ADJUSTING FOR ONLINE PARTICIPATING BIAS THROUGH PROPENSITY SCORING

Roberto Furlan[1]

Kantar Health, Epsom, UK

Roberto Corradetti

Department of Statistics and Applied Mathematics "Diego de Castro", University of Turin, Italy

Received: 14th August 2013/ Accepted: 18th October 2013

***Abstract**. Since the late 1990s, market research agencies have been increasingly using web surveys to investigate market, political, or social aspects. The main concern of web-based survey methodology is the presence of potential selection bias due to nonrandom selection. This work discusses how to reduce bias in web-based surveys with propensity scores and information from a supplementary non-RDD (Random Digit Dialling) small size telephone survey. We first illustrate the approach, then we present a case study including the setup, the analysis, and the main outcomes of a political study conducted over the Internet in Italy.*

Keywords: *Market research, Post-stratification, Propensity score, Selection bias, Web survey-CAWI*

1. INTRODUCTION

Since the late 1990s, market research agencies have been increasingly using Web-based surveys CAWI to investigate market, political, or social aspects. These surveys rely on large consumer panels and/or entry banners on popular web sites. This trend is due to the growing diffusion of the Internet and availability of fast Internet access (Cambiar, 2006; Epstein *et al.*, 2001; Fricker and Schonlau, 2002; Schonlau *et al.*, 2002). Web-based survey methodology is however affected by selection bias, here referred to as *Online Participating bias* (OP bias). The bias in web-based surveys is associated with nonrandom selection and nonrandom assignment. The presence of OP bias has been discussed by several researchers from the statistical community (Mitofsky, 1999; Rivers, 2000). They point out that

[1] Roberto Furlan, email: roberto.furlan@kantarhealth.com; roberto.furlan@gmail.com

surveys that are not based on probability sampling cannot produce trustworthy information and that it is not possible to statistically correct for the bias due to the gap between the target population and the collected sample. In this work we aim at showing how propensity scoring can be adopted for correcting such bias. Propensity scores have been developed for compensating for lack of random selection in different application domains (Rosenbaum and Rubin, 1983). We adapt it here to web-based surveys.

Specifically, we show how to reduce OP bias through Propensity Scores (PS). PS has been applied since the early 1980s to compensate for lack of random assignment to treatment groups and, thus, to reduce selection bias (Rosenbaum and Rubin, 1984, Rubin, 1997, D'Agostino, 1998; Keisuke *et al.*, 2003; Kurth *et al.*, 2006). The procedure proposed here for correcting OP bias in Web-based surveys combines PS methods with a parallel telephone survey. Some preliminary work in this area has been done in the United States by researchers working at RAND and Harris Interactive (Schonlau *et al.*, 2003; Terhanian *et al.*, 2001a, 2001b) and in Israel (Kenett *et al.*, 2003, 2006). However, the available literature is rather lacking in details and is focused on using an RDD (Random Digit Dialling) system — easily available in the United States but not allowed in some European countries like Italy. We consider here a PS approach combined with a non-RDD telephone survey in order to systematically adjust possible OP bias in open web-based surveys.

Section 2 discusses the potential of issues with web-based surveys and the characteristics of a control telephone survey that parallels the web-survey, Section 3 presents the PS approach from a theoretical point of view and in the context of web-based surveys. We then provide details on the setup, the analysis, and the main outcomes of the PS approach in a political study conducted over the Internet in Italy in 2004 (Section 4). The effectiveness of the selection bias adjustment provided by the PS approach is tested by a question regarding the political party vote at the Italian elections to the European Parliament of June 2004. To conclude, we compare the web-based survey outcomes with the actual results of the elections.

2. POTENTIAL AND ISSUES OF INTERNET-BASED SURVEYS

The diffusion of the Web in many countries all around the world and the increasing availability of broadband connections is leading to an impressive growth in the quality and quantity of Internet-based surveys. The Internet allows for collecting information from more and more population strata with multimedia questionnaires in areas like customer satisfaction and concept testing (Kenett *et al.*, 2006; Malhotra and Peterson, 2001; Weible and Wallace, 1998).

The move from traditional methods of surveying, such as face-to-face interviewing (PAPI – Paper-and-Pencil Interviewing, CAPI – Computer Assisted Personal Interviewing) and telephone interviewing (in particular CATI – Computer Assisted Telephone Interviewing) to Internet-based methods (CAWI – Computer Assisted Web Interviewing) became important when market research agencies realised that the Web can considerably reduce the time and the cost of collecting survey data. These two factors were also the main drivers associated with the shift in the 1980s from face-to-face to telephone surveys. However, that move was also supported by the fact that telephone surveys provided a better methodological choice than face-to-face surveys. In fact, telephone surveys have the intrinsic ability to easily reach low-density areas and ensure a more representative sample. However, for surveys involving visual and sensory stimuli, such as in concept testing, face-to-face surveys continued for years to be the only effective methodology with no valid alternatives. Only in the last decade, due to the growing diffusion of the Web and of high-speed connections, web-based surveys started presenting a first valid alternative to face-to-face surveys. The introduction and development of web-based surveys represent a methodological revolution, because online questionnaires can incorporate multimedia stimuli, such as visual cards, music or voice messages, videos, or even interactive tasks.

Besides the reduced cost (Einhart, 2003; Klein *et al.*, 2004) and the multimedia potential, the online approach presents several other advantages over face-to-face and telephone. These include: fast survey setup and execution; less intrusive process (allowing respondents to answer at their convenience, with the option to interrupt and later resume), and more accuracy due to the elimination of the so called interviewer error (Dillman, 2000; Levine *et al.*, 1999; McCullough, 1998). Consequently, a larger number of researchers began to apply web-based surveys (Cambiar, 2006; Epstein *et al.*, 2001; Fricker and Schonlau, 2002; Schonlau *et al.*, 2002).

Web-based surveys, however, need to be carefully evaluated for proper representation of the population frame (Bandilla, *et al.*, 2001; Faas, 2003). In this respect we can distinguish between three types of web-based surveys: 1) surveys naturally related to Internet users, 2) surveys focused on a more general statistical population, and 3) surveys using closed lists of emails such as lists of customers or employees. The first type of web-based surveys are a pure product of the Web age (e.g., measurement of opinions of the surfers of a particular web site). The second type of web-based surveys represent a migration of more traditional research studies (e.g., in concept testing). The third type of surveys are in fact a form of census since, typically, all customers or all employees will be asked to complete the questionnaire (Kenett, 2006). In the first case, it is possible to collect a representative sample, and web-based surveys should be definitely considered the best methodological approach. However, because there are no good lists of Internet users, even when the population of interest is Internet users, it is difficult to draw a probability sample. In the census type surveys a procedure such as the M-Test can be used to assess representativeness of the respondents (Kenett, 2006). For the second type of surveys, however, it is not possible to collect a probability sample and, as a consequence, there are justified concerns about bias in such web-based surveys. As an example, consider the potential bias from using banner advertisement to gather surfers, where respondents are affected by a strong and undesirable self-selection bias.

A better approach to conducting web surveys is to create a web community or panel of potential respondents (among others, All Global, Harris Interactive, SWG, Toluna). With this approach, the selection bias is strongly reduced, but not completely eliminated. Statistically speaking, community panels are considered to be better then simple lists of individuals, because survey participants are recruited from a large variety of sources and their demographic and attitudinal profiles are known to the panel managers. Therefore, it is possible to extract from the members' list a sample of respondents potentially representative of the target population. The extraction of the contact sample from the web community or panel list should be done by taking into consideration demographic, behavioural, and attitudinal variables according to their distributions in the target population.

It is worth noting that such a sample should not be considered a probability sample, but only a raw approximation. Even with meticulous selection, the sample collected from a web community or panel is still affected by some bias. This residual bias is not only due to the different response rate of the community members, but also to the fact that on-line respondents differ in some ways from off-line potential respondents. Generally speaking, people with online access tend to be younger, wealthier, with a better education, heavy consumers of technology products, tend to live in urban areas and in dual-parent families, and are of white or Asian/Pacific descent (Bethlehem, 2007; Couper, 2000; European Research into Consumer Affairs, 2001; Livraghi, 2011; Ministro per l'Innovazione e le Tecnologie, 2004; Palmquist and Stueve, 1996; Prandelli *et al.*, 2000; U.S. Department of Commerce, 2000; White, 2000). An additional source of bias in web-based surveys is the profile of the participants. Apparently, among those with online access, participants in web-based surveys tend to be more involved with technology gadgets and tend to have more free time. In this paper we refer to such bias as *Online Participating bias* (OP bias).

