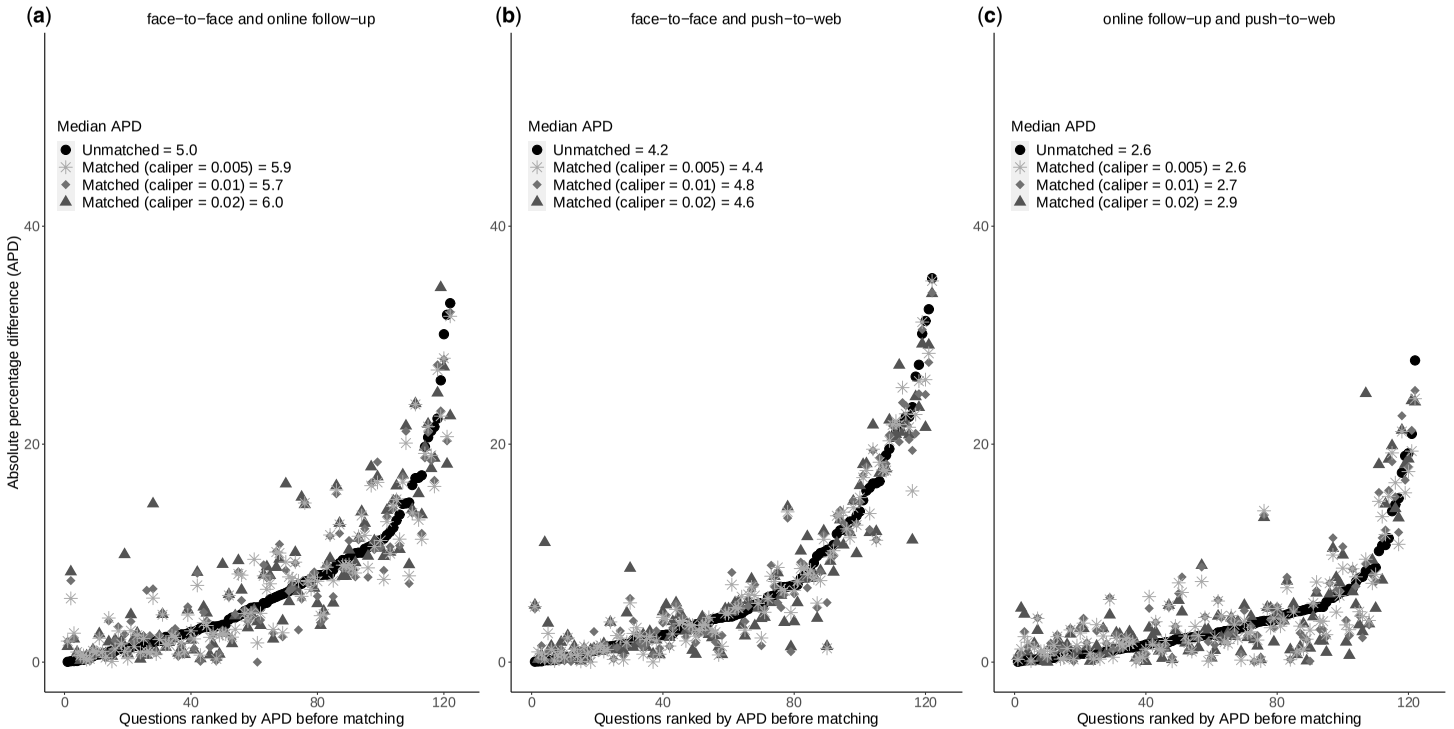

Figure 2. Estimated Mode Effects, as Measured by the Absolute Percentage Difference (APD) by Question before (Black Square) and after Matching (Gray Triangle) for (a) Face-to-Face and Online Follow-Up, (b) Face-to-Face and Push-to-Web, and (c) Online Follow-Up and Push-to-Web.

