Propensity Score Weight Adjustment for Dual-Mode or Dual-Frame Longitudinal Surveys

D.M. Hajducek[*] C. Boudreau[*] M.E. Thompson[*]

Abstract
Dual-mode surveys are use to increase response rates and/or reduce costs. Similarly, dual-frame surveys can yield substantial cost savings when sampling rare populations, or when different survey modes are used in each frame and one mode is much cheaper than the other. In longitudinal surveys however, respondents with a given set of characteristics might experience both higher attrition and greater propensity to answer via one of the two modes; thus increasing undercoverage of these respondents as time goes on. Similarly, in dual-frame surveys, one frame might be incomplete, bias, suffer from undercoverage, or have higher nonresponse than the other. Our proposed post-survey weight adjustment method utilises propensity scores to adjust the weights of respondents from the mode/frame suffering from undercoverage/bias by making their probability distribution closer to that of respondents in the other mode/frame.

Key Words: Propensity score adjustment; dual-frame; dual-mode; post-survey weighting; longitudinal survey

1. Introduction

The use of propensity scores (PS) in post-survey weighting has been proposed by various authors. Lepkowski et al. (1989), Göksel et al. (1991) and Smith et al. (2000) used PS to decrease bias arising from non-response and partial response. Battaglia et al. (1995), Hoaglin & Battaglia (1996), Duncan & Stasny (2001) and Garren & Chang (2002) used PS to reduce bias from incomplete coverage in telephone surveys. PS have also been used to adjust for non-probability sampling in Web surveys; see Taylor (2000), Varedian & Försman (2002), Schonlau et al. (2004), Lee (2006), Terhanian et al. (2000) and Terhanian & Bremer (2000).

The basic idea behind most of these PS weight adjustments is that respondents with similar propensity scores should have, on average, equal weights. Our proposed PS weight adjustment method is based on the same idea, and ultimately aims to uniformize the weights of mode/frame A respondents with those of mode/frame B respondents who have similar propensity scores.

In this section, we briefly introduce dual-mode surveys (section 1.1), dual-frame surveys (section 1.2) and propensity scores (section 1.3). We then detail, in section 2, our proposed propensity score weight adjustment method, referred to as the PSWA method from this point onward. In section 3, we conduct a simulation study to investigate the effectiveness of the PSWA method, and compare it with the one proposed by Lee (2006). The PSWA method is then applied to data from the International Tobacco Control (ITC) Project; section 4.1 is concerned with the ITC Canada Survey and section 4.2 with the ITC Netherlands Survey. Lastly, section 5 contains a few concluding remarks.

1.1 Dual-mode surveys

In the simplest dual-mode survey design respondents are sampled from a single frame, but then complete the survey using different modes or methods of data collection (e.g., phone, mail, face-to-face or Web). This is generally done to increase response rates and/or

[*] Dept. of Statistics & Actuarial Science, University of Waterloo, Waterloo ON, N2J 4Z1, Canada

reduce costs. For example, respondents to the American Community Survey (ACS) are first contacted by mail. Those who do not respond are followed-up by phone and, ultimately, a personal visit from ACS interviewers. Since about 50% of respondents completed the survey by mail, this allowed the 2010 ACS to achieve a final response rate of 97.5% at a much lower cost than if all fieldwork had been done by phone or face-to-face. More complex dual-mode survey designs utilize multiple frames in addition to multiple modes. One such survey is the Canadian Community Health Survey (CCHS) where three sampling frames (area, list and RDD) are used in conjunction with two modes (CATI and CAPI) of data collection. More information on dual and multiple-mode surveys can be found in Groves (1989), de Leeuw (2005) and Brick & Lepkowski (2008).

The utility of a propensity score weight adjustment method for dual-mode surveys arises when the survey is longitudinal and when attrition and mode are correlated; in other words, when respondents with a given set of characteristics/covariates have both higher attrition in one mode (say B) than the other, and lower (or higher) propensity to answer via that mode. For example, consider a dual-mode longitudinal survey, where the attrition of respondents from low socioeconomic status (SES) is greater in the Web mode (Mode B) than it is in the phone mode (Mode A), and where low SES respondents are considerably more likely to complete the survey by phone. In such a context, the weights of the low SES respondents in mode B will be lower than they should be. Since these low SES respondents are better represented in the phone mode, a propensity score weight adjustment would bring the weights of their Web counterparts up by making them, on average, equal to the weights of mode A respondents with the same propensity score; thus reducing bias.

1.2 Dual-frame surveys

A dual-frame survey consists of two sampling frames (A and B) which, combined together, cover the target population. Some units can belong to both frames (i.e., $A \cap B \neq \varnothing$), but independent probability samples are taken from each frame. Dual-frame designs are used to reduce non-coverage when no single frame includes all the units of the target population, or when constructing and/or sampling from such a frame would be prohibitively expensive.

As noted by Hartley (1962), a dual-frame design can be used to reduce sampling error by allowing more observations to be sampled for the same cost. One example of this is in the sampling of rare populations. In such a case, a dual-frame design might consist of: (i) an RDD frame (say A) which is complete, but where the cost per unit is high as only a small proportion of respondents have the rare characteristic under study, and (ii) a list frame (say B) consisting solely of individuals with the rare characteristic. Since the list frame is already screened for the characteristic of interest, the cost per unit of sampling is much lower than in the RDD frame. However, the list frame is incomplete, and estimation based solely on this frame are be biased. Another example of cost saving resulting from a dual-frame design is where different survey modes are used in each frame, and where one mode is much cheaper than the other but associated to a frame that is incomplete. This is the case of the International Tobacco Control (ITC) Netherlands Survey (see section 4.2), where frame A (an RDD frame) is complete but expensive to sample from, and frame B (a frame based on a Web panel) is much less expensive to sample from but incomplete.

The scenarios in which a propensity score weight adjustment for dual-frame surveys is advantageous are akin to that of dual-modes surveys. Hence, a propensity score weight adjustment is advantageous when respondents with a given set of characteristics/covariates are better represented in one frame (say A) than in the other (say B). The dual-frame analogue of the low SES example of section 1.1, would consist of a Web frame (frame B), where low SES people are underrepresented, and a representative RDD frame (frame A).

As in section 1.1, the weights of the low SES respondents in frame B will be lower than they should, and a propensity score weight adjustment would bring the weights of those respondents closer to their frame A counterparts; thus reducing bias.

