A Structured Approach for Reducing Bias in Surveys Using Online Access Panels

Yehuda Dayan – Survey Methodologist - Ipsos MORI
Neil Farrer – Technical Director - Ipsos MediaCT

Introduction

For many years, the gold standard of data collection in survey based research has been face to face interviews with a random sample of the target population (Yeager et al. 2009). A large dataset of such randomly selected respondents is still considered the most trusted source of information on a target population. This dominance of face to face research using probability samples has been steadily decreasing due largely to cost, practicality and speed of delivery. Increasingly there is pressure to move surveys to an online collection methodology. They are seen as more practical on these parameters.

Some online surveys have been done with probability samples of the population of interest (e.g., Moskalenko and McCauley 2009; Skitka and Bauman 2008). Here evidence suggests that data collection from probability samples via the Internet can yield results that are equally or more accurate than RDD telephone interviews or face-to-face interviewing with area probability samples (e.g., Chang and Krosnick in press; Smith 2003) possibly as a result of more candid responses in the absence of the interviewer (Dayan et al 2007).

However, the majority of online surveys are carried out with respondents collected through non random probability methods. Specifically, most commercial online surveys are run through Online Access Panels. There has been much research published indicating problems encountered when trying to compare information collected via Access Panels to the general population. A large part of these problems are associated with the samples taking part in the surveys and can be seen by errors or biases in results.

Attempts to remove these biases either through careful sample selection or via post survey weighting have had limited success (Yeager et al. 2009). One theory for the lack of success is due to there being several biasing effects that are not easily corrected via a method that attempts to correct all in one go. These can be broken down into three basic error types:

- Coverage – not everyone is connected to the Internet so cannot be online
- Self selection – not everyone on the Internet will want to join an online panel and complete surveys regularly
- Non response – not everyone who joins a panel and is invited to take part in a survey will choose to do so

In most academic and professional studies of this online panel inference problem researchers have tended to ‘bundle’ these three unique error terms into a single response probability and aim in one way or another to estimate, for any given survey, the respondent’s probability of selection from the population of interest to the achieved sample (Lee 2006).

A conceptual framework is presented in this paper that deconstructs the Online Access Panel sampling process into two main stages as follows:

- Panel Assembly - taking separate account of population biases of those that connect to the Internet and those who would join an Internet Access Panel
- Survey Sampling - taking into account panellists’ non-response error by demographic group. Additionally, further adjustment variables are also defined relating to the research topic to identify criteria that may be important for controlling the sample selection.

An immediate benefit of this framework is that it allows the model specification of the statistical adjustment to be error specific as well as to be informed by established theories for the two stages, and so potentially be more efficient. This deconstruction is parallel to that described by Groves and Couper (1998) for household non-response to non-contact and non-cooperation. Their main argument was that all well-specified response propensity models have different functional forms and so should be modelled separately. We agree and share the belief that the tendency both of the literature and of professionals to lump these phenomena together (along with the large non-response error) into one post survey adjustment model is probably one reason for the failure to achieve good, or at least consistent, estimates through the Online Access Panel platform.

Specifically, we expect that the causes of Internet connectivity and panel volunteering, the two main sources of error in panels, are different and sometimes even contradictive. For example there is a positive correlation between education and internet connectivity but within the online population there is a negative correlation between education and the likelihood of joining an Access Panel.

In this paper we show initial results of an application of the model to the GB case. We start by describing through the Total Survey Error framework the access panel assembly and survey stages. Here we focus on the specification of a Panel Assembly propensity model that includes two phases- Internet Connectivity followed by Panel Self Selection but detail how a non response stage is incorporated as well to create a Three phased adjusted estimator. For the estimation of the proposed model we require three sets of data with overlapping auxiliary data- (1) a large representative sample of the population that does not access the Internet, (2) a large representative sample of the Online connected part of the target population and (3) an Online Access panel. Using the Ipsos Online Access panel and the British National Readership Survey (NRS) we construct a panel assembly propensity model by estimating consecutively both components. We then test its effectiveness separately on an online survey recruited from the access panel applying the necessary weighting in stages.

The Online Access Panel problem through the Total Survey Error Framework

Total Survey Error (TSE) Models attempt to understand the properties of the survey process. Applied correctly, they can lead to better prediction and adjustment models. In the study of TSE components, the survey process can be broken down into four main sources of error: Frame, Sampling, Measurement and Non-response (Unit/Item).

This framework gives the statistician a structure that allows them to express statistical manifestations of individual error sources, and possibly the interdependency among these sources when they exist. It pushes our realisation that something must be done about these errors and motivates preventive strategies. Another product of a TSE perspective is the creation of Quality Indicators for the different aspects of the panel and survey process. Here we use the TSE to catalogue findings: gauge the relative contributions to total error, establishing the relative importance of (and thus the urgency in dealing with) individual components and rationally appropriating resources to deal with error sources.

Over the years the TSE model has evolved considerably (see Groves 1989, Groves et al 2004 for example). In this paper we focus our attention on the three main sources of error in Online Access Panels- Coverage, Self Selection and finally non-response error. For each of these errors we shall identify and estimate its associated bias and attempt to adjust for it through modelling the error process individually.