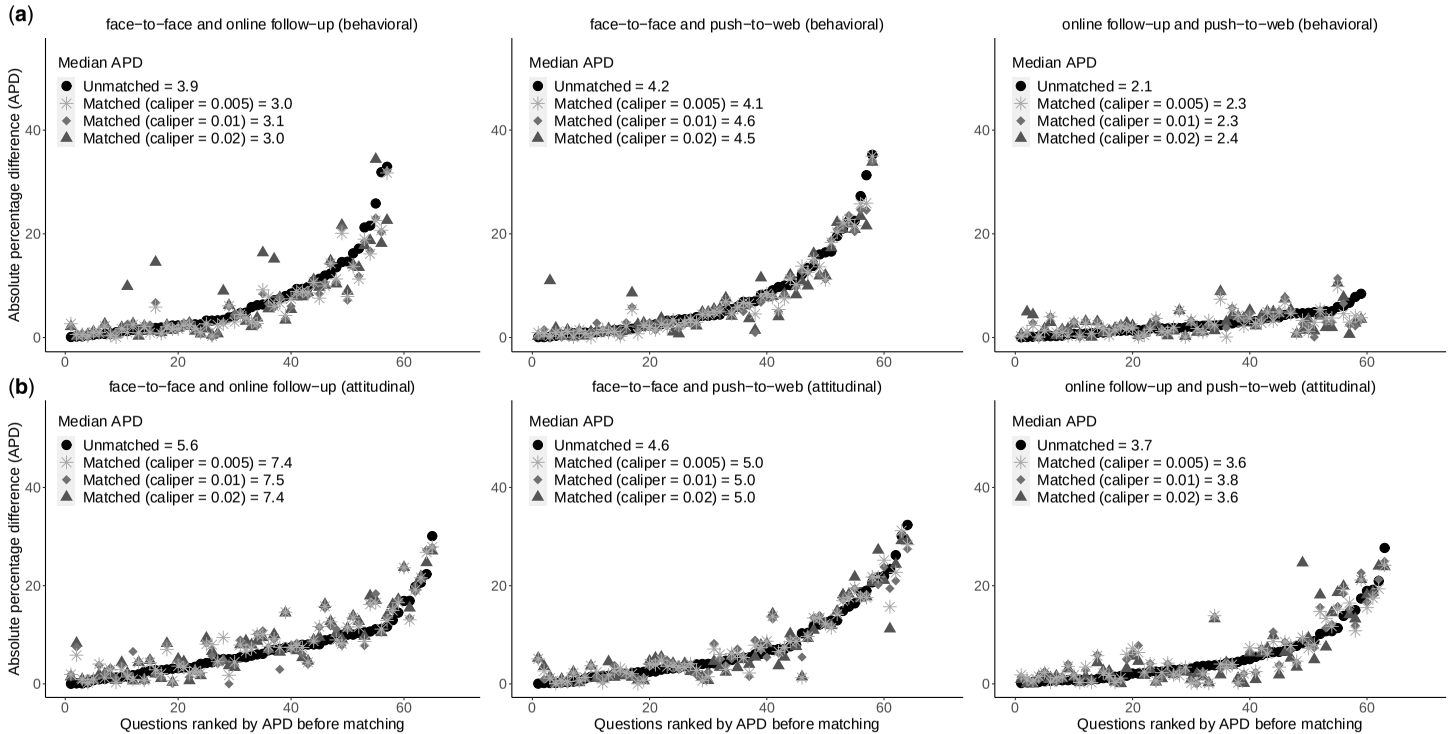

Figure 3. Estimated Mode Effects as Measured by the Absolute Percentage Difference (APD) Based on Question Type [Top Panel (a) Behavioral and Bottom Panel (b) Attitudinal] before and after Matching across Face-to-Face and Online Follow-Up, Face-to-Face and Push-to-Web, and Online Follow-Up and Push-to-Web.

Figure 3 assesses whether differences in mode effects before and after matching are related to question type. Overall, attitudinal questions do indeed tend to have higher mode effects across all sample comparisons both before and after matching compared to behavioral questions. Figure 3 shows the median APD differences for attitudinal questions in matched samples for face-to-face and online follow-up is 1.4 across all caliper widths and ranged between 0.4 and 0.9 for face-to-face and push-to-web, and 1.2 to 1.5 for online follow-up and push-to-web when compared to behavioral questions. We also find that the median APD for pure measurement effects after accounting for sample composition differences ranged between 0.6 and 0.7 percentage points in the matched face-to-face and online follow-up and 1.8 and 2.3 in for face-to-face and push-to-web for behavioral questions. The measurement effects increased to a range of 3.7–3. 8 in the face-to-face and online follow-up comparison and reduced to 1.2–1.4 percentage points in face-to-face and push-to-web for attitudinal questions. Overall, the comparison between behavioral and attitudinal questions in both the unmatched and matched samples for face-to-face versus online follow-up and face-to-face versus push-to-web were not statistically significant. However, they were statistically significant for the online follow-up versus push-to-web comparison.