1.3 Fundamental property of propensity scores

Propensity score methodology was introduced by Rosenbaum & Rubin (1983, 1984). In a nutshell, their methodology attempts to eliminate or reduce bias arising from the lack of randomization in observational studies; thus allowing the estimation of treatment effects as with randomized clinical trials (RCT's). The propensity score is defined as the conditional probability of treatment vs. control given a vector of observed covariates x; i.e.,

$$\mathrm{ps}(x) = \Pr(Z = 1|x) \ , \tag{1}$$

where $Z = 1$ if the individual is assigned to the treatment group and $Z = 0$ if the individual is assigned to the control group.

In their 1983 paper, Rosenbaum & Rubin show that (1) is a "balancing score" which means that, conditional on $\mathrm{ps}(x)$, Z and x are independent; i.e., $Z \perp x \mid \mathrm{ps}(x)$. More importantly, they show that under suitable conditions the difference between treatment and control means at any given value of the balancing score is an unbiased estimate of the average treatment effect at that value of the balancing score. More loosely, in subjects sharing the same propensity score, treatment allocation is independent of the observed variables x. Therefore, creating strata of subjects matched on their propensity scores allows one to replicate, conditional on observed variables, the design of a RCT.

In our PSWA method, Z does not correspond to treatment vs. control, but rather to mode A vs. mode B (section 2.1) or frame A vs. frame B (section 2.2).

2. Proposed weight adjustment method

2.1 Dual-mode surveys

Let mode B be the survey mode with the higher attrition and ensuing bias, and mode A the one without such a problem or where that problem is much less severe (i.e., the gold standard). In addition, let $Z_i = I(i^{\text{th}}$ respondent used mode $B)$ and w_i be the sampling weight of that respondent prior to the PSWA. Note that some post-survey weighting techniques (e.g., calibration or raking) might have been applied to the w_i's to reduce bias and other survey errors. Lastly, let $S^A = \{i \in S \mid Z_i = 0\}$ and $S^B = \{i \in S \mid Z_i = 1\}$; thus, S^A is the sub-sample of respondents who completed the survey via mode A (similarly for S^B), and $S = S^A \cup S^B$.

If a stratified design with different sampling fractions is used to sample respondents, the weights will vary considerably from stratum to stratum. The same situation is also likely to arise if post-stratification is used to compute the w_i's. In such situations, it is best to perform the PSWA on a per (post-)stratum basis. To this end, let S_h^A be the subset of (post-)stratum h $(h = 1, \ldots, H)$ respondents who completed the survey via mode A (similarly for S_h^B), and $S_h = S_h^A \cup S_h^B$. Our proposed PSWA method proceeds as follows:

Step 1: For (post-)stratum $h = 1$, fit a PS model to respondents from stratum S_h with Z_i as the dependent variable and x_i as the vector of covariates. Make sure that the model is properly balanced. Note that w_i should not be an element of x_i.

Step 2: Compute $\widehat{\mathrm{ps}}_i$ for $i \in S_h$ using the PS model developed in step 1.

Step 3: Divide respondents into Q percentiles based on their $\widehat{\text{ps}}_i$'s; thus creating $S_{h,1},\ldots,S_{h,Q}$, where $S_{h,q} = \{i \in S_h \mid \widehat{\text{ps}}_{(q-1)} < \widehat{\text{ps}}_i \le \widehat{\text{ps}}_{(q)}\}$ and $\widehat{\text{ps}}_{(q)}$ is the $(100\,q/Q)^{\text{th}}$ percentile.

Step 4: Compute the adjustment factors

$$f_{h,q} = \left(\frac{1}{n^A_{h,q}} \sum_{i\in S_{h,q}} (1-Z_i)\, w_i\right) \Bigg/ \left(\frac{1}{n^B_{h,q}} \sum_{i\in S_{h,q}} Z_i\, w_i\right) = \frac{\bar{w}^A_{h,q}}{\bar{w}^B_{h,q}}\ ,$$

for $q = 1,\ldots,Q$ and where $n^A_{h,q} = \sum_{i\in S_{h,q}} (1-Z_i)$ and $n^B_{h,q} = \sum_{i\in S_{h,q}} Z_i$.

Hence, $f_{h,q}$ is simply the average weight of mode A respondents in the (post-)stratum h/percentile q class divided by the corresponding average weight for mode B respondents.

Step 5: Compute the PS adjusted weights for the $n^B_{h,q}$ mode B respondents that are in $S_{h,q}$ by multiplying their w_i weights by $f_{h,q}$, yielding $w^{\text{ps}}_i = w_i\, f_{h,q}$.

Step 6: Repeat steps 1–5 for (post-)strata $h = 2,\ldots,H$.

Step 7: Re-calibrate and/or re-scale weights as required.

A few remarks are in order:

1. Our proposed method is slightly different from the one proposed by Lee (2006), where

$$f'_{h,q} = \frac{\sum_{i\in S_{h,q}} (1-Z_i)\, w_i \Big/ \sum_{i\in S_h} (1-Z_i)\, w_i}{\sum_{i\in S_{h,q}} Z_i\, w_i \Big/ \sum_{i\in S_h} Z_i\, w_i}\ ;$$

   thus, $f'_{h,q} = f_{h,q}\,(n^A_{h,q}\bar{w}^B_h)/(n^B_{h,q}\bar{w}^A_h)$ where $\bar{w}^A_h$ is the average weight of mode A respondents in (post-)stratum h (similarly for $\bar{w}^B_h$). We will refer to Lee's method as the Alt-PSWA method, and compare it to our proposed method in sections 3 and 4. Alternative versions for the computation of $f_{h,q}$ are given in Lee & Valliant (2008).

2. The most commonly used PS model is logistic regression, in which case logit(ps(x_i)) $= \alpha + \beta' x_i$. However, models such as GLM with a log-log link are also possible, and have been used to adjust weights for non-coverage in telephone surveys; see Garren & Chang (2002).

3. If a covariate is a strong predictor of Z_i, it can be advantageous to use that covariate as an extra post-stratification variable instead of a covariate in the PS model. For example, if gender is a strong predictor of mode, using $2H$ post-strata is likely to yield better results than using H (post-)strata. In sections 3.2 and 4.1, the covariate indicating if the respondent received an email invite to complete the survey via mode B is used in such a way. This obviously works best for categorical covariates, or covariates that can discretized into a few groups.