Web community and panel survey results are affected to a greater or lesser extent by these biases because of at least three decisions that people make in order to have their responses registered in a web-based survey:

1 The participant has to be connected to the Internet. This depends on the ability to use computers, the availability and the cost of a computer with Internet access.

2 The participant has to be a member of the specified community/panel. This depends on visibility of the community and the expected benefits from the registration;

3 The participant has to accept to respond to the survey. When the community or panel members receive an invitation to participate in a survey, they must respond positively and complete the questionnaire in full (interrupted questionnaires are usually not considered for the analysis). This decision depends on the interest of the participant in the topic of the survey, on the expected remuneration, on the length of the questionnaire, and on one's spare time and availability.

Each of these decisions leads to a different type of selection problem. Decision (1) leads to internet-users selection, decision (2) leads to a selection of specific Web community membership, and decision (3) leads to a full response to the web-based survey selection. The OP bias represents the combined effect of the bias arising from these three selection problems.

The presence of OP bias has been discussed in the literature (Bradley, 1999; Taylor, 2000). One main criticism of web-based surveys is that, since they are not based on probability sampling, they lack reliability. Some researchers (Mitofsky, 1999; Rivers, 2000) argue that it is not possible to correct for such bias, in particular when some of the bias is due to repeated participation of pre-selected respondents within a community. We suggest here that Propensity Scoring (PS) is capable of compensating for lack of random selection in web-based surveys. In the following section, after an introduction to this useful methodology, we adapt PS to web-based surveys.

3. THE PROPENSITY SCORES APPROACH

3.1 BACKGROUND

The PS approach was introduced by Rosenbaum and Rubin in the early 1980s (Rosenbaum and Rubin, 1983) and, since then, has been further explored and extensively used, particularly in healthcare research (D'Agostino, 1998; Keisuke *et al.*, 2003; Kurth *et al.*, 2006; Rubin, 1997). The approach is applied to observational data and it is basically a method for estimating treatment effects when treatment assignment is not random but can be assumed to be not confounded with relevant outcomes. In other words, the PS approach compensates for lack of random assignment to treatment groups and, thus, has the potential to reduce selection bias in web-based surveys.

Given a statistical unit, the propensity score is its probability of belonging to a particular treatment group, computed from a set of the characteristics of the unit, or covariates. PS allows the researcher to summarise, in a single score, the effect of a set of covariates on the probability of receiving a treatment and provides information for adjusting data from the treatment group based on data from a control group. To perform this adjustment, it is necessary to calculate propensity scores for the units in both the treatment and the control group. Based on the propensity scores, the two groups are stratified into a number of matching comparison groups, so that in each stratum there are two groups of units that have similar propensity scores. We can think of the units within each stratum as "randomly assigned" to the groups, as they are equally likely to belong to either the treatment or the control group. Finally, the treatment group proportions are weighted to be the same as the control group proportions, forcing the distribution of characteristics studied to be approximately the same within both groups.

Formally, the propensity score for unit i (i = 1, …, N) is its conditional probability of being assigned to treatment $\mathbf{Z}_i = 1$ *versus* control $\mathbf{Z}_i = 0$, given a vector of observed covariates $\mathbf{x}_i$:

$$PS\left(\mathbf{x}_i\right) = \Pr\left(\mathbf{Z}_i = 1 \middle| \mathbf{X}_i = \mathbf{x}_i\right) \tag{1}$$

The treatment assignment is assumed to be not confounded (Rosenbaum and Rubin, 1983), that is to say, the treatment is assumed to be statistically independent of potential outcomes conditional on the **X**s:

$$\begin{aligned} &\Pr\left(\mathbf{Z}_1 = \mathbf{z}_1, \ldots, \mathbf{Z}_N = \mathbf{z}_N \middle| \mathbf{X}_1 = \mathbf{x}_1, \ldots, \mathbf{X}_N = \mathbf{x}_N\right) = \\ &= \prod_{i=1}^{N} PS\left(\mathbf{x}_i\right)^{\mathbf{z}_i} \left(1 - PS\left(\mathbf{x}_i\right)\right)^{1-\mathbf{z}_i} \end{aligned} \tag{2}$$

The main limitation of PS, with respect to random assignment of treatments, lies in the inability to balance unobserved covariates (Joffe and Rosenbaum, 1999). Therefore, it is critical to have a set of covariates that have a sound rationale for inclusion and that are expected to control key biases.

3.2 PROPENSITY SCORES IN ONLINE RESEARCH

In the framework of this work, the procedure accounting for OP bias in a web survey is based on conducting a parallel telephone survey. The treatment group is represented by respondents who participate in the web survey, while the control group is represented by respondents who participate in the telephone survey.

A set of questions, the covariates of the model, has to be included in the questionnaires of both surveys. These questions – demographic, behavioural, and attitudinal variables – should be formulated in the same way in both questionnaires. The assumption is that large differences in the outcomes reflect biases in the Web sample because of the greater or smaller propensity of different respondents to be in the web-based survey. The two samples – online and telephone – do not need to be of similar size. However, the telephone survey should be large enough and representative of the target population, as its eventual biases would directly affect the adjustment for nonrandom selection applied to the online sample and, thus, the final outcome.

As presented in the previous section, the first step in applying PS is to adjust an online sample by estimating the propensity of each respondent of being in the web survey rather than in the telephone survey. This propensity is referred to as *sampling propensity score* and is computed by applying multivariate statistical tools. Either logistic regression or discriminant analysis can be used (Rosenbaum and Rubin, 1983). As a second step, respondents of both samples are classified into subclasses or strata based upon their PS. Cochran (1968) showed that a sub-classification with five equal-size strata is sufficient to remove over 90% of the selection bias associated with non-random aspects of the experiment. Rosenbaum and Rubin (1983) showed that even though the covariates have different distributions in the two samples, the distribution of covariates within each telephone group is approximately the same as the distribution of covariates within each corresponding online group. Consequently, if the online propensity proportions are weighted to be the same as the telephone propensity proportions, the distribution of characteristics studied will be approximately the same for both samples. In other words, the PS approach balances the distribution of the model's covariates in the online sample based on their distribution in the telephone sample. This is basically equivalent to a post-stratification of the online sample based on the distribution of a set of variables (the covariates) observed for a more representative sample which is provided by the telephone survey.

This approach has important advantages over traditional post-stratification approaches (i.e., target and rim wseighting described in Section 4.2). On the one hand, target weighting cannot be used when several variables are included in the post-stratification process, because the number of “cells” grows exponentially with the number of variables included, leading to the well-known problem of empty or low-populated cells. On the other hand, rim weighting should not be used on behavioural and attitudinal variables because they are likely to be inter-correlated and information on the joint distribution of the variables in the control telephone survey would be ignored, thus causing volatile weights to be generated. On the contrary, PS allows a comprehensive treatment of a large number of covariates by reducing observed characteristics to a single index, the sampling PS, which serves as the basis for a post-stratification. As a result, by applying carefully the PS methodology, it is possible to compensate for lack of random assignment and to significantly reduce the selection bias affecting the online sample.

In spite of its applicability, PS has so far been used only by few market research professionals to adjust online samples (Schonlau *et al.*, 2003; Terhanian *et al.*, 2001a; Terhanian *et al.*, 2001b), possibly because of the lack of availability of appropriate literature. PS, however, has been extensively applied in other fields, such as healthcare (U.S. General Accounting Office, 1995; Rosenbaum and Rubin, 1984) or education (Rosenbaum, 1986), as a valid alternative to random assignment, when this could not be used.

3.3 THE QUESTIONNAIRE

The main issue concerning the applicability of the PS approach lies in the appropriate choice of the questions with the covariates of the model, to be included in both the online and the control survey. In fact, this delicate choice directly affects the adjustment for nonrandom selection of the online sample. It is crucial to select a set of questions that have a sound rationale for inclusion and that are expected to capture adequately the differences between the Internet population and the general population.