Finally, figure 4 shows that, as expected, multicategory questions had larger mode effects both before and after matching compared to questions with only two answer categories. The APD estimates after matching across the three sample comparisons ranged from 2.4 to 4.7 times higher for multicategory questions when compared to binary questions. The median APD for the measurement effect, that is, the effect after "removing" selection differences, ranged between 0.5 and 0.8 percentage points for binary questions compared to 3.9 and 5.4 percentage points for multicategory questions in face-to-face and online follow-up. This same pattern is observed in the face-to-face and push-to-web comparison, where the range of pure measurement effect for binary questions (0.6–0.8) is narrower than for multicategory questions (0.3–4.0). Overall, the comparisons between binary and multicategory questions for both the unmatched and matched samples were statistically significant indicating true differences in means. Frequencies by question type and category are presented in figure S7 in the supplementary data online.

4. DISCUSSION

While face-to-face interviewing has long been considered to produce the highest quality data of all survey modes, the empirical evidence is not as clear-cut as one might expect. In large part, this is due to a lack of suitable criterion variables for making comparisons of accuracy between modes and the confounding of measurement and selection effects in "naïve" comparisons across survey modes (de Leeuw 1992; Tourangeau and Yan 2007; Vannieuwenhuyze

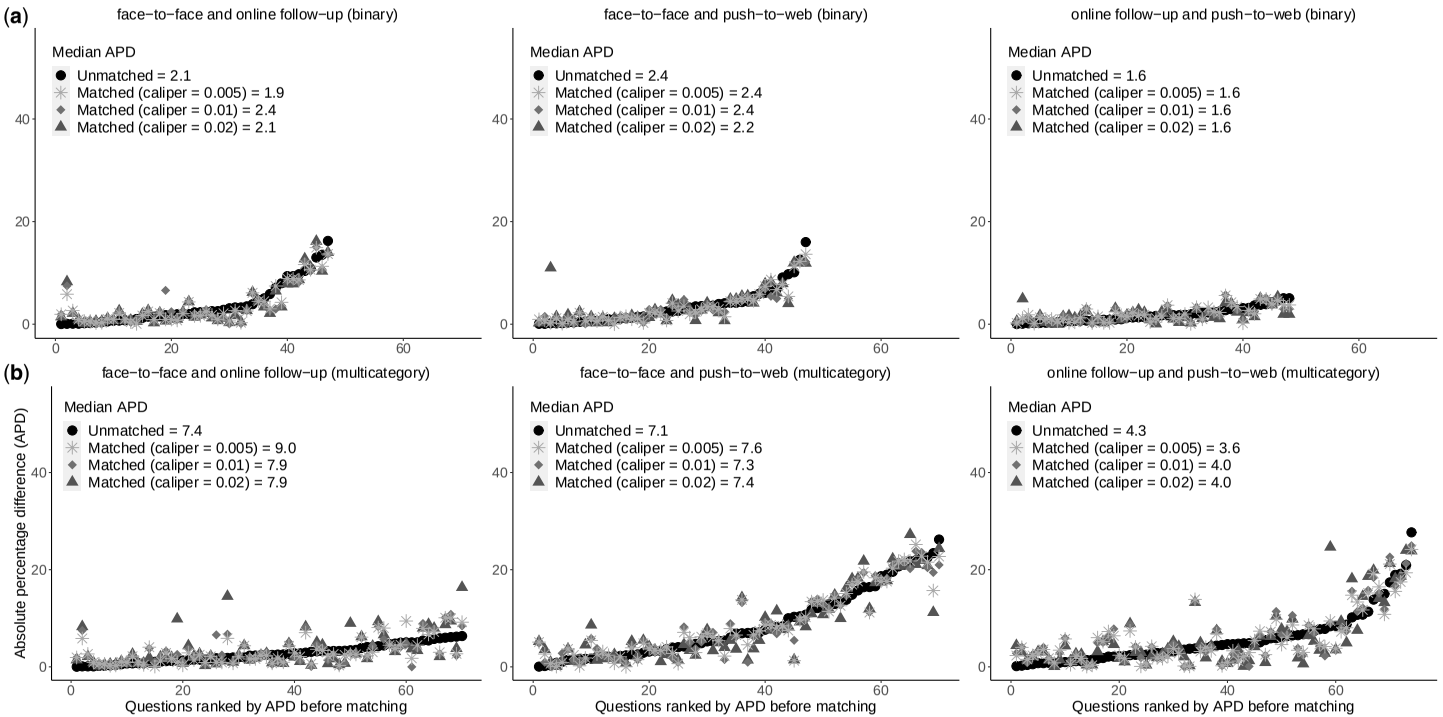

Figure 4. Estimated Mode Effects as Measured by the Absolute Percentage Difference (APD) Based on Question Category [Top Panel (a) Binary and Bottom Panel (b) Multicategory] before and after Matching across Face-to-Face and Online Follow-Up, Face-to-Face and Push-to-Web, and Online Follow-Up and Push-to-Web.