4. An obvious and important limitation of our proposed PSWA method is that it can only adjust for variables that were measured and included in the PS model and/or post-stratification.

5. In most single frame dual-mode surveys, the w_i's from both modes are pooled together for calibration/raking. Hence, in situations were the weights are re-calibrated or re-scaled (see step 7), the weights of respondents from mode A will also be modified. On the other hand, in dual-frame surveys, where calibration/raking is generally done on a per frame basis, the weights of respondents from frame A will not be modified by our proposed PSWA method.

6. If similar sampling fractions are used across strata, the w_i's will also be similar and those strata can be combined when performing the PSWA. Pooling will make the $f_{h,q}$'s, and thus the w_i^{ps}'s more stable. This strategy was used in sections 3.2 and 4.1.

2.2 Dual-frame surveys

The PSWA method for dual-mode surveys described above can also be applied to dual-frame surveys. One can follow the exact same steps as in section 2.1 (with the obvious exception that modes A and B are to be replaced with frames A and B). However, the context in which the method is applied varies, and is the topic of this section.

Let frame A be the one that is complete but expensive (i.e., the gold standard), and frame B the one that is less expensive to sample from but incomplete or biased. As in section 2.1, let Z_i be the indicator which is equal to 1 if the i^{th} respondent was sampled via frame B (and 0 otherwise), and w_i be the sampling weight of that respondent prior to the PSWA. Using sampling design $p^A(\cdot)$, a sample $S^A \subseteq U^A$ is drawn from frame A (similarly for S^B). In most surveys, one or both sampling designs will be stratified. Hence, let $U_1, \ldots, U_H$ be the strata obtained from the combination of the H^A strata from sampling design $p^A(\cdot)$ and the H^B strata from sampling design $p^B(\cdot)$. For example, if A is stratified by gender and B is stratified by age groups (18–24, 25–49 and 50+), then the U_h's would be the resulting 6 gender/age strata. Lastly, let $S_h^A = S^A \cup U_h$ (similarly for S_h^B) and $S_h = S_h^A \cup S_h^B$. Steps 1–7 of section 2.1 can then be applied.

3. Simulation study

The International Tobacco Control (ITC) Project carries out prospective longitudinal surveys of smokers in more than 20 countries, and aims to study the impact of national-level tobacco control policies and the behaviour/psychology of smoking. The ITC Canada (CA) Survey forms part of the larger ITC Four Country Survey, which started with Wave 1 in late 2002, in which around 2000 smokers are interviewed in each of Canada, United Kingdom, United States and Australia.

Respondents who completed the survey in Wave 7 (October – February 2009) were invited to participate in Wave 8 and interviewed in July – December of 2010. Invitations were sent by regular mail to participate by telephone or on-line. Email invitations were also sent to respondents with a valid email address.

The purpose of this simulation is to assess the effectiveness of the PSWA method by sampling from a generated population that is based on data from the ITC CA Wave 7 survey and the response patterns of this sample in Wave 8, where data collection was carried out via Computer Assisted Telephone/Web Interviews (CATI/CAWI). This survey is therefore an example of the simple case of dual-mode surveys discussed in section 1.1, where respondents are sampled from a single frame. The sample from Wave 8 consists of 1374 smokers, among which 747 responded by CATI (Mode A) and 627 responded by CAWI (Mode B). After discarding missing values, the sample of respondents was proportionally inflated 20 times, resulting in a population of 30582 individuals.

As with the ITC CA Survey, our population was divided into 14 strata (Thompson et al. (2006)). A stratified SRS sample was obtained through the proportional allocation method. Samples obtained in each iteration had a size of about $n = 2000$ after attrition. Sample retention and Web responses in the generated samples were simulated by binomial models using retention rates and Web propensity scores based on the ITC CA Waves 7 and 8 data. The following sub-sections describe the simulation of retention and Web responses.

Calibration to population smoking prevalence by age/sex/region was performed on the initial weights of individuals that were retained in the sample. Then PSWA adjustment was applied to the calibrated weights by groups defined by whether or not the person was invited by email to participate in the Web survey and were calibrated again.

The PSWA method was assessed on its own and also compared to the alternative Alt-PSWA proposed by Lee (2006) and briefly discussed in section 2. This was done through examination of the mean and coefficient of variation (CV) of the adjusted weights and through comparison of survey estimates produced with each set of weights. We present estimates (and mean squared error - MSE) of the average number of cigarettes smoked per day (CPD) and the proportion of individuals who had intentions to quit smoking (QUIT). Another way in which we compare the adjusted and unadjusted weights is by examining how highly they are correlated.

3.1 Simulation of retention

Retention rates used in the simulation were obtained by selecting a logistic model on data from Wave 7 of the ITC CA Survey. Results were obtained after a step-wise selection procedure that involved 20 variables: time in sample, smoking status, demographic (age, sex marital status, income, education, ethnicity), health related (general self assessed health, doctor visits, drinking habits and depression), exposure to anti-tobacco information (heard news about smoking, noticed anti-smoking information), quitting related (intention, sure would succeed), quitting reasons (health, medical advise, warning labels), salience and effects of warning labels (noticed, stopped from smoking) and knowledge about nicotine causing most cancers.

The income variable used in the selected retention model above was generated such that it would have a greater effect on both attrition and Web response propensity, hence triggering a greater need for the PSWA. Participants in the ITC CA sample were classified into two groups defined by whether or not the person was invited by email to participate in the Web survey. The main idea was to randomly assign low income to a reasonably large group of participants who did not receive an email invitation, especially those who were current smokers at the time of Wave 8 (non-smokers are identified as persons who quit smoking after they were recruited on Wave 1). Low, moderate and high income were simulated as observations from a multinomial random variable with probabilities (0.12, 0.12, 0.75) for both smokers and non-smokers in the group that was invited by email. For those who were not invited by email, the simulation probabilities for low, moderate and high income were (0.86, 0.07, 0.07) for smokers and (0.60, 0.20, 0.20) for non-smokers.

Retention was simulated using a binomial model based on the retention rates in each generated sample leading to a mild, moderate or severe level of attrition on the Web unsavvy sampled individuals (i.e., prone to respond by phone, defined by an estimated propensity of responding by Web ≤ 0.6). This is translated into various degrees of sample miss-represenation of the Web savvy population. The degree to which individuals were miss-represented in the sample is dictated by the magnitude of the induced effects of age and income in the probability of retention.