As a first recommendation, questions related to behaviours underestimated by the control telephone survey, such as those investigating travelling or dining out, should be carefully avoided, as they are likely to be biased. In fact, the probability of any individual being interviewed over the telephone is dependent on whether or not he/she is at home when called (we don’t consider interviews over mobile phones), he/she has a phone number listed in the telephone directory, and he/she is willing and able to answer to survey questions. Another recommendation is to avoid topics that are sensitive to social desirability bias, such as religion, sexual orientation, health status, illegal behaviour, politics, etc. In fact, if discussed over the telephone, they are likely to produce deceitful answers, because of the presence of a live interviewer (van Eunen, 1995).

3.4 THE TELEPHONE SURVEY

As eventual biases in the control survey would be directly reflected on the adjustments applied to the online sample, it is crucial to pay particular attention not only to the questionnaire (Section 3.3), but also to other important features of the telephone survey. In particular, there are two main aspects of the control survey to be considered: its recentness and its sample size.

Regarding the former aspect, it is recommended to base the adjustment on a telephone survey conducted in parallel with the web-based survey, as the information collected over the telephone is supposed to reflect the present-day population. Therefore, for a market research agency that conducts web-based surveys throughout the year, it seems reasonable and appropriate to run a control survey at least three or four times a year, possibly more. Running several control surveys per year would not only allow up-to-date information about the target population to be collected, but also to test new questions in order to keep the PS model up-to-date and, possibly, improve its efficacy. It is worth reminding the reader that nowadays population characteristics, in terms of behaviours and attitudes, tend to change at an impressive rate, in particular people's propensity of being online. Information only a few months old is sometimes notably outdated.

Regarding the size of the telephone survey, this should be neither too small (lack of accuracy that would be reflected on the PS adjustment) nor too large (too expensive). All the usual considerations concerning the sample size and the associated error apply (Cochran, 1977; Groves *et al.*, 2004; Kalton, 1983) and should be borne in mind when planning a control telephone survey. For example, assuming that the sample is drawn as a simple random sample from a large population and that the acceptable margin of error for the control survey outcomes is ±3.0% with a confidence level of 95%, the most conservative choice for the response distribution (e.g., 50%) would lead to a minimum recommended sample size of 1068. If the available budget did not allow running such a large sample or if the researcher preferred to get a larger sample for a more accurate PS-based adjustment, it would be possible to consider simultaneously the respondents information from two or three subsequent surveys, possibly covering a time period of not more than three or four months.

3.5 COST CONSIDERATIONS

Regarding the cost aspect, conducting parallel online and telephone surveys might appear to be more expensive than running exclusively telephone surveys.

In general, the marginal cost of adding respondents is lower for web-based surveys than for telephone surveys (Schonlau, 2002). However, fixed cost are usually higher for web-based surveys, as there are additional cost elements, not present in telephone surveys, such as recruitment, community maintenance, and advanced questionnaire design (i.e., offering multimedia stimuli). Therefore, it appears that web-based surveys are more economical than telephone surveys only when the sample size reaches a certain threshold, probably lying between a few hundred and a thousand respondents, depending on the exact cost structure of the agency involved in the survey process.

However, if one wants to reduce the selection bias affecting an online sample through PS, a recent parallel control telephone survey, investigating the same target population, is necessary. This means that PS-adjusted web-based surveys are more economical than pure telephone surveys only when enough web-based surveys are run for each control telephone survey and provided that their sample sizes reach a certain threshold. This consideration seems to limit the applicability of PS approach to market research agencies with a considerable workload. As a final remark, in deciding whether or not to undertake the online path, a market research agency should focus on all benefits offered by web-based surveys (see Section 2) and not only on the cost.

4. PROPENSITY SCORES APPLIED TO ITALIAN ELECTION 2004

4.1 THE PROJECT

In order to evaluate the benefit provided by the PS approach, we decided to test it on the vote at the Italian elections to the European Parliament of June 13, 2004, for which we had the actual outcome provided by the Italian government.

According to the PS approach, we needed a recent telephone survey to act as a control sample in the PS methodology. Because of budget restrictions, we could not afford to start an ad-hoc telephone survey for this project. We used a telephone sample collected by *SWG* through a CATI system in July, 2004. This was the only recent survey conducted on the relevant target population (Italian people with the right to vote, aged 18 and over) that has been made available to us. The telephone survey was realised based on a stratified (over the five geographical MACRO AREAS[2]) and quota (GENDER and AGE) sample design. Information from the Italian Census 2001 was used to develop the design, and the phone numbers were

[2] NORD-OVEST: Liguria, Lombardia, Piemonte, Val d'Aosta; NORD-EST: Emilia Romagna, Friuli-Venezia Giulia, Trentino, Alto Adige, Veneto; CENTRO: Lazio, Marche, Toscana, Umbria; SUD: Abruzzo, Basilicata, Calabria, Campania, Molise, Puglia; ISOLE: Sardegna, Sicilia.

selected from the Italian public directory. Exactly 927 interviews were collected in less than one week, during after-work hours (5.30pm to 9.30pm).

One could object that a stratified and quota sample is not a probability sample and, thus, should not be used to develop the PS weights to correct the bias in the online sample, as it is itself possibly biased. It is necessary to point out that while in many countries (e.g., the United States) it is possible to use a RDD system to collect a telephone sample, in Italy this is not possible because of a stricter privacy law. Italian legislation (updated in 2010 by the D.P.R. 178/2010) does not allow contacting respondents unwilling to receive phone calls for market research purpose. Therefore, a stratified and quota sample such as the SWG one, even though not being strictly a probability sample, represents the best practicable solution. Because of this, it is the most popular sampling approach among the Italian market research agencies (Autorità per la Garanzia nelle Comunicazioni, 2013).

The web-based survey was conducted in September, 2004, by using the SWG *Research Online Community* (SWG, 2013) to collect the data. About two thousand community members were selected for participation based on a stratified sample design (MACRO AREAS, GENDER and AGE) reflecting information provided by the Italian Census 2001. They received an invitation by email to take part in the web-based survey. In order to increase the response rate, email reminders were sent after one week to those who did not yet participated in the survey. In ten days, 793 interviews were collected, with an overall response rate of about 40%. This sample is not a probability sample representative of our target population, as any sample collected through community members is at most representative of the community population.

Many questions were present in the telephone survey. However, for the PS approach we considered only the demographic, behavioural, and attitudinal variables mentioned in Section 4.2. Therefore, only these questions were repeated in the Web questionnaire and they were formulated exactly in the same way in both questionnaires to optimise the results. In the online questionnaire we also included a question about the political party voted at the Italian elections to the European Parliament of June 2004. In fact, as stated earlier, we intended to use this specific, external criterion variable to evaluate the effectiveness of the selection bias adjustment provided by the PS approach.

4.2 THE ANALYSIS

As a first step, we checked how closely the two survey populations were to the two target populations, on the basis of socio-demographic characteristics and we made the appropriate adjustments.

In the telephone sample, the distributions of GENDER, AGE, and MACRO AREA were quite close to the target distributions. However, almost 17% of respondents had a college degree, while only 7.8% of the target population actually holds such a degree. In order to adjust the telephone sample for this bias and some secondary differences with respect to the theoretical distributions of GENDER, AGE, and MACRO AREA, we performed a post-stratification based on these four variables ([GENDER x AGE x MACRO AREA] + EDUCATION LEVEL[3], here referred to as GAM+E weighting). After having recoded AGE into six categories and EDUCATION LEVEL into two categories, we used both the *cells* and the *marginal* weighting algorithm available in SPSS Quantum (SPSS Limited, 2002) to obtain the GAM+E weights.

In cells weighting, also known as *target weighting*, the survey population is divided into a set of mutually exclusive and exhaustive categories or "cells". In our study, these cells were obtained by combining in all possible ways the categories of GENDER x AGE x MACRO AREA (please note the notation "x"). The sample is then weighted by the ratio of the target population proportion in each cell to the corresponding sample population proportion (Smith, 1991). The cells weights obtained are here referred to as GAM weights. The marginal weighting, also know as *rim weighting*, is a procedure for adjusting the sample in such a way that the sample marginal proportions match the target population marginal proportions on a number of categorical characteristics at the same time (in our study, GAM and EDUCATION LEVEL — E). These weights are computed by an iterative proportional fitting algorithm (Little and Wu, 1991; Sharot, 1986). The marginal weights obtained are here referred to as GAM+E weights (please note the notation "+"). These two weighting procedures remove a lot of the non-response bias from the differences between respondents and non-respondents. However, the bias from the differences related to characteristics not taken into account by the weighting algorithm cannot be removed.