A sample survey process on a commercial Online Access Panel has four major conceptual and operational stages. These are described in figure 1. The errors associated with this survey process are shown in figure 2 and taken from Lee (2004)

Figure 1 – Typical sample survey process for an Internet Access Panel

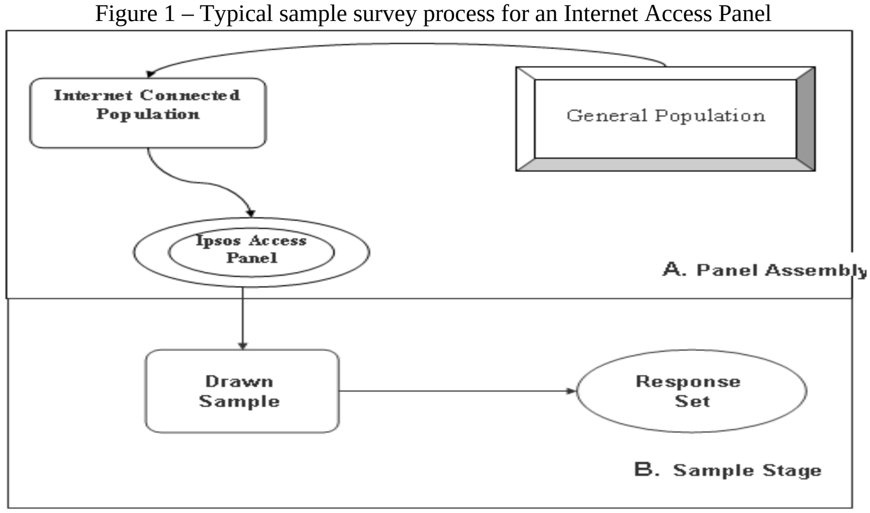
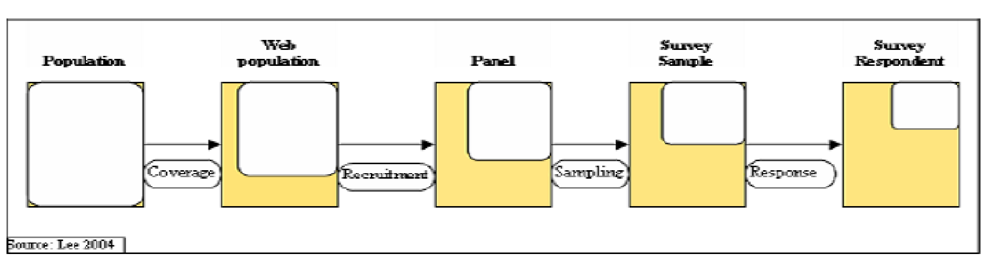


Figure 2 – Sources of error in an online survey



To this we can also add measurement error in some cases linked to how questions are asked in a self-completion online survey (Carroll, Collins, Farrer 2009).

To explore this we have created a unique dataset that merges the National Readership Survey (NRS) and the Ipsos Online Access Panel over 100 social demographics, lifestyle, internet and area level variables. This dataset includes:

- Approximately 9,600 NRS respondents who are non-Internet users and 19,800 NRS respondents who are Internet users. These respondents represent the 2008 survey dataset. Respondents from Scotland were removed for data processing reasons.
- Approximately 30,000 Ipsos Online Access Panel respondents covering England and Wales. These represented around 10% of the total online access panel base available at that time. The information available on these panellists is limited to basic demographics and limited media usage information covering radio listening. Unfortunately newspaper and other readership is not currently screened on this panel.

Our first application of the dataset will be to estimate the two main biases- Coverage and Self Selection. What are their absolute and relative sizes and what can we learn from them on the Online and Panel population?

The 2008 NRS dataset weighted to normal population targets for England and Wales shows that the estimated proportion of the population who have had access to the Internet in the last year is 71%. The places people access the internet are varied and includes home (92%), work (40%), educational establishment (13%) and increasingly via mobile phone (16%). More than nine in ten (94%) now have a broadband connection.

Internet users are referred to in this paper as the online population, non-Internet users as the offline population. The panel population profiles are taken from the latest screened Internet statistics for those panel members.

<u>Coverage and selection bias</u>

In this section we review the coverage and self selection bias. Separating and measuring these two biases will give three main benefits:

(1) Allow us to assess the size of each error, which then in relative terms will give an indication of where to put more emphasis in adjustment.
(2) The isolation of the errors will show us the direction and functional form of our adjustment. That is, we suspect that the two errors do not necessarily bias our main estimates in the same direction (over versus under estimation) and such an understanding will guide as to the components and functional form of the adjustment tool whether it be some type of weighting or matching
(3) This deconstruction can identify optimally the operational side of the panel company as to where they can focus the recruitment of new panellists.

With many demographics, it is possible to improve the self selection bias by targeted recruitment. Typically, recruitment may be via banner advertising or via email contact lists. Using the right website or list will influence the panel composition, though in general our panel recruitment follows general non-minority targeting principles.