To summarize, income has been simulated such that low levels are associated with the

Web un-savvy group, which is more likely to respond by phone. This results in younger people with low income who are prone to respond by phone to be more likely to become lost to follow up. The various degrees of induced attrition are referred to as simulation scenarios I, II and III, which correspond to an average of 5% under-representation, and 12% and 27% over-representation of the Web savvy population, respectively.

3.2 Simulation of Web responses

Web responses were simulated according to a binomial model based on propensity scores that were estimated from fitting a logistic model to the probability of responding by Web (vs. phone). Web responses were simulated in order for their relationship to the simulated income to remain consistent with the Waves 7 and 8 of the ITC CA Survey. Modeling was done separately by groups, determined by whether the person was invited by email to participate in the Web survey, using variables that involve demographic characteristics, smoking status (smoker or non-smoker), health (whether feeling down with little interest and/or diagnosed with depression, whether visited a doctor, and self assessed health), and frequency of alcohol drinking.

3.3 Results

As mentioned earlier, we denoted by A and B the sets of participants who responded by phone and Web, respectively. The PSWA and Alt-PSWA methods were implemented as in section 2. Since this is a self-weighing design, no stratification was used in the adjustments.

Table 1 shows the results obtained across the three simulation scenarios over 1000 iterations. It shows the sample sizes for the telephone and Web respondents n_q^A and n_q^B for quartile groups $q = 1, 2, 3, 4$; the means (and CV's) of the calibrated weights before the adjustment for telephone and Web respondents, $\bar{w}_q^A$ and $\bar{w}_q^B$; the means of the weights adjusted by PSWA and Alt-PSWA, $\bar{w}_q^{\text{ps}}$ and $\bar{w}_q^{\text{ps}'}$; and the means after the adjustment and re-calibration, $\bar{w}_q^{\text{ps}+c}$ and $\bar{w}_q^{\text{ps}'+c}$. The adjustment factors under PSWA and Alt-PSWA are denoted by f_q and f'_q, respectively.

The values of the PSWA adjustment factor f_q indicate that the amount of relative discrepancy between the means of the weights from phone and Web respondents (hence, the need for the PSWA) is quite low, ranging from 0.91 to 1.07 across the three scenarios, but remaining stable around the value of one. The adjustment factor under Alt-PSWA f'_q on the contrary, has a wider range going from 0.50 to 2.01.

The dispersion of the weights is not compromised by the PSWA method (before and after re-calibration), as the overall CV's remain fairly unchanged compared to those of the initial calibrated weights. The CV's under the Alt-PSWA method however, show substantial increment. The mean of the weights after the PSWA and before re-calibration $\bar{w}_q^{\text{ps}}$ is quite stable in the sense that it remains fairly close to the mean of the weights of phone respondents $\bar{w}_q^A$, while the means under the Alt-PSWA $\bar{w}_q^{\text{ps}'}$ are farther apart.

Table 2 shows the bias (and MSE) given by the PSWA and Alt-PSWA for each simulation scenario. The columns refer to the weights used for estimation: initial (stratified SRS), calibrated before the adjustments, and re-calibrated after the adjustments. Although the effect on CPD was greater, an overall pattern of increased bias across simulation scenarios can be seen for both CPD and QUIT estimates under the initial weights. For CPD, calibration of the weights gives a substantial reduction of bias with respect to the initial weights, with the PSWA giving further and systematic reduction by 1, 3 and 7% across scenarios. In

Table 1: Summary of the distribution of unadjusted and PS adjusted weights over 1000 iterations.

				Calibrated		PSWA			Alt-PSWA		
	Quartile	n_q^A	n_q^B	$\bar{w}_q^A$	$\bar{w}_q^B$	f_q	$\bar{w}_q^{\text{ps}}$	$\bar{w}_q^{\text{ps}+c}$	f_q'	$\bar{w}_q^{\text{ps}'}$	$\bar{w}_q^{\text{ps}'+c}$
I	1	350	153	0.977	1.013	0.965	0.977	0.978	2.014	2.040	1.944
				(0.148)	(0.165)		(0.165)	(0.167)		(0.165)	(0.159)
	2	288	211	0.986	0.995	0.991	0.986	0.984	1.227	1.221	1.223
				(0.133)	(0.140)		(0.140)	(0.141)		(0.140)	(0.163)
	3	240	259	1.008	0.999	1.009	1.008	1.005	0.847	0.846	0.866
				(0.136)	(0.135)		(0.135)	(0.136)		(0.135)	(0.160)
	4	175	323	1.036	1.010	1.027	1.036	1.033	0.505	0.509	0.533
				(0.138)	(0.135)		(0.135)	(0.135)		(0.135)	(0.165)
	Overall	1053	947	0.996	1.004	-	1.008	1.006	-	1.004	1.003
				(0.142)	(0.143)		(0.144)	(0.144)		(0.552)	(0.518)
II	1	361	142	0.953	1.023	0.933	0.953	0.955	1.995	2.039	1.935
				(0.178)	(0.189)		(0.189)	(0.193)		(0.189)	(0.182)
	2	299	201	0.969	0.988	0.981	0.969	0.967	1.226	1.211	1.209
				(0.166)	(0.174)		(0.174)	(0.175)		(0.174)	(0.198)
	3	251	249	1.008	0.992	1.016	1.008	1.004	0.858	0.851	0.87
				(0.181)	(0.177)		(0.177)	(0.177)		(0.177)	(0.206)
	4	188	311	1.083	1.031	1.05	1.083	1.075	0.529	0.546	0.573
				(0.191)	(0.192)		(0.192)	(0.190)		(0.192)	(0.229)
	Overall	1099	902	0.992	1.009	-	1.017	1.013	-	1.009	1.007
				(0.186)	(0.186)		(0.192)	(0.191)		(0.546)	(0.511)
III	1	373	130	0.937	1.029	0.911	0.937	0.936	2.007	2.064	1.949
				(0.225)	(0.240)		(0.240)	(0.245)		(0.240)	(0.229)
	2	311	188	0.959	0.967	0.991	0.959	0.951	1.256	1.214	1.205
				(0.222)	(0.227)		(0.227)	(0.228)		(0.227)	(0.244)
	3	260	239	1.017	0.979	1.039	1.017	1.006	0.87	0.851	0.868
				(0.242)	(0.238)		(0.238)	(0.237)		(0.238)	(0.268)
	4	196	302	1.125	1.049	1.074	1.125	1.109	0.534	0.559	0.59
				(0.245)	(0.255)		(0.255)	(0.253)		(0.255)	(0.300)
	Overall	1140	860	0.994	1.009	-	1.03	1.02	-	1.009	1.005
				(0.244)	(0.247)		(0.256)	(0.254)		(0.571)	(0.535)

contrast, bias under Alt-PSWA has a substantial increment. In the case of QUIT, the PSWA offers no improvement in bias while Alt-PSWA shows a slight increment.