In spite of the great accuracy in the selection of respondents from the SWG community, the web-based survey population did not come out to be as close to the target population as we hoped. In fact, socio-demographic information and response rate did not come out to be independent variables, as we got a higher response rate

[3] The symbol "x" between two variables indicates that the sample has been weighted based on these variables through a cell weighting algorithm. The symbol "+" between two variables indicates that the sample has been weighted based on these variables through a marginal weighting algorithm.

for young, highly educated, and living in the north males[4]. In order to adjust the online sample for these major socio-demographic differences, we post-stratified the Web sample in the same way as the telephone sample, thus obtaining both the GAM and the GAM+E weights.

As a second step, we estimated the probability of each respondent being in the Web survey rather than in the telephone survey. Either logistic regression or discriminant analysis can be profitably used (Rosenbaum and Rubin, 1983), however we chose the former as it is typically done in current practice (Pasta, 2000; Schröder *et al.*, 2006; Stürmer *et al.*, 2005). In particular, we used the logistic regression framework of generalised linear model (GLM). Given n_t respondents in the telephone survey, n_w respondents in the Web survey ($n_t + n_w = n$), and q covariates, the model can be written as:

$$E(\mathrm{y}) = \boldsymbol{\mu}, \quad \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}, \quad logit(\boldsymbol{\mu}) = \boldsymbol{\eta} \tag{3}$$

where we have the following:

- $\mathbf{y}$ is an n-dimensional response variable, whose generic element $\mathbf{y}_i$ assumes value 0 for the telephone data and value 1 for the online data;
- $\mathbf{X}$ is a design matrix of the type (n x p), where p is the number of columns necessary to estimate the main and interaction effects of the q covariates;
- $\boldsymbol{\beta}$ is a p-dimensional vector of parameters;
- $\boldsymbol{\eta}$ is the systematic component of the model and it is referred to as the linear predictor;
- $logit\ (\boldsymbol{\mu}) = ln\ (\boldsymbol{\mu}\ /(1-\ \boldsymbol{\mu}))$ is the logit transformation of the mean response variable and represents the link function. It serves to link the random or stochastic component of the model, the probability distribution of the response variable $\mathbf{y}$, to the systematic component of the model $\mathbf{X}\boldsymbol{\beta}$**.**

Regarding the covariates included in the GLM, we tested all main effects and first-order interaction models considering employment status (v8: employed/ not employed), family size (v7: “How many people are there in your family, including yourself?”), and five attitudinal-behavioural variables (v1 “characteristics that describe you more often…” - v1a: “security seeker or change seeker”; v1b: “risk

[4] At the moment of our survey, the SWG team was developing a system of individual scores, based on the manifested response rate and to be used in the selection stage, directed to take into account the complex relationship between response rate and socio-demographic information. When this system is available, it should allow obtaining far more accurate samples than is possible now.

taker or risk avoider"; v1c: "sensitive to social pressure or not sensitive to social pressure"; v2: "regarding all the TV news programs, magazines, newspapers, and computer information services available, do you feel overloaded with information or do you prefer having a lot of information available?"; v3: "most companies today want to know about the individual interests and lifestyle of their customers so they can tailor their information services and products to each customer's personal preferences. In general, do you see such personalization as a good thing or a bad thing"?). Please note that we applied the GAM+E weights in the fitting process (prior weights), in order to take into account the adjustment provided by the initial socio-demographic post-stratification (inclusive of the education level information). The most significant variable in the model was v2 (attitude toward information), closely followed by v8 (employed/ not employed). All other variables were significant with a *p*-value less than 0.03.

We also tested all first-order interaction models, but no interaction appeared to be meaningful. Therefore, we decided to continue our analysis with the full main effects model. Results are provided in Table 1. Please note that all variables except family size (v7) are categorical dichotomous: only the effect for the second level has been estimated for these variables (the effect of the first level is estimated by the intercept). Family size has been considered as continuous with range 1 to 6.

In the last column the odds ratio estimates (OR) are presented. An OR (among the others, Agresti, 2002) indicates the likelihood that the related variable's level is associated with being in the Web survey. If the odds ratio is larger than 1, it is likely that the level is associated with being in the web survey. If the odds ratio is smaller than 1, it is likely that the level is associated with being in the telephone survey. The largest ORs are for variables v2 and v8 which were also the most significant variables in the model. Since there are no interaction terms in the model, the interpretation is straightforward. Respondents who prefer having a lot of information (v2) have a 2.15 times higher chance of being in the Web survey than of being in the telephone survey (as compared with respondents who feel overloaded with information). Similarly, respondents who are not employed (v8) have a 1.68 (1/ 0.60) times higher chance of being in the telephone survey than of being in the Web survey (as compared with respondents who are employed). As family size (v7) has been considered as continuous, its OR relates to the ratio of the odds for a one-unit change in family size. So we can say for a one-unit increase in family size, we expect to see about 12% (1/0.89=1.12) increase in the odds of being in the telephone survey than of being in the Web survey.

Table 1: Regression information for the GLM used for the PS estimation.

	Estimate	Std. Error	t value	Pr(>|t|)	OR
(Intercept)	-0.0210	0.2357	-0.089	0.929	
v1a change seeker	0.0520	0.0230	2.261	0.024	1.05
v1b risk avoider	-0.3277	0.1374	-2.384	0.017	0.72
v1c not sensitive to social pressure	0.2669	0.1193	2.237	0.026	1.31
v2 prefer having a lot of information	0.7640	0.1224	6.242	0.000	2.15
v3 bad thing	0.2513	0.1150	2.185	0.029	1.29
v7 family size (continuous)	-0.1152	0.0428	-2.689	0.007	0.89
v8 not employed	-0.5173	0.1005	-5.146	0.000	0.60

As a third step, we transformed the linear predictors by the inverse of the link function to obtain the fitted mean values. The fitted mean values can be interpreted as the probability of each respondent being in the Web survey rather than in the telephone survey, thus they can be referred to as sampling PS. We estimated the kernel density of these values with a smoothing kernel and a smoothing bandwidth of 0.06. A plot of the estimated kernel density, for both the telephone and the online sample, is provided in Figure 1. As one might expect, online respondents have a probability of being online significantly greater than telephone respondents. Consequently, it was necessary to adjust the Web sample for this important bias. We did this by applying the PS approach earlier explained.

We classified respondents of both surveys into subclasses based on their sampling PS, leading to the variable PS_CLASS, a new variable that summarises the information contained in the covariates included in the GLM. It is expected to adjust the online data for the fact that respondents have a different probability to be in the online rather than in the telephone survey (sources of OP bias (1) and (2) - Section 2). More precisely, five subclasses of equal size were constructed at the quintiles of the sample distribution of the PS. We used a five-group classification as, according to Cochran (1968), it is sufficient to remove over 90% of the selection bias associated with non-random aspects of the experiment. Then, based on the GAM+E-weighted samples, we weighted the online propensity proportions of PS_CLASS to be the same as the telephone propensity proportions of PS_CLASS. In other words, we computed a weight for each of these five subclasses of PS_CLASS; this set of five weights represents the target distribution of PS_CLASS.