Some demographic measures are though not possible to influence recruitment and targeted profiles of panel members may show high levels of self selection error if joining a panel is a discriminator on this variable. Having a very large pool of respondents of course does help the sample selections to take these differences into account when known.

The profile of the offline population versus the online population shows large differences in almost all major social demographic characteristics (table 1). For example, non-Internet users are more likely to be in certain regions (table 2), Female, disabled, widowed, less likely to be working, from lower social grades and older.

The online population is younger, is more likely to be single and there are more likely to be children under 15 in the household. The latter however does not show a perfectly linear relationship with number of children (table 3) and there may be interactions with ethnicity and social grade. Internet users are more likely to be working full time and be more educated (table 4).

These results reinforce our perception of the profile of online and offline populations. As the online percentage in the population increases, we expect the differences between online and offline populations will become more extreme on certain variables.

Note – in tables	Coverage Bias	=	(Online pop % – offline pop %)/ Offline pop %
	Self Selection Bias	=	(Panel % - online pop %)/ Online pop %

Table 1 – Example demographic groups of the offline, online population and Ipsos Access Panel members with bias indicators

Demographics	Offline Population	Online population	Panel	Coverage Bias %	Self Selection Bias %
Age (mean)	61.2	39.8	42.9	35	8
Income band (group)	3.8	6.7	5.7	75	14
% Main Shopper	74	63	58	14	8
Mean number of people in household	2.2	3.0	2.9	35	3
Mean number working in household	1.5	1.7	1.5	8	11
Mean number of adults aged 15+	2.0	2.5	2.3	24	9
Mean number of Children under 15	0.3	0.6	0.5	135	10
Mean number of males	1.0	1.3	1.1	33	15

Source – NRS 2008 and Ipsos Online Access Panel

Table 2 – Region at location of the offline, online population and Ipsos Access Panel members with bias indicators

Standard Region	Offline Population %	Online population %	Panel %	Coverage Bias %	Self Selection Bias %
North	8.7	4.8	5.6	45	17
North-West	12.5	12.6	13.3	1	6
Yorkshire and Humberside	11.5	9.7	10.9	16	12
West Midlands	11.7	10.0	9.0	15	10
East Midlands	9.0	8.6	10.0	4	16
East Anglia	3.7	4.7	12.9	30	174
South-West	10.1	10.5	10.6	4	1
South-East	16.2	22.4	17.2	38	23
Greater London	9.0	16.7	9.8	86	41
Wales	7.5	4.8	5.6	36	17

Source – NRS 2008 and Ipsos Online Access Panel

Table 3 – Number of children by Internet availability in household

Number of children in household	Offline population %	Online population %	Ratio Internet to non-Internet
0	86.2	65.0	0.8
1	6.2	16.6	2.7
2	4.6	13.1	2.8
3	2.0	4.1	2.1
4	0.8	1.0	1.3
5+	0.2	0.2	1.0

Source – NRS 2008

Table 4 – Qualification levels of the offline, online population and Ipsos Access Panel members with bias indicators

Qualification	Offline Population %	Online population %	Panel %	Coverage Bias %	Self Selection Bias %
Secondary school or earlier	53.9	37.3	41.3	31	11
Higher education below university degree level	26.7	27.6	19.1	3	31
University degree (Including polytechnic or college degree)	19.3	35.1	39.6	82	13

Source – NRS 2008 and Ipsos Online Access Panel

The self selection panel joining bias works differently in many instances to the coverage bias. Panel members are more likely to be female, older, have lower income, not full time working (table 1). There are some large regional differences on panel members (table 2).

Level of education is an example of where the biases work in different ways (table 4). The online population is educated to a higher level and is almost twice as likely to have a university degree than the offline population. Those educated to secondary school are under-represented (54% offline versus 37% online). On joining an online access panel, those in the middle education band to Higher Education level below degree are under-represented. One reason may be that the incentive effect may have appeal to the less educated group. Such a difference highlights our issue around deconstructing the errors.

There are also differences between offline, online and panel populations on demographic areas that cannot be corrected by normal demographic weighting. For example, the presence of landline and mobile phones shows large differences relating to panel selection (table 5). Mobile phone only households are a relatively new component of the survey mix and this segment of the population has recently been described in the UK market (Stead, Farrer 2009).

Table 5 – Telephone ownership of the offline, online population and Ipsos Access Panel members with bias indicators

Phone Ownership	Offline Population %	Online population %	Panel %	Coverage Bias %	Self Selection Bias %
Mobile Only	10.3	8.8	2.8	15	68
Landline only	30.2	3.1	1.9	90	39
Both	54.2	83.2	95.2	54	14
Neither	3.1	.4	.1	87	75

Source – NRS 2008 and Ipsos Online Access Panel

The difference between offline and online profiles is a general indication of differences in profile around technology based ownership. This difference is shown in table 6.

The difference in profile of telephone ownership between online and panel populations is an indication of the more transient nature of mobile phone only households. We know they have more recently moved house for example and it may be that a panel that is recruited over a long period cannot keep up with frequent changes on such measures that of course results in biases in both coverage and self-selection.