In terms of the variability of the estimates, the PSWA keeps a steady MSE compared to the estimates under the calibrated weights, while Alt-PSWA shows a slight increase.

The correlation of the PSWA weights with the initial calibrated weights ranges from 0.992 to 0.987 across scenarios, while that of Alt-PSWA ranges from 0.374 to 0.598; thus indicating that the former weights are much closer to the original ones.

Table 2: Bias (MSE) of CPD and QUIT over 1000 iterations.

				Re-calibrated	
Variable		Initial	Calibrated	PSWA	Alt-PSWA
CPD	I	0.129 (0.0540)	0.118 (0.0508)	0.118 (0.0505)	0.122 (0.0598)
	II	0.254 (0.1050)	0.127 (0.0557)	0.123 (0.0549)	0.155 (0.0674)
	III	0.348 (0.1576)	0.115 (0.0501)	0.107 (0.0484)	0.160 (0.0664)
QUIT	I	-0.003 (12x10^{-3})	0.002 (12x10^{-3})	0.002 (12x10^{-3})	0.004 (15x10^{-3})
	II	-0.007 (15x10^{-3})	0.003 (11x10^{-3})	0.003 (11x10^{-3})	0.005 (14x10^{-3})
	III	-0.010 (20x10^{-3})	0.004 (12x10^{-3})	0.004 (12x10^{-3})	0.005 (15x10^{-3})

True values in generated population: 16.84 for CPD and 77.1% QUIT.

4. Applications to ITC Surveys

4.1 ITC Canada Survey

This section provides results from implementing the PSWA method on data from Wave 8 of the ITC CA Survey. The effects of the PSWA and Alt-PSWA methods are assessed by comparing: (i) the means (and CV's) of the adjusted and unadjusted weights, (ii) biases and standard errors (SE's) of the estimates of CPD and QUIT, (iii) the correlation between unadjusted and adjusted weights, and (iv) non-parametric density estimates of the distribution of the weights.

Table 3 shows results on the mean (and CV) of initial weights of telephone and Web respondents: $\bar{w}^{A}_{h,q}$ and $\bar{w}^{B}_{h,q}$, as well as the adjustment factor and mean (and CV) under PSWA: f_{hq}, $\bar{w}^{ps}_{h,q}$, and Alt-PSWA: f'_{hq}, $\bar{w}^{ps'}_{h,q}$. The subscripts h, q indicate strata for email invitation and percentile groups, respectively. Since recalibration was carried out by pooling weights from both modes together, the PS adjustments not only gave new values for Mode B but also for Mode A. Table 4 shows the means and CV's of the re-calibrated weights under the PSWA for modes A and B: $\bar{w}^{A+c}_{h,q}$, $\bar{w}^{\text{ps}+c}_{h,q}$; and similarly, under Alt-PSWA: $\bar{w}^{A+c'}_{h,q}$ $\bar{w}^{\text{ps}+c'}_{h,q}$.

Adjustment factors under the Alt-PSWA, f'_{hq}, show more variability across percentile groups compared to those under the PSWA f_{hq}. This is consistent with the simulation results in Table 1. Adjustment factors under the PSWA range from 1.15 to 1.21 and 0.89 to 1.07 for the email invitation "Yes" and "No" groups, while those from Alt-PSWA range from to 0.62 to 2.08 and from 0.65 to 2.94.

The CV's of the weights across quartiles and tertiles in Table 3 are unchanged compared to those from Mode B for both PSWA and Alt-PSWA. Overall though, Alt-PSWA gives higher values than PSWA: 1.052 vs. 0.794 and 0.829 vs. 0.711, for email invitation groups "Yes" and "No", respectively. Similarly, Table 4 shows that, re-calibration after combining the weights from email invitation groups, gives similar CV values under both PSWA and

Alt-PSWA for both Modes A and B across quartiles, but a higher CV for Mode B under Alt-PSWA.

Comparisons of survey estimates will show only whether the estimates are similar or not. In the latter case there is no way of knowing which should be preferred; however, in the former case it is still possible to select the method that gives the most accurate estimate. The estimates of CPD in this example are quite similar, being under the unadjusted, PSWA and Alt-PSWA re-calibrated weights: 15.949, 15.904 and 15.839, respectively. In the same order, the SE's are 0.342, 0.341 and 0.351. In consistency with the effects of the PSWA, it gives (slightly) lower variability compared to Alt-PSWA. Therefore, for CPD the difference between methods is negligible and so is the case for the QUIT variable, with estimates of 0.759, 0.751 and 0.751 (with SE's 0.016, 0.017 and 0.017).

Table 3: Summary of the distribution of unadjusted and PS adjusted weights, ITC CA Survey Wave 8.

Invite	Quartile/			Calibrated		PSWA		Alt-PSWA	
Email	Tertile	$n^A_{h,q}$	$n^B_{h,q}$	$\bar{w}^A_{h,q}$	$\bar{w}^B_{h,q}$	$f_{h,q}$	$\bar{w}^{\text{ps}}_{h,q}$	$f'_{h,q}$	$\bar{w}^{\text{ps}'}_{h,q}$
Yes	1	18	24	1.707	1.414	1.208	1.707	1.945	2.750
				(0.938)	(0.498)		(0.498)		(0.498)
	2	42	50	1.120	0.973	1.151	1.120	2.077	2.020
				(0.797)	(0.687)		(0.687)		(0.687)
	3	41	110	1.220	1.042	1.171	1.220	0.937	0.977
				(0.770)	(0.898)		(0.898)		(0.898)
	4	56	228	1.064	0.810	1.182	1.064	0.624	0.561
				(0.673)	(0.780)		(0.780)		(0.780)
	Overall	157	412	1.193	0.977		1.150		0.977
				(0.810)	(0.794)		(0.794)		(1.052)
No	1	123	19	0.800	0.746	1.073	0.800	2.942	2.194
				(0.612)	(0.570)		(0.570)		(0.570)
	2	200	52	0.884	0.829	1.068	0.884	1.739	1.440
				(0.705)	(0.688)		(0.688)		(0.688)
	3	248	144	1.0652	1.189	0.896	1.065	0.653	0.777
				(0.743)	(0.716)		(0.716)		(0.716)
	Overall	571	215	0.945	1.063		0.998		1.063
				(0.726)	(0.735)		(0.711)		(0.829)

The correlation between the unadjusted (first time calibrated) and re-calibrated weights after the PSWA is 0.98 while the correlation between the unadjusted and Alt-PSWA re-calibrated weights is 0.86. This indicates that the former weights are closer to the initial, calibrated weights.