and Loosveldt 2013; Burkill et al. 2016; Villar and Fitzgerald 2017). An increasingly popular way of addressing this confounding problem is to use PSM to remove, or at least reduce, sample composition differences between samples collected in different modes. The logic here is if we can remove differences in sample composition, then what is left can be attributed to the measurement properties of the different modes. The aim of this study has been to assess the suitability of PSM for this purpose using a specially designed study in which data were collected using different modes at the same time but using the same questionnaire. We addressed the following two research questions: (i) how effective is PSM in removing selection effects in face-to-face and online surveys? and (ii) do question characteristics affect mode effects after matching?

The results of our analyses revealed nontrivial differences in estimates between face-to-face and online samples before and after matching. Smaller differences were found where both samples were collected online but employed different sampling designs (online follow-up and online push-to-web samples). We also found larger measurement effects in the face-to-face versus online follow-up comparison than in the face-to-face versus push-to-web sample comparison. Attitudinal and multicategory questions were more prone to measurement effects when compared to behavioral and binary questions, respectively. These findings are consistent with Villar and Fitzgerald (2017) who found that behavioral questions tend to show higher levels of agreement between face-to-face and online surveys compared to attitudinal items. However, it should be noted that the majority of attitudinal questions in our study have multiple response options, so we cannot draw a strong conclusion about their independent effects.

Matching on common observed demographic variables did little to reduce the magnitude of differences between estimates produced in face-to-face and online modes; median mode effects between face-to-face and online samples were above 2.5 percentage points, after accounting for sample composition differences. This provides more evidence that the majority of the raw mode effects appear to be attributable to measurement differences across face-to-face and online samples rather than to selection effects (Tourangeau et al. 2013; Shino et al. 2022).

These conclusions, however, are subject to the strong caveat that matching may not remove all the compositional differences across samples as there is always the possibility that unobserved factors are missing from the vector of matching variables in the prediction models. For example, the inclusion of paradata such as contact information, device type, and response times may improve the quality of the matching (Kreuter et al. 2010; West 2011; Callegaro 2013). Unfortunately, variables of this type were not available in our dataset but would represent a useful extension in future studies. While the predictive power obtained in the fitted propensity score models was low, we do not expect this to have influenced the ability of the models to adjust for

selection bias since the covariates we included in the models were either true confounders of the survey mode and outcome variables or were related to the outcome only (Rubin and Thomas 1996; Brookhart et al. 2007). Moreover, high predictive power is not essential for successful matching (Alves 2022). In this regard, we found that matching was successful in removing only around a quarter of the selection effects between the two online surveys. We can attribute this difference to sample composition with high confidence because, for this comparison, mode-related measurement differences can be ruled out *a priori* as exactly the same questionnaire was administered in the same mode. This means that some important variables must indeed have been missing from the matching vector. That being said, it is possible that previous survey experience for the online follow-up participants may have influenced their responses and may have accounted for some of measurement differences. However, given the relatively long time lag between the face-to-face and online follow-up surveys, large effects of this nature do not seem likely.

Contrary to expectation, the median APD between face-to-face and online sample comparisons *increased* after matching. This implies that selection and measurement effects are moving in opposite directions for some variables (Schouten et al. 2013; Tourangeau 2017). For example, a lower response rate in an online sample might result in higher estimates for volunteering when compared to a face-to-face survey with a higher response rate if the tendency to volunteer is correlated with the response propensity. At the same time, social desirability bias is likely to be higher in the face-to-face mode, increasing the volunteering estimate and offsetting the compositional difference, partially or fully. In this scenario, successful matching between samples would increase and not reduce the magnitude of the difference in estimates of volunteering. Clearly, the possibility that this phenomenon might affect some (but not all) variables complicates the interpretation of pre- and postmatching estimates.