Table 6 – Technology items owned on the offline, online population and Ipsos Access Panel members with bias indicators

Items owned	Offline Population %	Online population %	Panel %	Coverage Bias %	Self Selection Bias %
Home Cinema	33	53	52	54	-2
Game Console	15	49	52	218	7
PDA	1	10	13	830	33
PC/Laptop	29	92	84	201	-9
MP3 Player	12	56	60	333	8
Web Cam	4	26	39	591	51

Source – NRS 2008 and Ipsos Online Access Panel

There are also differences between offline, online and panel populations on survey measurements related to media research.

Examples for national daily newspapers are shown in table 7. A survey suffering from online coverage bias would overestimate some titles and underestimate others. Overall though, offline populations are more likely to be readers of national daily newspapers than online ones.

Table 7 – Internet and non Internet user profiles for daily newspapers – measure is Almost Always Frequency

Newspaper	Offline population %	Online Population %	Difference %	Coverage Bias %
Financial Times	0.1	1.0	0.8	706
The Guardian	0.6	2.6	2.0	320
The Independent	0.3	1.3	0.9	277
The Times	1.6	4.0	2.4	153
The Daily Telegraph	3.6	3.7	0.1	3
Daily Mail	11.1	8.3	2.7	25
Daily Express	3.9	2.0	1.8	47
Daily Mirror	10.1	5.5	4.6	45
The Sun	15.6	12.6	3.0	19
Any National Daily	43.3	35.3	8.0	18

Source NRS 2008

Finally, similar results for listening to national UK radio stations are shown in table 8 where we are able to compare offline, online and panel profiles. This table is important as these are the only media statistics that are presently collected of all Ipsos Online Access Panel members. As expected, the radio stations appealing to younger listeners (BBC Radio 1) and men (BBC Radio 5 live) show the largest differences between offline and online populations. The upmarket radio stations of BBC Radio 3 and Classic |FM show the largest self selection bias.

Table 8 – Radio listening in the offline, online population and Ipsos Access Panel members with bias indicators

Radio station	Offline Population %	Online population %	Panel %	Coverage Bias	Self Selection Bias
BBC Radio 1	10	30	29	220	-6
BBC Radio 2	20	26	34	33	27
BBC Radio 3	3	4	5	35	37
BBC Radio 4	13	18	19	42	4
BBC Radio 5 live	5	12	15	126	20
Classic FM	9	11	16	28	48
Talk Sport	3	6	7	158	8

*Source – NRS 2008 and Ipsos Online Access Panel

Methods for Reducing Error in Online Panels

The challenge of reducing the bias associated with online access panels has taken several shapes and forms. The following are possible approaches.

In these examples, we signify *y* as the attribute we are measuring and *p* the propensity to participate in a survey.

Direct estimation

A simple approach for a set of panel results would be to use the panel mean directly. Bethlehem (2009) and others have shown that under a general set-up the expected bias of the direct mean of the panel is simply $C(p, y)$, the covariance (unstandardised correlation). This suggests that the bias is determined by the average propensity in the population, where the higher the likelihood of population members to participate, the lower the bias. The bias is also a function of the relationship between the target attribute and the volunteering propensity. Thus the bias will diminish the more individual panel volunteering probabilities are similar, or the more similar the attribute measured is across the population. The bias will diminish also when the association between the attribute and the probability is low. This last case explains why some attributes may have a high online survey potential whereas others more technology oriented are unlikely to be estimated well by an unadjusted panel sample.

This is not an appropriate method for readership and media research as shown by the differences between offline and online populations in table 7.

Population based auxiliary information

In most surveys researchers do attempt to improve on the direct estimation and use known population information. This information is usually the population profile on publicly available key social demographic indicators such as age, gender, region and social grade or some equivalent measure. The aim is to adjust the panel sample before and / or after data collection on these variables.

Post survey weighting adjustments are popular methods in statistical practice that use auxiliary data to (a) reduce the variance of sample estimates, by taking advantage of the relationship between the target attribute *y* and the weighting variables in what is in fact an approximate Analysis of Variance type of procedure, and (b) it can be used to implicitly model the unknown response probability behaviour by in effect classifying the population into homogeneous groups where response behaviour within these classes are similar. An optimal weighting design will attempt to capture both (a) and (b). This is sometimes called Double Robustness (Litle and Vartivarian 2003).

There are some weaknesses with this method. First the population counts of the weighting classes need to be known as well as collected in the survey itself. There is also the fear of running into large stratification matrices. An even more serious threat is that there is no real reason to believe that both *y* and the response propensity variation are grouped within the same strata. Lastly expert knowledge will probably contradict that the same functional form, a simple ANOVA model in this case, is appropriate for both mechanisms- *y* and *p*.

Yeager et al. (2009) have produced an extensive comparison of post stratification or outgo balance on population figures.