Figure 1(a) shows non-parametric density estimates of the distributions of initial and re-calibrated weights after adjustments, by mode. For the Web mode, the CAWI PSWA distribution is closer to the reference (CATI) than CAWI Alt-PSWA, while for the phone mode it is CATI Alt-PSWA the one that is substantially closer. It seems that in this case the PSWA method keeps the CATI PSWA and CAWI PSWA at similar but moderate distances from CATI, while Alt-PSWA is more extreme, keeping CATI Alt-PSWA very close to the reference but leaving CAWI Alt-PSWA the farthest away.

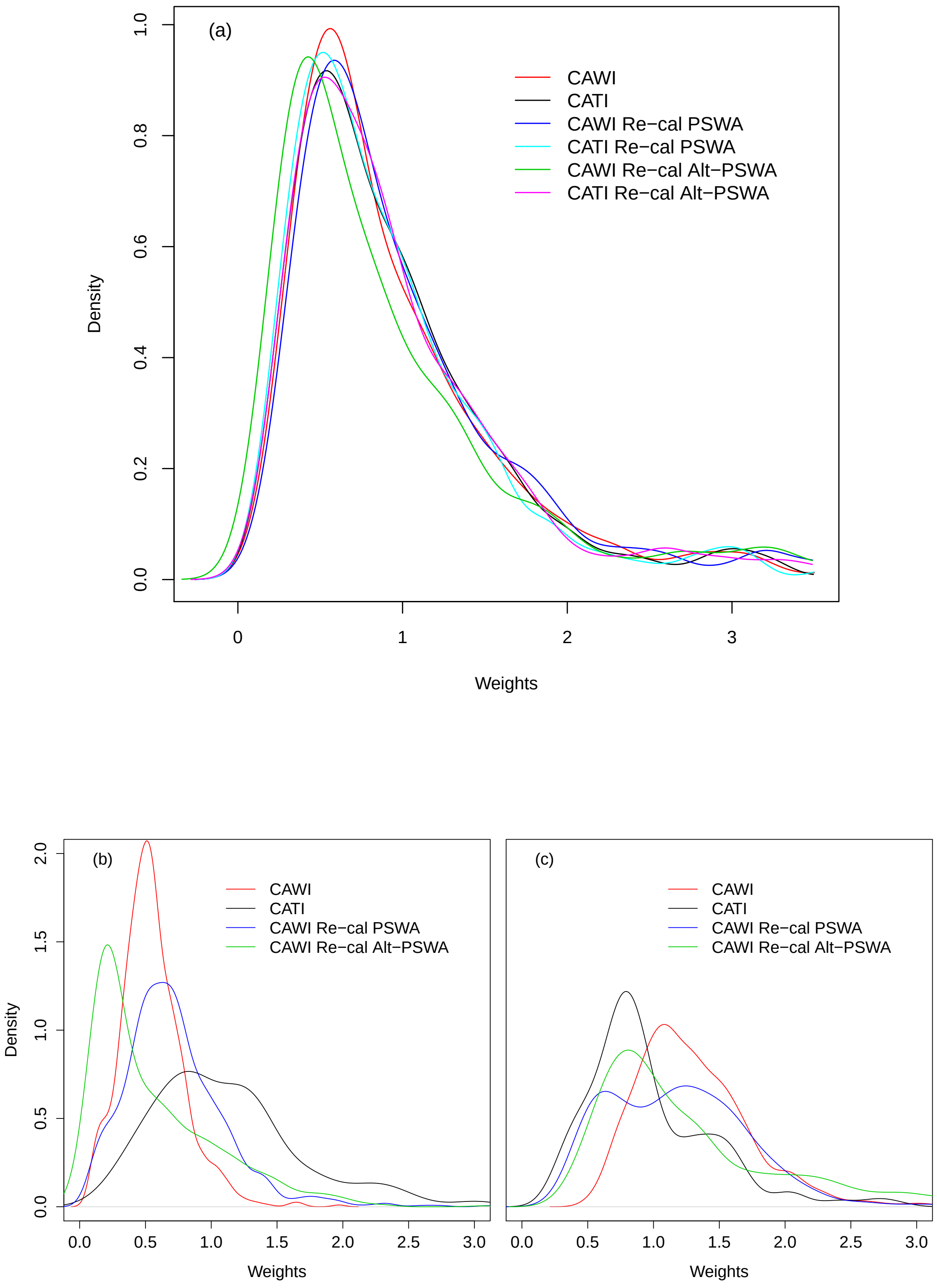

Figure 1: Density estimates of distribution of weights from: (a) ITC CA Survey Wave 8, (b) ITC NL Survey Wave 1, Age ≤ 30, (c) ITC NL Survey Wave 1, Age > 30.

Table 4: Mean (CV) by mode of ITC CA Survey Wave 8 re-calibrated weights.

			PSWA		Alt-PSWA	
Quartile	n_q^A	n_q^B	$\bar{w}_{h,q}^{A+c}$	$\bar{w}_{h,q}^{\text{ps}+c}$	$\bar{w}_{h,q}^{A+c'}$	$\bar{w}_{h,q}^{\text{ps}'+c}$
1	141	43	0.882	1.246	0.861	2.314
			(0.848)	(0.631)	(0.806)	(0.531)
2	242	102	0.891	0.958	0.899	1.684
			(0.737)	(0.698)	(0.712)	(0.711)
3	289	254	1.045	1.088	1.125	0.894
			(0.749)	(0.816)	(0.778)	(0.857)
4	56	228	1.015	1.016	1.098	0.586
			(0.665)	(0.778)	(0.661)	(0.796)
Overall	728	627	0.960	1.052	0.997	1.008
			(0.761)	(0.775)	(0.768)	(0.947)

$\bar{w}_{h,q}^{A+c}$ and $\bar{w}_{h,q}^{A+c'}$ are the means of re-calibrated weights from phone respondents after the PSWA and Alt-PSWA methods, respectively.