Our findings here are consistent with recent developments in the causal inference literature which have found that PSM can increase rather than reduce imbalance (King and Nielsen 2015). In any event, the fact that the matching had such a small effect on the difference in estimates between face-to-face and online samples implies that the larger part of the difference between modes is due to measurement rather than sample composition differences. We also found that different caliper specifications resulted in very similar APD estimates across matched samples, suggesting that stricter matching criteria did little to change the substantive findings. We must also point out that while nearest neighbor matching with replacement should reduce bias because less information is discarded, it may also increase the variability of measurement effects due to reused cases since this method tends to decrease the effective sample size (Smith and Todd 2005). However, sensitivity analyses conducted using an optimal matching approach (not presented here) produced substantively similar conclusions.

Of course, our analysis is based on a single study in one country with a focus on a particular substantive area. Generalization to other contexts, topic

areas, and survey designs should therefore be done with caution. Additionally, while the basic methodological principles remain unchanged, the data used for the analysis were collected 10 years ago and there have been technological changes in that time which could potentially affect some of our findings. For example, UK internet penetration was 92 percent in 2014, slightly lower compared to the current 97 percent (Petrosyan 2003). During this time, internet use among adults aged 75+ has nearly doubled from 29 percent (2013) to 54 percent (2020) indicating that this group may have been underrepresented in the online surveys (Prescott 2021).

The results of this study have two main implications for survey practice. First, we provide further evidence that there are substantial mode differences between online surveys and surveys conducted using a face-to-face mode of data collection. Therefore, survey designers and commissioners must be cautious when switching a survey from one mode to another. Second, we have shown that PSM is not a straightforward solution for mitigating differential selection effects in surveys conducted in different modes. This is because quite large differences are evident across mode samples even after matching. Additionally, the potential for measurement and selection effects to move in the same or opposite directions for different variables complicates the interpretation of results using this method, as does the possibility of missing important variables from the matching vector. Overall, our findings suggest that PSM and related approaches are likely to be of only limited utility in separating measurement and sample composition differences across surveys conducted in different modes. Researchers should use these methods only with caution and with sensitivity to the problems we have identified here.

Supplementary Materials

Supplementary materials are available online at academic.oup.com/jssam.

DATA AVAILABILITY

Data for the online follow-up survey are not publicly available but may be obtained from a third party, Kantar Public, who collected the data. The main face-to-face survey and the push-to-web survey can be accessed via the UK Data Service: main face-to-face survey: https://doi.org/10.5255/UKDA-SN-7836-1 and push-to-web survey: https://doi.org/10.5255/UKDA-SN-7900-1.

REFERENCES

AAPOR (2016), *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys* (9th ed.), Alexandria,VA: American Association for Public Opinion Research

(AAPOR). Available at https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf

Alves, M. F. (2022), "Causal Inference for the Brave and True," São Paulo, Brazil. Available at https://matheusfacure.github.io/python-causality-handbook/01-Introduction-To-Causality.html

Austin, P. C. (2008), "A Critical Appraisal of Propensity-Score Matching in the Medical Literature between 1996 and 2003," *Statistics in Medicine*, 27, 2037–2049.

———. (2009a), "Balance Diagnostics for Comparing the Distribution of Baseline Covariates between Treatment Groups in Propensity-Score Matched Samples," *Statistics in Medicine*, 28, 3083–3107.

———. (2009b), "Some Methods of Propensity-Score Matching Had Superior Performance to Others: Results of an Empirical Investigation and Monte Carlo Simulations," *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 51, 171–184.

———. (2011), "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies," *Multivariate Behavioral Research*, 46, 399–424.

Austin, P. C., Grootendorst, P., and Anderson, G. M. (2007), "A Comparison of the Ability of Different Propensity Score Models to Balance Measured Variables between Treated and Untreated Subjects: A Monte Carlo Study," *Statistics in Medicine*, 26, 734–753.

Berzelak, N., and Vehovar, V. (2018), "Mode Effects on Socially Desirable Responding in Web Surveys Compared to Face-to-Face and Telephone Surveys," *Advances in Methodology and Statistics*, 15, 21–43.

Bethlehem, J. (2010), "Selection Bias in Web Surveys," *International Statistical Review*, 78, 161–188.

Bethlehem, J., and Biffignandi, S. (2021), *Handbook of Web Surveys*, Hoboken, NJ: John Wiley & Sons.

Brigham, G. S., Feaster, D. J., Wakim, P. G., and Dempsey, C. L. (2009), "Choosing a Control Group in Effectiveness Trials of Behavioral Drug Abuse Treatments," *Journal of Substance Abuse Treatment*, 37, 388–397.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Stürmer, T. (2007), "Variable Selection for Propensity Score Models," *American Journal of Epidemiology*, 163, 1149–1156.