Using auxiliary information from a reference sample

A key assumption when considering whether an adjustment is effective or not is the statistical condition Missing at Random (MAR). MAR simply assumes that the available auxiliary information (here the weighting variables) explain the relationship between *y* and *p* well. This means that our weighting variables should be able to predict the causes of response propensity as well as being strong drivers of the attribute *y*.

However, given that normally our available information of the population is limited to a small number of demographics that may or may not be interlocked and can come from a variety of different sources, this can be an unrealistic assumption resulting in ineffective adjustment. Post weighting the level of bias will remain the same.

A common approach taken to try and meet the MAR assumption in recent years has been to run a smaller reference offline survey parallel to the online one. This survey is normally a random probability one in which we assume results are unbiased or, more realistically, are much smaller. Through this reference sample the researcher may collect data that are possibly more relevant to the response probability of a respondent to an access panel.

Bethlehem (2008), Schonlau et al (2002) and others have described in detail the strengths and weaknesses of this approach. When the reference sample is used simply as a means of estimating the population profile we achieve an adjusted estimator which is of the same form as the post stratification estimator above. That means that the quality of the bias reduction is a function of the explanatory power of the weighting variables which we assume to be equal to or greater than the publicly available socio demographics. On the other hand this estimator results in sharp reduction of the effective sample size. In fact the standard error of the adjusted estimator is of the order of the reference sample. Schonlau et al (2002) indicate that in terms of Mean Square Error (MSE), a measure of accuracy taking into account both variance and bias, a more efficient estimation procedure would be to allocate the Online panel cost to a larger probability based sample.

This trade off between bias and variance/standard error is identical for any method where a larger convenience sample is adjusted by a smaller random probability sample.

Matching panellists to an external database

One slight variation on the previous approach is to sample from the Online Access Panel using a ‘mirror’ sample drawn from available governmental or commercial databases that are, at least, seen to be more representative of the target population. A good example of this method is YouGov Polimetrix in the USA (Rivers 2005). In this approach, before sampling from the company panel, they first find a database that can enumerate the target population. This database can be considered as a sampling frame.

On a readership study, the target would be the general population (or all adults) and for enumeration of this target population to use consumer databases compiled by commercial vendors such as CACI or Experian. Then create a target sample from the population list using the necessary stratification and selection rules. Finally match each member of this target sample to a similar member from the online panel by means of some statistical distance metric.

This method does attempt to overcome two weaknesses described previously by (a) using a usually very large platform (i.e. database) as a reference sample and so there is only a small reduction in the effective sample size and (b) getting closer to the MAR assumption by using a possibly richer population platform data source.

However, we still may be cautious of whether the commercial database is indeed an unbiased sample of the population itself, the accuracy of this database and question whether the matching variables and matching technique really capture the necessary profile of the population resulting in a substantially smaller bias. There are a number of such population lists available in the UK and they are often used as sample discriminators for offline samples.

Propensity score adjustment

The propensity score adjustment is simply an optimal approximation to an exact adjustment method. It is a method initially suggested by Rosenbaum and Rubin (1984) and has since emerged as one of the most popular adjustment methods in the fields of medicine, economics and social research. It has recently been adopted as a possible adjustment method in online research. Lee (2006) and Duffy (2005) provide comprehensive reviews and examples.

Consider the example shown in table 9 where there are eight population profile variables from an available large random probability collected database.

Table 9 – Example of population variables that may be used in weighting adjustment

Adjustment Variable	**Number of categories**
Age bands	9
Gender	2
Income Bands	10
Social Grade	6
Standard Region	8
Employment status	6
Education Level	5
Social involvement index	8
Household Tenure	5

It is reasonable to consider that these adjustment variables jointly, and not separately, may explain a large part of why a random unit in our population decides to join an Online panel. For example, it is the combination of age, income and education that defines the 'Silver Surfers' of the population rather than independently. In such a case we would then aim to recreate as a post adjustment the joint distribution of the population on these individual and household attributes. However, even with a relatively small number of 8 attributes we have reached a problem normally defined as the curse of dimensionality. Simply put, the 8 variables create over a million adjustment cells which is not possible.

The propensity score aims to resolve this problem. By definition, the propensity score is simply the conditional probability of being a member of the panel rather than in the reference sample. Normally we use a logistic regression to estimate this probability. When applying this method we use the estimated propensity score rather than the raw variables for adjustment.

Thus, given the additional flexibility, the propensity score allows in terms of volume of adjustment cells, we would expect the MAR assumption to be more plausible. However, usually only a small reference sample is used for propensity score estimation and there could still be a strong reduction in effective sample size.

Summary and Conclusions

From our review we can note that all current practices model the panellist propensity to participate in a panel survey in one step. Furthermore, when modelling this process the analysts usually use, to meet the MAR assumption, available population statistics or smaller reference samples. Lastly, in most cases the practitioner applies a combination of post stratification/ rim weighting on population percentages and some statistical multivariate adjustment such as propensity score but not with a clear separation as to which model *y* and which model *p* as well as the order of the two adjustments, see for example Lee (2004), Dever (2008) for practical examples.