At this point, we computed the final individual weights for the online respondents considering both the socio-demographic variables (GENDER, AGE,

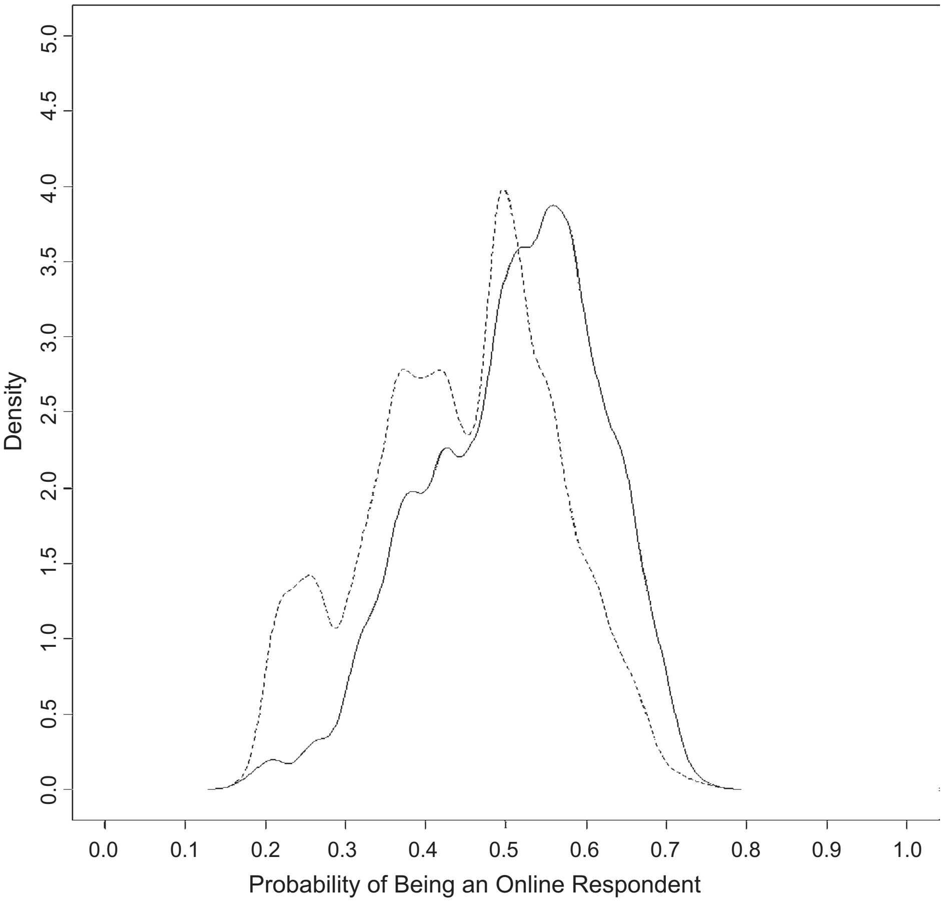

Figure 1: Density plot of online respondent probability for the CATI (dashed line) and the Web sample (solid line) — GAM+E weights

MACRO AREA, and EDUCATION LEVEL) and the PS_CLASS information. In computing these final weights, we had to keep working with the socio-demographic variables because we used them all along in the PS process; excluding them now would not allow adjusting the sample for these important dimensions. These final individual weights have been computed through the marginal weighting algorithm available in SPSS Quantum (SPSS Limited, 2002). Because of their composition, they should be referred to as GAM+E+PS weights (Figure 2); however, for the sake of brevity we refer to them as PS final (PSF) weights.

PSF = GAM + E + PS

VARIABLES INVOLVED:	Gender, age macro area, education level	attitudinal-behavioral, employment status, family size
SOURCE:	official (Italian Census 2001)	telephone survey
MOTIVATION FOR THE ADJUSTMENT:	socio-demographic information and response rate are not independent variables	respondents have a different probability to be in the online rather than in the telephone survey

Figure 2:weights in the PS approach: variables involved, target distributions source and reasons for the corrections.

It is worth mentioning that we could have used an alternative approach: considering the socio-demographic variables as covariates in the PS model and not weighting separately for them in each step (i.e., model estimation; computation of PS_CLASS weights; computation of final individual weights for the online respondents). However, this approach would not have allowed obtaining exact marginal distributions for these socio-demographic variables, but only marginal distributions more or less close to the theoretical distributions. Final users (market researchers and their clients) expect the sample to be exactly adjusted based on known figures (they are usually concerned with GENDER, AGE, and MACRO AREA/REGION information, sometimes also with EDUCATION LEVEL) so they feel reassured when the sample is (or, rather, appears) ‘representative’ of the target population.

In Section 3.2, we mentioned that according to Rosenbaum and Rubin (1983) the distribution of covariates within each telephone subclass is expected to be similar to the distribution of covariates within each corresponding online subclass. Table 2 shows that the covariates have actually reasonably similar distributions for both samples in each of the five strata. We have also mentioned that with the PS approach the distribution of the observed covariates is expected to be approximately the same for both the telephone and the online samples. Using the GLM framework previously reported in this section and the five-group classification of the sample distribution of the PS, we have been successful in balancing the observed covariates.

The main results are presented in Table 3. The distributions of variables in the GAM+E-weighted online sample substantially differ from the distributions of the same variables in the GAM+E-weighted telephone sample (in particular for v2 and v8, for which the differences are large – these are the variables with the lowest *p*-value in the GLM). However, when the PSF weights are applied, the distributions of variables in the online sample are much closer to the telephone sample ones, suggesting that the PS approach here adopted was appropriate to balance the covariates.

Table 2: Distribution of GAM+E-weighted covariates included in the GLM for PS estimation by telephone/online sample and sub-classification group (%).

	Group 1		Group 2		Group 3		Group 4		Group 5	
GLM covariates	Tel.	Online	Tel.	Online	Tel.	Online	Tel.	Online	Tel.	Online
v1 – psychographic characteristics										
(a) security seeker	79.0	85.4	58.1	58.0	72.6	72.6	57.8	60.9	47.8	40.2
change seeker	21.0	14.6	41.9	42.0	27.4	27.4	42.2	39.1	52.2	59.8
(b) risk taker	4.4	7.3	11.0	19.7	9.6	13.9	20.9	16.1	53.3	35.6
risk avoider	95.6	92.7	89.0	80.3	90.4	86.1	79.1	83.9	46.7	64.4
(c) sensitive to social pressure	87.5	91.9	86.4	85.4	80.4	75.8	70.7	84.5	52.2	42.5
not sensitive to social pressure	12.5	8.1	13.6	14.6	19.6	24.2	29.3	15.5	47.8	57.5
v2 - information available										
feel overloaded with information	82.7	68.3	22.5	20.4	2.3	4.4	0.0	1.2	0.0	0.6
prefer having a lot of information	17.3	31.7	77.5	79.6	97.7	95.6	100.0	98.8	100.0	99.4
v3 - personalization										
good thing	85.2	79.7	81.1	82.2	82.6	87.3	67.8	69.9	49.5	50.6
bad thing	14.8	20.3	18.9	17.8	17.4	12.7	32.2	30.1	50.5	49.4
v7 - family size										
1	2.6	0.8	6.6	2.5	5.0	2.5	6.0	15.5	13.0	23.9
2	17.1	10.7	32.9	38.2	18.2	9.6	31.0	23.0	23.9	27.2
3	25.3	12.3	25.9	23.6	31.4	44.6	21.6	32.8	45.7	36.1
4	39.8	50.0	28.1	27.4	32.7	36.9	25.0	17.8	13.0	11.7
5	10.0	20.5	3.9	7.6	10.5	5.1	16.4	10.9	3.3	1.1
6+	5.2	5.7	2.6	0.6	2.3	1.3	0.0	0.0	1.1	0.0
v8 - employment status										
employed	27.7	26.0	22.0	17.2	64.8	65.8	81.0	89.7	94.6	97.2
not employed	72.3	74.0	78.0	177.0	35.2	34.2	19.0	10.3	5.4	2.8

Table 3: Distribution of covariates included in the GLM for PS estimation (%).

	Telephone Sample	Online Sample	
GLM covariates	GAM+E	GAM+E	PSF
v1 – psychographic characteristics			
(a) security seeker	66.7	61.9	70.4
change seeker	33.3	38.1	29.6
(b) risk taker	14.3	19.3	14.0
risk avoider	85.7	80.7	86.0
(c) sensitive to social pressure	79.9	74.5	82.4
not sensitive to social pressure	20.1	25.5	17.6
v2 - information available			
feel overloaded with information	30.3	16.1	32.7
prefer having a lot of information	69.7	83.9	67.3
v3 - personalization			
good thing	77.9	73.0	79.2
bad thing	22.1	27.0	20.8
v7 - family size			
1	5.5	10.0	5.1
2	23.6	22.3	20.8
3	28.5	30.8	27.5
4	30.5	27.0	33.7
5	8.7	8.2	9.1
6+	3.1	1.6	3.8
v8 - employment status			
employed	48.4	62.2	49.7
not employed	51.6	37.8	50.3

As a final step, we wanted to re-estimate the probability of each respondent being in the Web survey rather than in the telephone survey, but this time using the PSF weights for the online sample. Therefore, we fitted the same main-effects GLM as previously done, but in the fitting process we used the PSF weights for online respondents, in order to take into account the additional adjustment provided by the PS approach. We obtained the fitted mean values for the new GLM and we estimated their kernel density as previously done. A plot of the estimated kernel density, for both the telephone and the online sample, is provided in Figure 3. As one might expect, after applying the PS approach, telephone and online respondents have a similar probability of being online. Consequently, it is possible to assert that the PS approach could properly adjust the online sample toward the telephone sample, with respect to the set of covariates included in the GLM. It is worth noting that the variables in the PSF-weighted online sample and the variables in the telephone sample will actually have a comparable distribution only if the set of covariates included in the PS model can sufficiently explain the OP bias.