Burkill, S., Copas, A., Couper, M. P., Clifton, S., Prah, P., Datta, J., Conrad, F. G., Wellings, K., Johnson, A. M., and Erens, B. (2016), "Using the Web to Collect Data on Sensitive Behaviours: A Study Looking at Mode Effects on the British National Survey of Sexual Attitudes and Lifestyles," *PloS One*, 11, e0147983.

Callegaro, M. (2013), "Paradata in Web Surveys," in *Improving Surveys with Paradata: Analytic Uses of Process Information*, ed. F. Kreuter, Hoboken, NJ: John Wiley & Sons, pp. 259–279.

Campanelli, P., Sturgis, P., and Purdon, S. (1997), *Can You Hear Me Knocking? And Investigation into the Impact of Interviewers on Survey Response Rates*, Southampton, England, United Kingdom: National Centre for Social Research.

Capacci, S., Mazzocchi, M., and Brasini, S. (2018), "Estimation of Unobservable Selection Effects in On-Line Surveys through Propensity Score Matching: An Application to Public Acceptance of Healthy Eating Policies," *PloS One*, 13, e0196020.

Chen, J. W., Maldonado, D. R., Kowalski, B. L., Miecznikowski, K. B., Kyin, C., Gornbein, J. A., and Domb, B. G. (2022), "Best Practice Guidelines for Propensity Score Methods in Medical Research: Consideration on Theory, Implementation, and Reporting. A Review," *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, 38, 632–642.

Cuong, N. V. (2013), "Which Covariates Should Be Controlled in Propensity Score Matching? Evidence from a Simulation Study," *Statistica Neerlandica*, 67, 169–180.

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

———. (2005), "To Mix or Not to Mix Data Collection Modes in Surveys," *Journal of Official Statistics*, 21, 233–255.

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., and Messer, B. L. (2009a), "Response Rate and Measurement Differences in Mixed-Mode Surveys Using Mail,

Telephone, Interactive Voice Response (IVR) and the Internet," *Social Science Research*, 38, 1–18.
Dillman, D. A., Smyth, J. D., and Christian, L. M. (2009b), *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method* (3rd ed.), Hoboken, NJ: John Wiley & Sons.
Goldenbeld, C., and De Craen, S. (2013), "The Comparison of Road Safety Survey Answers between Web-Panel and Face-to-Face; Dutch Results of SARTRE-4 Survey," *Journal of Safety Research*, 46, 13–20.
Granger, E., Watkins, T., Sergeant, J. C., and Lunt, M. (2020), "A Review of the Use of Propensity Score Diagnostics in Papers Published in High-Ranking Medical Journals," *BMC Medical Research Methodology*, 20, 132–139.
Groves, R. M., and Peytcheva, E. (2008), "The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis," *Public Opinion Quarterly*, 72, 167–189.
Gu, X. S., and Rosenbaum, P. R. (1993), "Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms," *Journal of Computational and Graphical Statistics*, 2, 405–420.
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, 111–121.
Heerwegh, D., and Loosveldt, G. (2008), "Face-to-Face versus Web Surveying in a High-Internet-Coverage Population: Differences in Response Quality," *Public Opinion Quarterly*, 72, 836–846.
Ho, D., Imai, K., King, G., Stuart, E., and Whitworth, A. (2018), "Package 'MatchIt'." Version.
Imbens, G. W. (2004), "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *The Review of Economics and Statistics*, 86, 4–29.
Kantar Public (2022), "Community Life Survey," Technical Report 2020/21.
Khazaal, Y., Van Singer, M., Chatton, A., Achab, S., Zullino, D., Rothen, S., Khan, R., Billieux, J., and Thorens, G. (2014), "Does Self-Selection Affect Samples' Representativeness in Online Surveys? An Investigation in Online Video Game Research," *Journal of Medical Internet Research*, 16, e2759.
King, G., and Nielsen, R. (2015), *Why Propensity Scores Should Not Be Used for Matching*, 617, 1–36. Available at https://gking.harvard.edu/files/gking/files/psnot.pdf.
Klausch, T., Schouten, B., and Hox, J. J. (2017), "Evaluating Bias of Sequential Mixed-Mode Designs against Benchmark Surveys," *Sociological Methods & Research*, 46, 456–489.
Kreuter, F., Couper, M., and Lyberg, L. (2010), "The Use of Paradata to Monitor and Manage Survey Data Collection," in *Proceedings of the Joint Statistical Meetings*, American Statistical Association, pp. 282–296.
Kreuter, F., Presser, S., and Tourangeau, R. (2008), "Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity," *Public Opinion Quarterly*, 72, 847–865.
Krosnick, J. A. (1991), "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys," *Applied Cognitive Psychology*, 5, 213–236.
Leite, W. (2017), *Practical Propensity Score Methods Using R*, Thousand Oaks, CA: SAGE Publications.
Lenis, D., Nguyen, T. Q., Dong, N., and Stuart, E. A. (2017), "It's All about Balance: Propensity Score Matching in the Context of Complex Survey Data," *Biostatistics*, 20, 147–163.
Linden, A. (2015), "Graphical Displays for Assessing Covariate Balance in Matching Studies," *Journal of Evaluation in Clinical Practice*, 21, 242–247.
Lugtig, P., Lensvelt-Mulders, G. J. L. M., Frerichs, R., and Greven, A. (2011), "Estimating Nonresponse Bias and Mode Effects in a Mixed-Mode Survey," *International Journal of Market Research*, 53, 669–686.
Malay, S., and Chung, K. C. (2012), "The Choice of Controls for Providing Validity and Evidence in Clinical Research," *Plastic and Reconstructive Surgery*, 130, 959–965.
Maslovskaya, O., Struminskaya, B., and Durrant, G. (2022), "The Future of Online Data Collection in Social Surveys: Challenges, Developments and Applications," *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185, 768–772.