In light of this we propose to track the modular operational process of sampling from the Ipsos Online Access Panel, and decompose the inferential model accordingly, using a rich auxiliary dataset to meet the MAR assumption and a large nationally representative reference sample with an online/offline indicator. Finally we offer a structured post panel adjustment approach to incorporate modelling *y*.

Our method - 'Tracking the error trail'

We propose to update the well established modern model-assisted survey sampling approach to our panel correction question. The model is a two phased or double sampling method. We propose an estimator assuming an unknown sampling mechanism. Here we briefly describe the link between the conceptual TSE framework described previously to this model. We then present an estimator which is unbiased under the model assumptions which has also an approximately unbiased estimator of the variance- used for the calculation of confidence intervals.

The TSE framework described earlier suggests that the path of a panellist to finally accept an invitation to a survey and respond to the web questionnaire can be conceptualized as a series of conditional events. Neyman (38) was the first to offer an estimator based on a phased data collection design- at first we select, say, *n* out of *N* population units with a certain probability and ask them a small number of questions. Then, in the second phase we select *m* out of the *n* sampled units with a certain random probability design (e.g. after stratifying on information collected in first phase) for a more detailed survey. So instead of only one sampling error as we have in the normal single-phase case, we have two sampling errors. Further research in the field then updated this estimator to include the option of adding auxiliary information and also uses as a framework for non response, see Sarndal and Swensson (1987).

The classic single-phase estimator divides each sampled unit by its inclusion probability. The two-phase estimator is an extension and divides each observation by the product of (a) probability of being selected in the first phase and (b) the conditional probability of being selected in the second phase given[1].

As noted earlier the errors we identify in the panel survey process are; Coverage, Self Selection, Sampling and Response. With these data collection phases our aim is to allow derivation from the respondent set of panellists achieved to the total target population. After modelling, the correction process will be an exact reverse picture of the survey process. At each stage we adjust for the specific error.

In its simplest form our estimate of the total population for attribute *y* would then be

$$\hat{t}_{y\pi^*} = \sum_r y_k / \pi_k^* \quad \text{where} \quad \pi_k^* = \theta_k \rho_k \psi_k$$

This means that each panellist which has responded to our invite and has reported on their individual value of variable *y* (say readership of Newspaper X) will simply be divided by the product of three probabilities.

- θ_k is the internet acceptance probability- the propensity of an individual *k* in GB population for having used the Internet at least once in the last year.
- ρ_k is the individual conditional self selection probability. That is the propensity an individual *k*, who uses the Internet at least once a year, has to join the IIS online access panel.
- ψ_k is the probability of panellist *k* to reply to the survey invitation as some panellists are more likely to reply than others.

To estimate these unknown theoretical probabilities we have created a three layered dataset combining large representative samples of the British Offline and Online populations from the NRS and the Ipsos Online Access Panel.

The Ipsos Access Panel part of the dataset includes a sample of active panel members who have provided answers to recent screening surveys on general topics. This information, along with responses at original recruitment information act as the basis for the access panel part of the dataset. This restriction of a complete set of responses to such surveys has reduced the number of panellist records available from over 350,000 to approximately 30,000.

After sourcing the two data sources, a long data processing stage followed where a considerable amount of recoding of the two sources of data into common values was required. For example individual characteristics such as Social Grade, qualification levels, income and even ethnicity are asked in different ways or have different categories of answers.

[1] Which is not equal to the individual inclusion probability.

An immediate concern is that this data processing (recoding) stage, along with possible mode errors (Face to Face vs. Self Completion) may confound our self selection results. We have taken the step of sampling around 2,000 Ipsos Online Access panellists of the 30,000 we are using in this exercise and sending them a survey asking suspected questions with possible measurement in a common format[2].

We estimate Connectivity (the probability of accessing the internet) by a logistic regression model where the dependent variable is a flag indicating whether a random NRS respondent has or has not accessed the Internet in the last 12 months. Estimation of the model is done over the available created NRS-Access Panel dataset.

The Self Selection probability (the probability of an individual joining the access panel) is estimated also by a logistic regression. Here the target is the flag variable indicating whether the unit is an 'Online' NRS respondent or an IIS panellist.

Lastly, the response probability which measures how likely an IIS panellist is to reply to an invitation to complete a survey by Ipsos is estimated by a combination of logistic regression and classification trees on long term panellist behaviour information collected by the Ipsos Access Panel management team[3].

Testing the new estimation model on a combined NRS and Access Panel dataset

The data processing procedure created a combined NRS-Access Panel dataset of over 60,000 observations on more than 100 individual, household and (small) area level variables (these variables have been derived from Office of National Statistics based estimates and geodemographic lists such as Mosaic). We use this dataset to estimate two propensity models- Connectivity to the Internet and Self Selection to the Access Panel.

The estimation of these propensities was through a sequential logistic regression modelling procedure. At first we run a logistic regression to estimate the propensity to connect to the Internet using the Online and Offline parts of the weighted dataset. Following this we segment the 'raw' estimated propensities into 8-10 propensity groups. This segmentation is a standard approach that protects from likely model misspecification and random errors. These segments create a broad more stable set of coverage error weights. We than apply these weights when running the second (conditional) Self Selection logistic model. The resulting estimated probabilities are also segmented into 8-10 propensity groups.