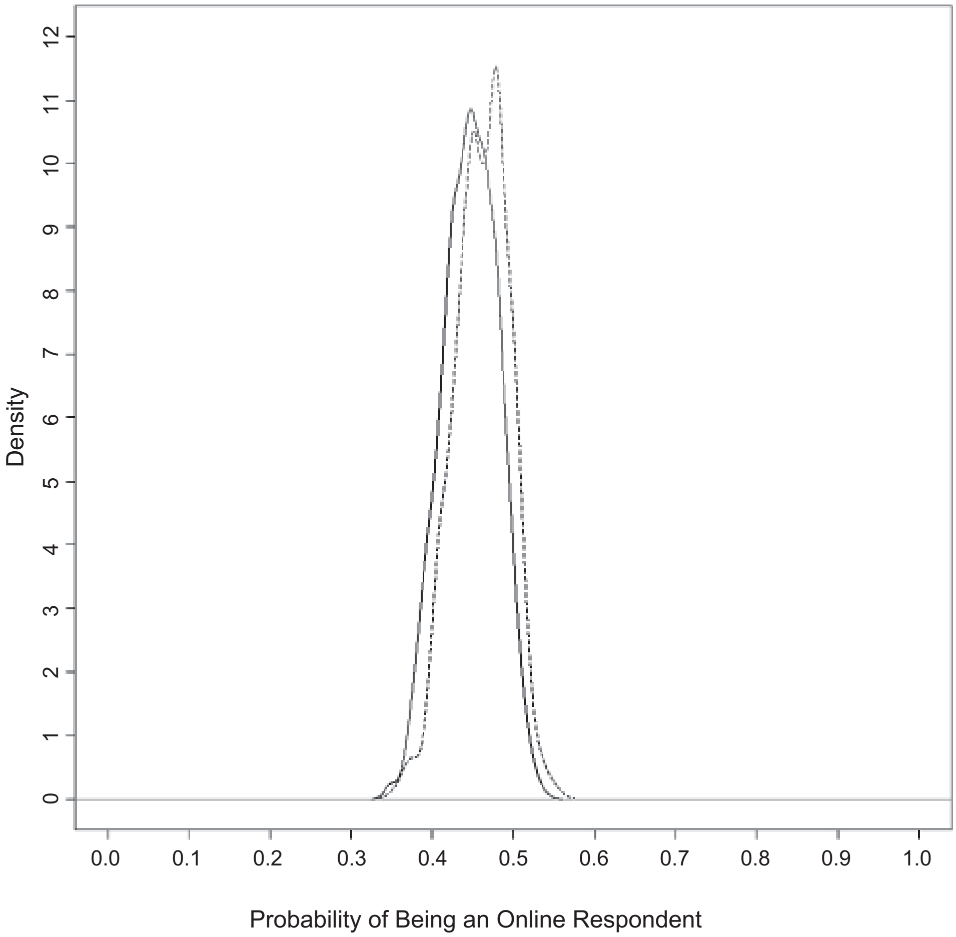

Figure 3: Density plot of online respondent probability for the CATI (dashed line) and the Web sample (solid line) — PSF weights

4.3 EFFECTS OF WEIGHTING ON BIAS

As stated earlier, in order to be able to evaluate the effectiveness of the selection bias adjustment provided by the PS approach, we included, in the on line questionnaire, a question regarding the political party voted at the Italian elections to the European Parliament of June 2004 to be used as an external criterion variable.

From the actual results of the elections, we found out that 26.9% of the eligible electors were not interested in voting and did not go to the polling stations, while 7.7% of the eligible electors did not express an acceptable vote (blank or void votes). We decided to focus the attention only on valid votes (65.4%) because this is what people are usually interested in. In addition, because of the presence of several small parties due to the highly fragmented party system, and because of the limited size of the online sample, we decided to collapse together all parties that received less than 3% of valid votes.

Based on the question included in the online questionnaire, we estimated the election results by using the unweighted data and also by applying the three vectors of weights previously computed: GAM, GAM+E, and PSF. In this way, we could evaluate the effectiveness of different post-stratification approaches in predicting the election results. The results of our analysis are reported in Table 4. In both the unweighted and GAM-weighted data, the left-wing parties (*United in the Olive Tree*, *Communist Refoundation Party*) are heavily overestimated, while the right-wing parties (*Forza Italia*, *Union of Christian Democrats*, *National Alliance*, *Northern League*) are heavily underestimated. This important bias probably originates from the tendency of left-wing oriented people to be more willing to participate to a political survey than right-wing ones. This bias is considerably reduced by the GAM+E post-stratification and, even further, by the PS approach. In fact, only reasonable differences are still present between the PSF-weighted sample results and the actual elections outcomes, all of them well within the margin of error at 95% confidence level. The difference in accuracy is quantified by the sum of the squared differences between the estimated and the actual results as reported in the last row of Table 4.

A graphical representation of the results reported in Table 4 is provided in Figure 4, based on a multidimensional scaling (MDS) analysis. Firstly, we computed a distance matrix containing the Euclidean distances between the columns of Table 4. Then, we used the function *cmdscale* available in the package R (R Development Core Team, 2013) to perform a classical multidimensional scaling (Cox and Cox, 1994; Seber, 1984) on the distance matrix, and we plotted the first two dimensions.

Table 4: Italian elections to the European Parliament of June, 2004 (%). Sample estimates vs. actual results. Standard errors are reported in brackets.

Political Party	Unweigh.	GAM	GAM+E	PSF	Actual Results
United in the Olive Tree	36.3	35.6	34.8	33.6	31.1
(*Uniti nell'Ulivo*)	[.019]	[.024]	[.031]	[.038]	
Communist Refoundation Party	9.4	9.0	7.9	6.9	6.1
(*Rifondazione Comunista*)	[.012]	[.015]	[.018]	[.020]	
Forza Italia	16.3	16.8	18.7	18.8	21.0
	[.015]	[.019]	[.025]	[.031]	
Union of Christian Democrats	3.8	3.9	4.4	5.5	5.9
(*Unione Democratica Cristiana*)	[.008]	[.010]	[.013]	[.018]	
National Alliance (*Alleanza Nazionale*)	9.1	9.0	9.9	10.3	11.5
	[.011]	[.015]	[.020]	[.024]	
Northern League	2.8		3.2	3.5	3.8 5.0
(*Lega Nord*)	[.007]	[.009]	[.012]	[.015]	
Others	22.3	22.5	20.7	21.1	19.5
	[.016]	[.021]	[.026]	[.032]	
∑é(estimated - actual results)^2	82.8	68.9	31.0	17.4	

The reader can easily see that the initial GAM adjustment is not effective in removing the right-wing/left-wing bias earlier described. The introduction of the education information in the weighting process (GAM+E weights) reduces this bias significantly, but it is the PSF weighting that removes most of the bias. The residual difference between the PSF-weighted results and the actual election results can be explained as residual OP bias and/or as sampling error due to the limited sample sizes of both the online and the telephone survey.