Medaglio, D., Stephens-Shields, A. J., and Leonard, C. E. (2022), "Research and Scholarly Methods: Propensity Scores," *Journal of the American College of Clinical Pharmacy*, 5, 467–475.

Nguyen, T. L., Collins, G. S., Spence, J., Daurès, J. P., Devereaux, P. J., Landais, P., and Le Manach, Y. (2017), "Double-Adjustment in Propensity Score Matching Analysis: Choosing a Threshold for considering Residual Imbalance," *BMC Medical Research Methodology*, 17, 78–78.

Petrosyan, A. (2003), "Share of Individuals using the Internet in the United Kingdom (UK) from 2002 to 2023." Available at https://www.statista.com/statistics/1124328/internet-penetration-uk/

Pintor, J. B. K., McAlpine, D., Beebe, T. J., and Johnson, P. J. (2015), "Propensity Score Matching to Measure the Effect of Survey Mode on Reports of Racial and Ethnic Discrimination in Health Care," *Medical Care*, 53, 471–476.

Plewis, I., Ketende, S., and Calderwood, L. (2012), "Assessing the Accuracy of Response Propensity Models in Longitudinal Studies," *Survey Methodology*, 38, 167–171.

Prescott, C. (2021), *Internet Users, UK: 2020*, Newport, Wales, United Kingdom: Office for National Statistics. Available at https://www.ons.gov.uk/businessindustryandtrade/itandinternetindustry/bulletins/internetusers/2020

Revilla, M. A., and Saris, W. E. (2013), "A Comparison of the Quality of Questions in a Face-to-Face and a Web Survey," *International Journal of Public Opinion Research*, 25, 242–253.

Roberts, C. (2007), *Mixing Modes of Data Collection in Surveys: A Methodological Review*, Southampton, England, United Kingdom: National Centre for Research Methods (NCRM) Publications. Available at https://eprints.ncrm.ac.uk/id/eprint/418/

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

Rubin, D. B., and Thomas, N. (1996), "Matching Using Estimated Propensity Scores: Relating Theory to Practice," *Biometrics*, 52, 249–264.

Schafer, J. L. (1999), "Multiple Imputation: A Primer," *Statistical Methods in Medical Research*, 8, 3–15.

Schonlau, M., and Couper, M. P. (2006), "Selection Bias in Web Surveys and the Use of Propensity Scores."

Schouten, B., van den Brakel, J., Buelens, B., van der Laan, J., and Klausch, T. (2013), "Disentangling Mode-Specific Selection and Measurement Bias in Social Surveys," *Social Science Research*, 42, 1555–1570.

Shino, E., Martinez, M. D., and Binder, M. (2022), "Determined by Mode? Representation and Measurement Effects in a Dual-Mode Statewide Survey," *Journal of Survey Statistics and Methodology*, 10, 183–202.

Smith, J. A., and Todd, P. E. (2005), "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?," *Journal of Econometrics*, 125, 305–353.