4.2 ITC Netherlands Survey

Also a part of the ITC Project, the ITC Netherlands (NL) Survey is a prospective longitudinal study of about 2200 smokers. Although fieldwork for Wave 6, the latest of the ITC NL Survey was completed in June 2012, this example is concerned with Wave 1 which ran through March–April 2008.

The ITC NL Survey uses a dual-frame sampling design, and fieldwork was conducted by the Dutch survey firm TNS NIPO. Frame A consists of a traditional stratified RDD design, and slightly over 400 respondents were interviewed from that frame using computer assisted telephone interviews (CATI). Frame B is the Web portion of the TNS NIPObase, which consists of over 140000 respondents who have agreed to participate in TNS NIPO research on a regular basis. After stratifying on age ($\leq$ 30 vs. $>$ 30), slightly over 1800 respondents were interviewed from that frame using computer assisted Web interviews (CAWI). Two important points must be made about the design of the ITC NL Survey. First, the dual RDD/Web frame was explicitly conceived with the aim of using the RDD frame to adjust for non-coverage and bias in the Web frame. Second, members of the TNS NIPObase were randomly selected (mostly by mail and RDD), and are thus not a panel of self-selected volunteers (Boudreau (2009)). Hence, the design of the ITC NL Survey is quite different from those of volunteer panel surveys described in section 1.

As Table 3, Table 5 shows results on the mean (and CV) of initial weights of telephone, adjustment factors, and mean (CV) of adjusted weights of Mode B. The subscripts h, q indicate strata $h = 1, 2$ for age and tertile groups, respectively. In contrast with the ITC CA Survey which illustrates a dual-mode survey (within a single frame), this example illustrates a dual-frame, therefore only weights within Frame B were adjusted and re-calibrated and the weights from Frame A remained unchanged. Means and CV's for re-calibrated weights after the adjustments are shown in Table 6.

Overall CV's under the PSWA (column $\bar{w}_{h,q}^{\text{ps}}$) are not too far from the CV of initial weights by Web ($\bar{w}_{h,q}^{B}$), while the Alt-PSWA counterparts show quite an increase from 0.433 to 0.856 for ages below 30 and from 0.409 to 0.649 for ages above. Similarly with re-calibrated weights by mode in Table 6.

The values of the PSWA factor $f_{h,q}$ in this example indicate a greater need for the adjustment, compared to the ITC CA Survey example. In consistency with the previous results however, the PSWA factor has a lesser variation across tertile groups compared to the Alt-PSWA $f'_{h,q}$. It ranges from 1.66 to 2.55 and from 0.69 to 0.76 for the age below and above 30 groups, while the latter ranges from 0.36 to 1.98 and from 0.67 to 1.67. Also in consistency with previous results, the CV's given by the PSWA do not change substantially from those of the the initial, calibrated weights, while those from the Alt-PSWA method are substantially higher.

Table 5: ITC NL Survey Wave 1 weight distribution summary for stratified adjustments with $h = 1, 2$ for age group, before re-calibration (CV's in parentheses)

				Initial		PSWA		Alt-PSWA	
Age	Tertile	$n^A_{h,q}$	$n^B_{h,q}$	$\bar{w}^A_{h,q}$	$\bar{w}^B_{h,q}$	$f_{h,q}$	$\bar{w}^{ps}_{h,q}$	$f'_{h,q}$	$\bar{w}^{ps'}_{h,q}$
≤ 30	1	43	194	0.995	0.598	1.663	0.995	1.982	1.185
				(0.508)	(0.335)		(0.335)		(0.335)
	2	18	212	1.372	0.538	2.551	1.372	1.164	0.626
				(0.533)	(0.419)		(0.419)		(0.419)
	3	12	389	1.177	0.536	2.195	1.177	0.364	0.195
				(0.257)	(0.486)		(0.486)		(0.486)
	Overall	73	795	1.118	0.552		1.185		0.552
				(0.502)	(0.433)		(0.457)		(0.856)
> 30	1	152	271	1.013	1.458	0.695	1.013	1.671	2.435
				(0.499)	(0.369)		(0.369)		(0.369)
	2	91	324	0.883	1.304	0.677	0.883	0.815	1.063
				(0.509)	(0.400)		(0.400)		(0.400)
	3	88	430	1.000	1.311	0.763	1.000	0.669	0.877
				(0.485)	(0.437)		(0.437)		(0.437)
	Overall	331	1025	0.974	1.348		0.967		1.348
				(0.500)	(0.409)		(0.413)		(0.649)

Table 6: Mean (CV) by mode of ITC NL Wave 1 re-calibrated weights.

				PSWA	Alt-PSWA
Tertile	n^A_q	n^B_q	$\bar{w}^A_q$	$\bar{w}^{\text{ps}+c}_q$	$\bar{w}^{\text{ps}'+c}_q$
1	195	465	1.009	1.003	1.851
			(0.499)	(0.579)	(0.530)
2	109	536	0.964	1.016	0.903
			(0.554)	(0.518)	(0.495)
3	100	819	1.021	0.987	0.580
			(0.459)	(0.635)	(0.825)
Overall	404	1820	1.000	1.000	1.000
			(0.503)	(0.587)	(0.821)

The CPD estimates produced by initial, PSWA and Alt-PSWA are: 15.427, 15.364 and 15.604 with SE's 0.248, 0.247 and 0.282, respectively. This again reinforces the notion discussed so far, about the PSWA giving less variability compared to Alt-PSWA. Estimates for QUIT are somewhat dissimilar, but with the SE's between adjustment methods is unchanged (in the same order): 0.812, 0.813 and 0.820 with SE's 0.010, 0.011 and 0.011.

Plots (b) and (c) in Figure 1 show non-parametric density estimates of the distributions weights by adjustment method and mode, for the two age groups ≤ 30 and > 30, respectively. In both cases, the CAWI PSWA distribution is closer to the reference CATI than CAWI Alt-PSWA.