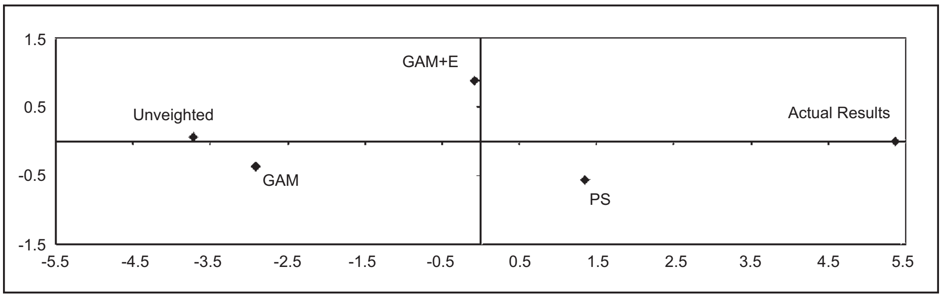

Figure 4: MDS representation of the differences between the estimated and the actual election results

4.4 EFFECTS OF WEIGHTING ON ACCURACY

A well-known disadvantage of weighting is that the variance of estimates increases, which means that the precision of estimates are reduced by the weighting process. Thus, weighting makes confidence intervals wider and differences will have to be larger to be recognised as statistically significant. It is evident that it is very important to be able to quantify the extent by which the standard error of an estimate increases due to the weighting process. The factor by which the variance is increased due to the weighting process is usually referred to as *design effect* (variance of estimate in the weighted data = variance of estimate in the unweighted data × design effect) (Kish, 1965; 1995). Kish (1992) has shown that the design effect of weighting (DEFFW) can be estimated from the sample by computing the relative variance of the weights and adding one. It is given by:

$$DEFFW = \frac{n\sum_{i=1}^{n}\mathbf{W}_i^2}{\left(\sum_{i=1}^{n}\mathbf{W}_i^2\right)} \tag{4}$$

where $\mathbf{W}_i$ is the weight of respondent *i* and *n* is the number of respondents in the sample. The DEFFW index assumes values in [1, *n*]. It assumes value 1 when all weights are equal to each other (i.e., there is no adjustment at all), while it equals the number of sample units when the difference among the weights is maximised (i.e., all units but one have null weight).

In order to estimate the negative impact on the precision of estimates of the PSF weighting process, we computed the DEFFW index for the PSF weights ($\text{DEFFW}_{\text{PSF}}$ = 4.03). The DEFFW indexes for the socio-demographic GAM+E weights ($\text{DEFFW}_{\text{GAM+E}}$ = 2.72) and for the GAM weights ($\text{DEFFW}_{\text{GAM}}$ = 1.66) have also been computed in order to get a benchmark. The results show the extent to which the negative effect on variance increases (i.e., the design factor increases) as the weighting becomes more and more complex because more and more variables are taken into consideration. Therefore, while the bias of the estimates has been considerably reduced by the PS approach (Section 4.3), their variance has increased, leading to less accurate estimates. In other words, the bias reduction came at the cost of an increase in variance, confirming the conclusions of previous studies (Lee, 2006). It is worth noting that the strongest increase in variance occurred when the adjustments were most effective in reducing the bias (i.e., GAM+E-weighted data versus unweighted data). This indicates the important contribution of the socio-demographic variables and education level information to the overall weighting process. This has also another important meaning: if the community is properly managed and the invitation/ selection process takes into account the individual response rate, then the online sample is likely to be properly selected. Therefore, the sample distribution of the socio-demographic and education level variables would be approximately the same as the distribution of the same variables in the target population, and, consequently, the initial GAM+E adjustment would not be required. The PSF weights would be based exclusively on the variables included in the GLM of the PS approach and would not be inclusive of the GAM+E weights as in the case history presented in this article. As a consequence, the DEFFW index for the PSF weights would be smaller, thus the effective sample size would be closer to the actual sample size.

One can conclude that, given an appropriate selection of the online sample and a competent application of the PS approach, including the choice of the calibrating variables, PS-based weights have the potential to significantly reduce the OP bias without necessarily leading to an excessive increase in variance. The larger online sample required to compensate for the increase in variance is likely to be more than compensated by the financial benefits and the multimedia opportunities that Web-based survey offer.

5. CONCLUSIONS AND FURTHER RESEARCH

In this article, we examined PS as a method to overcome non-sampling bias affecting internet-based surveys. This is an important topic to market research professionals because online samples are almost always affected by biases associated with nonrandom selection and nonrandom assignment. We developed the preliminary work done by some researchers over the last few years (Kenett *et al.*, 2003, 2006; Schonlau *et al.*, 2003; Terhanian *et al.*, 2001a, 2001b) who proposed PS adjustment for Internet-based surveys based on the results from a parallel telephone survey. In particular, we discussed the proposed methodology in more detail and we presented an application based on a non-RDD system for the parallel telephone survey.

Our application was based on a political study conducted over the Internet in Italy in 2004, with a target population of Italian adults. The propensity scores were computed based on a suitable CATI sample collected a couple of months earlier than the online sample. We illustrated the detailed procedure to derive the PSF weights and we presented the main results. In order to offer the cleanest possible picture of the PS approach and its benefits, we made continuous comparisons with the traditional socio-demographic (GAM and GAM+E) post-stratifications. A question regarding the political party voted at the Italian elections to the European Parliament of June 2004 was used as an external criterion variable to evaluate the effectiveness of the selection bias adjustment provided by the PSF weighting and compare it with the adjustments provided by the socio-demographic weightings. The evident bias affecting the web-based survey was reduced to a satisfactory level only by the PSF adjustment, while the socio-demographic weights alone were not very effective. It is important to highlight that through the PS approach the bias was reduced but not completely cancelled.

We showed how the PS approach removed substantial bias from the online sample and, thus, improved the estimation of the variable of interest, although this occurred at the cost of an increase in variance. Our analysis was limited to a political study, but, if properly applied, PS approach can benefit other research sectors for which data are collected through online samples (Schonlau *et al.*, 2003). The key is that the covariates of the model should be able to compensate for the selection bias resulting from an online sample consisting exclusively of Internet respondents. Therefore, the success of PS approach hinges on including, in both the online and the control sample, a set of questions that are expected to capture adequately the differences between the Internet population and the general population. As a consequence, more work should be done in this direction, for example by identifying different sets of variables appropriate for different research sectors to be investigated by web-based surveys (politics, durables, fast moving consumer goods, telecommunications, financial services, etc).

More work should also be done in investigating alternative ways to derive the PSF weights, for example by applying a statistical tool different from logistic regression, or by classifying respondents into strata in a different way from that suggested by Cochran (1968), for instance using strata formed by balancing on the inverse variance of the strata treatment effects (Huppler Hullsiek and Louis, 2002).

It is important to highlight the importance of having an appropriate and representative control survey (i.e., unbiased, large, recent, and on the same target population), to avoid introducing harmful and additional bias into the online sample. For our application we had to use an initial post-stratification adjustment to balance the control sample in terms of the socio-demographic information, in particular education level. Ideally, this initial adjustment should not be necessary. A similar consideration should be made for the online sample. In fact, more effort should be spent to obtain a socio-demographically balanced online sample, in order to avoid or at least to contain the initial adjustment (Section 4.4). We believe that more work should also be done to study the effects of the sample size and of the sampling method for the control telephone survey, possibly through a simulation study where the exact relationships between the covariates and the actual behaviour is known.

ACKNOWLEDGEMENT

We would like to thank SWG S.r.l., Trieste, Italy for providing both the web and the telephone sample and thus giving us the opportunity to study and apply the propensity scores approach to a real-life case study.

REFERENCES

Agresti, A. (2002). Categorical Data Analysis. New York: Wiley-Interscience.

Autorità per la Garanzia nelle Comunicazioni (2013). *Elenco dei Documenti Relativi ai Sondaggi*. http://www.agcom.it/Default.aspx?message=contenuto&DCId=302. Last access: 29/08/2013.

Bandilla, W., Bosnjak, M., and Altdorfer, P. (2001). Effekte des Erhebungsmodus? Ein Vergleich Zwischen einer Web-basierten und einer Schriftlichen Befragung Zum ISSP-Modul Umwelt. In *ZUMA Nachrichten*, 49**:** 7-28.

Bethlehem, J. G. (2007). *Reducing the Bias of Web Survey Based Estimates*. Discussion paper 07001, Statistics Netherlands, Voorburg/Herleen, The Netherlands.

Bradley, N. (1999). Sampling for Internet surveys: An examination of respondent selection for Internet research. In *Journal of the Market Research Society*, 41(4): 387-394.

Cambiar (2006). *The online research industry. An update on current practices and trends*. Technical report. http://www.sigmavalidation.com/tips/06_05_01 The Online Research Industry 2006.pdf. Last access: 29/08/2013.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. In *Biometrics*, 24: 295-313.

Cochran, W. G. (1977). *Sampling techniques.* Third edition. New York: John Wiley & Sons.

Couper, M. P. (2000). Web surveys: a review of issues and approaches. In *Public Opinion Quarterly*, 64: 464-494.