The first phase (Coverage) logistic regression model was run over 30,000 available observations from the NRS while the estimation of the conditional logistic regression model between the online NRS and Ipsos Access Panel dataset was estimated at over 50,000 observations of which 30,000 were IIS panellists and 20,000 were NRS respondents who are online. External datasets covering geodemographics were also added to try and increase the pool of information available.

It has been assumed that there is no overlap between the two data sources. Whilst not strictly accurate, the online access panel does form only a very small percentage of the population (30,000 available in total represents less than 0.1% of the population).

We end with two sets of profile distributions which can then be used for weighting (equivalently for sampling or matching) to adjust the two profiles (see Lee 2006 for more detailed review of the basic steps of the weighting approach).

As with any modelling exercise there are a number of factors to take into consideration. For propensity score type estimation, there is a 'triangle' trade off of bias and variance between:

(1) The number of variables to include in the model and their functional form (e.g. do we include interactions),
(2) The number of segments to group these estimated probabilities
(3) The shape of this grouping. That is, whether equal sizes in terms of observations or equal distance in terms of propensity scores.

More variables and interactions and smaller propensity segments should increase the accuracy of the estimates however will probably increase the variance and so consistency of results.

The aim is to find an optimal balance between an increase in the quality of prediction (and the plausibility of approximating the MAR assumption) while defending against an increase in sensitivity to model misspecification and variance of the propensity scores.

---

[2] The sampling stage has been completed on the 9th September 2009 and with it we will easily cross examine direct and recoded responses to inspect this possible data processing error.
[3] We do not use this adjustment here as the experiment is directly on the panel.

Results of the Coverage Bias adjustment models

The Connectivity propensity is the basis for our coverage adjustment stage and is modelled using the NRS only part of the data. The coverage model derived from logistic regression has identified key variables shown in table 10.

Table 10 – Key variables to correct for Coverage Bias (Internet Connectivity)

Internet Connectivity Model	
Individual Age	Adults With No Low Qualification Rate (ONS)
Phone Owner	Net Income Estimate1 (ONS)
Social Grade	Rank Education Score Rank Areas (ONS)
Employment Status	
Qualifications	Topic of Interest
House Tenure	Photography
Gender	Personal Finance
Income	Business/ Company news
Ethnicity	Education
	Jobs/ appointments

This model achieved a high level of classification rate (84% overall, 91% correct classification for the online population and 72% for the Offline segment) and an average of 90% bias reduction on the variables in the model.

Table 11 shows, for a selection of daily and Sunday newspapers, the following

- The original NRS estimate for the offline population using the original survey rim weighting to match the sample estimates on the population profile of age, gender, social grade and standard region. We are interested in looking at how well a random probability online sample can hope to estimate the readership behaviour of a non-covered offline segment of the same population.
- The original NRS estimates for the online population
- The unadjusted offline bias is simply the difference between the online and the offline populations. For example, the first title shown has a readership level of 0.1% in the offline population and a readership of 1.0% in the online population. This gives a bias of 0.8% between online and offline populations and a relative difference of 706%. We shall use the average relative absolute difference as a measure of a group coverage bias, the coverage bias of the unadjusted Online sample is 199% for the daily newspapers selected and 135% for the Sunday newspapers
- The adjusted bias after applying a rim weight on the online population to match the profile of the weighted offline sample. Following our previous example, the first title listed has coverage bias reduced dramatically and in relative terms it has dropped to only 92%, a dramatic change which implies a strong association between the likelihood to read the specific paper with the fundamental socio-demographic indicators used in rim weighting. The average coverage bias of daily newspapers reduced so 79% and Sunday papers to 77%.
- The propensity score adjusted bias using the variables created from the logistic regression model. We are able to adjust even further by loading the adjustment model with far more variables and multivariate relationships than the rim weighting can cope with. The more aggressive propensity adjustment further reduces the relative bias of our example first newspaper from 92% to 11%. This means that an adjusted Online sample will overestimate the readership levels of the Offline population by only 10% a sizable reduction from the initial 706% we started off with. The average coverage bias drops to 41% and 55% for daily and Sunday newspapers. Whilst this shows a further improvement in the offline modelled readership, not all titles are improved above the rim weighting solution.

The same analysis looking at the coverage bias for the national radio station listening estimates is shown on table 12. The same sequence of improvement for rim adjusted and propensity score adjusted relative bias is seen with the average bias across all stations reducing from 92% to 53% to 27%.