Stuart, E. A. (2010), "Matching Methods for Causal Inference: A Review and a Look Forward," *Statistical Science*, 25, 1–21.

Sturgis, P., Williams, J., Brunton-Smith, I., and Moore, J. (2017), "Fieldwork Effort, Response Rate, and the Distribution of Survey Outcomes: A Multilevel Meta-Analysis," *Public Opinion Quarterly*, 81, 523–542.

Suzer-Gurtekin, Z. T., Valliant, R., Heeringa, S. G., and de Leeuw, E. D. (2018), "Mixed-Mode Surveys: Design, Estimation and Adjustment Methods," in *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)*, eds. T. P. Johnson, B. Pennell, I. A. L. Stoop, and B. Dorer, Hoboken, NJ: John Wiley & Sons, pp. 409–430.

Tijdens, K., and Steinmetz, S. (2016), "Is the Web a Promising Tool for Data Collection in Developing Countries? An Analysis of the Sample Bias of 10 Web and Face-to-Face Surveys from Africa, Asia, and South America," *International Journal of Social Research Methodology*, 19, 461–479.

Tourangeau, R. (2017), "Mixing Modes: Tradeoffs among Coverage, Nonresponse, and Measurement Error," in *Total Survey Error in Practice*, eds. P. Biemer, E. de Leeuw, S.

Eckman, B. Edwards, F. Kreuter, L. Lyberg, N. C. Tucker, and B. T. West, Hoboken, NJ: John Wiley & Sons, pp. 115–132.
Tourangeau, R., Conrad, F. G., and Couper, M. P. (2013), *The Science of Web Surveys*, New York City, NY: Oxford University Press.
Tourangeau, R., and Yan, T. (2007), "Sensitive Questions in Surveys," *Psychological Bulletin*, 133, 859–883.
Vannieuwenhuyze, J., and Loosveldt, G. (2013), "Evaluating Relative Mode Effects in Mixed-Mode Surveys: Three Methods to Disentangle Selection and Measurement Effects," *Sociological Methods and Research*, 42, 82–104.
Vijverberg, W. P. M. (2011), "Testing for IIA with the Hausman-McFadden Test," Bonn, Germany. Available at https://docs.iza.org/dp5826.pdf
Villar, A., and Fitzgerald, R. (2017), "Using Mixed Modes in Survey Data Research: Results from Six Experiments," in *Values and Identities in Europe: Evidence from the European Social Survey*, ed. M. Breen, Oxford, England, United Kingdom: Routledge, pp. 273–310.
Voogt, R. J. J., and Saris, W. E. (2005), "Mixed Mode Designs: Finding the Balance between Nonresponse Bias and Mode Effects," *Journal of Official Statistics*, 21, 367–387.
Wang, J., and Donnan, P. T. (2001), "Propensity Score Methods in Drug Safety Studies: Practice, Strengths and Limitations," *Pharmacoepidemiology and Drug Safety*, 10, 341–344.
Weisberg, H. F. (2005a), "Measurement Error Due to Interviewers: The Debate over Interviewing Style," in *Total Survey Error Approach: A Guide to the New Science of Survey Research*, Chicago, IL: University of Chicago Press, pp. 45–71. https://doi.org/10.7208/chicago/9780226891293.003.0004
———. (2005b), *The Total Survey Error Approach: A Guide to the New Science of Survey Research*, Chicago, IL: University of Chicago Press. https://doi.org/10.7208/chicago/9780226891293.001.0001
Weitzen, S., Lapane, K. L., Toledano, A. Y., Hume, A. L., and Mor, V. (2005), "Weaknesses of Goodness-of-Fit Tests for Evaluating Propensity Score Models: The Case of the Omitted Confounder," *Pharmacoepidemiology and Drug Safety*, 14, 227–238.
West, B. T. (2011), "Paradata in Survey Research," *Survey Practice*, 4, 1–8.
Williams, D., and Brick, M. J. (2018), "Trends in U.S. Face-to-Face Household Survey Nonresponse and Level of Effort," *Journal of Survey Statistics and Methodology*, 6, 186–211.
Williams, J. (2017), *Community Life Survey Disentangling Sample and Mode Effects*, London, England, United Kingdom. Available at https://assets.publishing.service.gov.uk/media/5a82c535ed915d74e34037bd/Disentangling_sample_and_mode_effects_on_the_Community_Life_Survey_-_Nov_2017_revision.pdf

Journal of Survey Statistics and Methodology, 2024, 00, 1–26
https://doi.org/10.1093/jssam/smae017
Article