Cox, T. F. and Cox, M. A. A. (1994). *Multidimensional scaling*. London: Chapman & Hall.

D'Agostino, R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. In *Statistics in Medicine*, 17(19): 2265-2281.

Dillman, D. A. (2000). *Mail and Internet surveys: The Tailored design method*. New York: John Wiley & Sons.

Einhart, N. (2003). The opinion catcher. In *Business* 2.0, 4(4), 87.

Epstein, J., Klinkenberg, W. D., Wiley, D., and McKinley, L. (2001). Insuring sample equivalence across Internet and paper-and-pencil assessments. In *Computers in Human Behavior*, 17(3): 339-346.

van Eunen, E. A. (1995). *Interviewing for market and opinion research*. Amsterdam, The Netherlands: ESOMAR.

European Research into Consumer Affairs (2001). *Preventing the digital television and technological divide*. Technical report. http://www.net-consumers.org. Last access: 23/02/2006.

Faas, T. (2003). Offline rekrutierte access panels: Königsweg der online-forschung? In *ZUMA-Nachrichten*, 53(27): 58-76.

Fricker, R. D. Jr., and Schonlau, M. (2002). Advantages and disadvantages of Internet research surveys: Evidence from the literature. In *Field Methods*, 14(4): 347-367.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2004). *Survey methodology.* New York: Wiley.

Huppler Hullsiek, K., Louis, T. A. (2002). Propensity score modeling strategies for the causal analysis of observational data. In *Biostatistics*, 2(4): 179-193.

Joffe, M. M. and Rosenbaum P. R. (1999). Propensity scores. In *American Journal of Epidemiology*, 150: 327-333.

Kalton, G. (1983). *Introduction to survey sampling.* Beverly Hills: Sage Publications.

Keisuke, H., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. In *Econometrica*, 71(4): 1161-1189.

Kenett, R. S., Kaplan, O., and Raanan, Y. (2003). Statistical properties of Internet based market research surveys. In *proceedings of the third annual conference of ENBIS, the European Network for Business and Industrial Statistics*, Barcelona, Spain.

Kenett, R. S. (2006). On the planning and design of sample surveys. In *Journal of Applied Statistics*, 33(4): 405-415.

Kenett, R. S., Kaplan, O., and Raanan, Y. (2006). Surveys with new technologies: is it the end of telephone interviews? [in Hebrew]. In *Kesher Haeihut*, 53-54: 6-8.

Kish, L. (1965). *Survey Sampling*. New York: Wiley and Sons.

Kish, L. (1992). Weighting for unequal P_i . In *Journal of Official Statistics*, 8: 183-200.

Kish, L. (1995). Methods for design effects. In *Journal of Official Statistics*, 11: 55-77.

Klein, D., Roster, C. A., Albaum, G., and Rogers, R. D. (2004). A comparison of response characteristics from Web and telephone surveys. In *International Journal of Market Research*, 46(3): 359-373.

Kurth, T., Walker, A. M., Glynn, R. J., Chan, K. A., Gaziano, J. M., Berger, K., and Robins, J. M. (2006). Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. In *American Journal of Epidemiology*, 163(3): 262-270.

Lee, S. (2006). Propensity Score adjustment as a weighting scheme for volunteer panel Web surveys. In *Journal of Official Statistics*, 22(2): 329-349.

Levine, P., Ahlhauser, B., and Kulp, D. (1999). Pro and con: Internet interviewing. In *Marketing Research*, 11(2): 33-6.

Little, R. J. A. and Wu Mei-Miau (1991). Models for contingency tables with known margins when target and sampled populations differ. In *Journal of the American Statistical Association*, 86: 87-95.

Livraghi, G. (2011). *Dati sull'Internet in Italia*. Technical report. http://www.gandalf.it/dati/dati3.htm. Last access: 29/08/2013.

Malhotra, N. K. and Peterson, M. (2001). Marketing research in the new millennium: emerging issues and trends. In *Marketing Intelligence & Planning*, 19(4): 216-232.

McCullough, D. (1998). Web-based market research users in new age. In *Marketing News*, 32(19): 27-28.

Ministro per l'Innovazione e le Tecnologie (2004) *Rapporto Statistico sulla Società dell'Informazione in Italia*. Technical report. http:// http://www.academia.edu/1216304/Rapporto_Statistico_sulla_Societa_dell_Informazione_in_Italia. Last access: 29/08/2013.

Mitofsky, W. J. (1999). Pollsters.com. In *Public Perspective*, 10(24): 24-26.

Palmquist, J. and Stueve, A. (1996). Stay plugged into new opportunities. In *Marketing Research*, 8(1): 13-15.

Pasta, D. J. (2000). Using propensity scores to adjust for group differences: examples comparing alternative surgical methods. In *proceedings of the Twenty-Fifth Annual SAS Users Group International Conference*, Indianapolis, Indiana.

Prandelli, E., Spreafico, C., and Pol, A. (2000). Osservatorio Internet Italia: l'utenza Internet 2000, *I-LAB Centro di Ricerca sull'Economia Digitale*. http://www.unibocconi.it (PDF version of document downloaded March 8, 2006).

R Development Core Team (2013). *R: A language and environment for statistical computing*. R foundation for statistical computing, Vienna, Austria. URL http://www.R-project.org.

Rivers, D. (2000). Fulfilling the promise of the web. In *Quirks Marketing Research Review*: 34-41.

Rosenbaum, P. R. (1986). Dropping out of high school in the United States: an observational study. In *Journal of Educational Statistics*, 11(3): 207-224.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of propensity score in observational studies for casual effects. In *Biometrika*, 70(1): 41-55.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. In *Journal of the American Statistical Association*, 79: 516-524.

Rubin, D. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, 127(8): 757-763.

Schonlau, M., Fricker Jr., R. D., Elliott, M. N. (2002). *Conducting Research Surveys Via E-mail and the Web*. Santa Monica, CA: RAND.

Schonlau, M, Zapert, K., Payne, L. S., Sanstad, K., Marcus, S., Adams, J., Spranca, M., Kan, H., Turner, R., Berry, S. (2003). A comparison between a propensity weighted Web survey and an identical RDD survey. In *Social Science Computer Review*, 21(10): 1-11.

Schröder, H., Caputo, A., Debling, D., and Stürmer, T. (2006). Propensity-score based methods: an application to data on cognitive function in the elderly. In *proceedings of the Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS)*, Leipzig, Germany.

Seber, G. A. F. (1984). *Multivariate Observations*. New York: John Wiley & Sons.

Sharot, T. (1986). Weighting survey results. In *Journal of the Market Research Society*, 28: 269-284.

Smith, T. M. F. (1991). Post-stratification. In *The Statistician*, 40(3): 315-323.

Stürmer, T., Schneeweiss, S., Avorn, J., and Glynn, R. J. (2005). Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. In *American Journal of Epidemiology*, 162(3): 279-289.

SPSS Limited (2002). *Quantum User's Guide – Volume 3 Advanced Tables*, SPSS Limited, London.

SWG (2013). *La Community di SWG*. http://www.swg.it. Last access: 29/08/2013.

Taylor, H. (2000). Does Internet research work? comparing online survey results with telephone survey. In *International Journal of Market Research*, 42(1): 51-63.

Terhanian, G., Marcus, S., Bremer, J., and Smith, R. (2001a). Reducing error associated with non-probability sampling through propensity scores: evidence from election 2000. In *Joint Statistical Meeting 2001*, Atlanta, Georgia, USA.

Terhanian, G., Taylor, H., Siegel, J., Bremer, J., and Smith, R. (2001b). The accuracy of Harris Interactive's pre-election polls of 2000. In *AAPOR 2001 Annual Conference*, Montreal, Quebec.

U.S. Department of Commerce (2000). Fall through the Net: toward digital inclusion; a report on Americans' access to technology tools. Technical report. *USDC*, Washington D.C.

U.S. General Accounting Office (1995). Breast conservation versus mastectomy: patient survival in day-to-day medical practice and in randomized studies. Technical report. *USGAO*, Washington D.C.

Weible, R. and Wallace, J. (1998). Cyber research: the impact of the Internet on data collection. In *Marketing Research*, 10(3): 19-31.

White, E. (2000). Market research on the Internet has its drawbacks. In *The Wall Street Journal*, March 2, B4.