5. Concluding Remarks

The simulations of section 3 and examples of section 4 indicate that the PSWA method is effective in terms of making the weights of mode/frame B respondents "closer" to those of mode/frame A respondents. The density estimates of Figure 1 best illustrate this. Moreover, this was achieved while not compromising the variability of the weights. Since more variable weights result in the loss of precision, this is an important consideration. The PSWA weights are also very highly correlated with the initial calibrated weights, whereas the Alt-PSWA method yielded weights that are somewhat different.

In terms of bias, the PSWA method yielded modest reduction for descriptive statistics CPD and QUIT (there was also some gains in terms of MSE for CPD). One important reason why these these gains in bias (MSE) are marginal is the design of our simulation study. As describe in section 1.1, our PSWA method works best when the proportion of respondents with the given set of characteristics/covariates (e.g., low SES) is high in the lower PS quartiles and low in the upper PS quartiles. Though our intent was to simulate income for respondent who did not receive an email invite to achieve that, the proportion of low income respondents remained somewhat spread out over the quartiles. Hence, it is not surprising that the PSWA method did not achieve important reduction in bias.

Lastly, we considered two ways of grouping respondents in quartiles for the simulations of section 3. The first consisted in pooling the propensity scores of respondents in the two sets of quartiles that correspond to the two models given by the post-stratification groups (e.g., email invitation); so that the propensity scores within each quartile may belong to both models. The other consisted in performing the adjustments in the set of quartiles produced by each model separately. Since both approaches gave similar results, we showed only results under the former. These and other approaches on quartile grouping have to our knowledge not yet been explored. With real data, the choice of percentile grouping (pooled vs. separate, and quartiles vs. tertiles) may depend on practical issues during the modeling stage, such as the number of respondents in each group, and balance of the PS model. It is recommended to assess the different options at hand on a case by case basis.

References

Battaglia, M. P., Malec, D. J., Spencer, B., Hoaglin, D. C. & Sedransk, J. (1995), Adjusting for noncoverage of nontelephone households in the National Immunization Survey, *in* 'Proceedings of the Section on Survey Research Methods', ASA, pp. 678–683.

Boudreau, C. (2009), Construction and use of sampling weights for the International Tobacco Control (ITC) Netherlands Survey, Technical report, ITC Project.

Brick, J. M. & Lepkowski, J. M. (2008), Multiple mode and frame telephone surveys, *in* Lepkowski et al. (2008), chapter 7, pp. 149–169.

de Leeuw, E. D. (2005), 'To mix or not to mix data collection modes in surveys', *Journal of Official Statistics* **21**, 233–255.

Duncan, K. B. & Stasny, E. A. (2001), 'Using propensity scores to control coverage bias in telephone surveys', *Survey Methodology* **27**, 160–168.

Garren, S. T. & Chang, T. C. (2002), 'Improved ratio estimation in telephone surveys adjusting for noncoverage', *Survey Methodology* **28**, 63–76.

Göksel, H., Judkins, D. R. & Mosher, W. D. (1991), Nonresponse adjustment for a telephone follow-up to a national in-person survey, *in* 'Proceedings of the Section on Survey Research Methods', ASA, pp. 581–586.

Groves, R. M. (1989), *Survey Errors and Survey Costs*, Wiley, New York, NY.

Hartley, H. (1962), Multiple frame surveys, *in* 'Proceedings of the Social Statistics Section', ASA, pp. 203–206.

Hoaglin, D. C. & Battaglia, M. P. (1996), A comparison of two methods of adjusting for noncoverage of non-telephone households in a telephone survey, *in* 'Proceedings of the Section on Survey Research Methods', ASA, pp. 497–502.

Lee, S. (2006), 'Propensity score adjustment as a weighting scheme for volunteer panel web surveys', *Journal of Official Statistics* **22**, 329–349.

Lee, S. & Valliant, R. (2008), Weighting telephone samples using propensity scores, *in* Lepkowski et al. (2008), chapter 8, pp. 170–183.

Lepkowski, J. K., Kalton, G. & Kasprzyk, D. (1989), Weighting adjustment for parital nonresponse in the 1984 SIPP panel, *in* 'Proceedings of the Section on Survey Research Methods', ASA, pp. 296–301.

Lepkowski, J., Tucker, C., Brick, J. M., de Leeuw, E. D., Japec, L., Lavrakas, P. J., Link, M. W. & Sangster, R. L., eds (2008), *Advances in Telephone Survey Methodology*, Wiley, Hoboken, NJ.

Rosenbaum, P. R. & Rubin, D. B. (1983), 'The central role of the propensity score in observational studies for causal effects', *Biometrika* **70**, 41–55.

Rosenbaum, P. R. & Rubin, D. B. (1984), 'Reducing bias in observational studies using subclassification on the propensity score', *Journal of the American Statistical Association* **79**, 516–542.

Schonlau, M., Zapert, K., Simon, L. P., Sanstad, K., Marcus, S., Adams, J., Spranca, M., Kan, H., Turner, R. & Berry, S. (2004), 'A comparison between responses from a propensity-weighted web survey and an identical RDD survey', *Social Science Computer Review* **22**, 128–138.

Smith, P. J., Battaglia, M. P., Huggins, V. J., Hoaglin, D. C., Rodén, A.-S., Khare, M., Ezzati-Rice, M. & Wright, R. A. (2000), 'Overview of the sampling design and statistical methods used in the National Immunization Survey', *American Journal of Preventive Medicine* **20**, 17–24.

Taylor, H. (2000), 'Does internet research work? comparing online survey result with telephone survey', *International Journal of Market Research* **42**, 58–63.

Terhanian, G. & Bremer, J. (2000), 'Confronting the selection-bias and learning effects problems associated with internet research', Research paper: Harris Interactive.

Terhanian, G., Bremer, J., Smith, R. & Thomas, R. (2000), 'Correcting date from online survey for the effects of nonrandom selection and nonrandom assignment', Research paper: Harris Interactive.

Thompson, M., Fong, G., Hammond, D., Boudreau, C., Driezen, P., Hyland, A., Borland, R., Cummings, K., Hastings, G., Siahpush, M., Mackintosh, A. & Laux, F. (2006), 'Methods of the International Tobacco Control (ITC) Four Country Survey', *Tobacco Control* **15 (Suppl III)**, 12–18.

Varedian, M. & Försman, G. (2002), 'Comparing propensity score weighting with other weighting methods: a case study on web data', Paper presented at the *Annual Meeting of the American Association for Public Opinion Research*, St. Petersburg Beach, FL.