Table 11 – Coverage bias statistics for daily and Sunday newspapers

Newspaper	NRS original Offline Estimate %	NRS Online Estimate %	Coverage Bias					
			Unadjusted		Adjusted Rim		Propensity Score	
			Bias %	Relative %	Bias %	Relative %	Bias %	Relative %
Financial Times	0.1	1.0	0.8	706.0	0.1	92.4	0.0	10.5
The Guardian	0.6	2.6	2.0	320.3	1.2	194.2	0.7	94.5
The Independent	0.3	1.3	0.9	277.3	0.5	156.2	0.2	51.6
The Times	1.6	4.0	2.4	152.8	1.8	113.9	1.8	86.7
Daily Mail	11.1	8.3	-2.7	24.8	2.9	26.1	1.1	8.8
Daily Express	3.9	2.0	-1.8	47.5	0.3	6.6	-1.2	26.2
Daily Mirror	10.1	5.5	-4.6	45.1	-1.3	13.1	-1.4	15.5
The Sun	15.6	12.6	-3.0	19.3	-5.1	32.5	-4.4	33.2
Any National Daily	43.3	35.3	-8.0	18.4	1.3	3.0	-2.2	5.0
Average National Daily				199.1		79.4		40.9
Independent on Sunday	0.3	1.1	0.9	345.0	0.4	150.3	0.3	98.5
The Observer	0.7	2.5	1.8	256.1	1.2	170.8	1.1	138.7
The Sunday Times	2.0	6.6	4.6	237.5	2.9	146.6	1.4	56.2
The Sunday Post	1.2	0.2	-1.0	80.3	-0.7	54.5	-0.7	52.5
Sunday Express	3.9	2.2	-1.7	42.6	0.3	6.9	-0.6	13.0
Sunday Mirror	8.7	5.4	-3.3	37.9	-0.9	10.2	-0.7	9.2
The People	4.7	1.9	-2.9	60.8	-2.2	46.3	-1.7	41.5
NOTW	15.0	15.0	-3.2	21.5	-5.0	33.6	-3.5	28.0
Any Sunday	39.5	39.5	-4.4	11.2	3.0	7.6	1.5	3.7
Average Sunday				135.2		77.4		54.7

Table 12 – Coverage bias statistics for national radio stations

Radio station	NRS original offline estimate %	NRS Online Estimate %	Coverage Bias					
			Unadjusted		Adjusted Rim		Propensity Score	
			Bias %	Relative %	Bias %	Relative %	Bias %	Relative %
BBC Radio 1	9.6	30.7	21.1	219.8	2.8	29	0.2	3.1
BBC Radio 2	20.1	26.7	6.6	33	8.4	41.6	4.6	21.4
BBC Radio 3	2.7	3.6	0.9	34.6	1.9	69.7	1	27.6
BBC Radio 4	13	18.3	5.4	41.5	8.1	62.5	4.7	27
BBC Radio 5 Live	5.4	12.1	6.8	126.5	2.9	53.4	0.9	16.2
Classic FM	8.5	10.9	2.4	27.8	5.9	69	4.1	39.1
Talk Sport	2.8	7.1	4.4	157.9	1.2	43.8	0.3	15
None of These Stations	29.2	14.8	-14.4	49.2	-11.3	38.7	-4.6	16.4
Average station				91.6		52.7		21.3

Other variables covered on the NRS questionnaire have also been looked at, including financial variables, car ownership, business information and so on. All tend to follow a significant trend of improvement. However in no measure can we say that all coverage biases are removed.

Results of the self selection bias adjustment models

Whilst not as large as the coverage biases observed, there are significant self selection biases reported previously between being online and joining/ participating in an online access panel.

With the aim of identifying the size of these biases, along with theories on the causes of why people in the population participate in volunteer access panels, we have modelled Self Selection to the online access panel. The estimation of the model was with the newly constructed online NRS-Access Panel dataset described previously.

A logistic regression model predicting the self selection variable (Online vs Panel) as the dependent variable was run conditioned on the calculated coverage weights of the previous phase. The variables entering the model are shown in table 13.

Table 13 – Key variables to correct for Self Selection Bias (Response)

Self Selection to the IIS Panel Model		
Individual Age	Frequency of using Internet	Mosaic Group
Phone Owner	Access Work	Full Time HH
Social Grade	Access Education	HH Size
Employment Status		
Qualifications		
Gender	Home Cinema	Adults With No Low Qualification Rate (ONS)
Income	Game Console	Net Income Estimate1 (ONS)
Ethnicity	PDA	Rank Education Score Rank Areas (ONS)
Marital Status		

The resulting model is a combination of 17 variables that include individual characteristics such as income, marital status, technology and Internet related issues – Internet frequency of usage, early adopter indicators and household and three area level indicators.

There is a difference between this model and that of the coverage, in size and functional form- in the sense the independent variables, even if similar have different size and sign of coefficient. This implies that there is merit in the separation of the two processes as they are unique.

Compared to the coverage model, the summary statistics of the model effectiveness are less favourable. The logistic regression overall classification rate is 76%, with only a 53% correct classification of the Online population and almost 90% on the panel. This imbalance is a symptom of a relatively weak model. This and other methodology concerns may explain the problematic adjustment results shown here.

Table 14 compares three different possible estimations of the online population against the panel participating population covering radio listening information.

- An unadjusted figure - direct use of the online access panel
- Applying a rim weighting adjustment on age, gender, social grade and standard region
- The self selection propensity score model from variables as selected in table 13

As noted previously, our dataset does not currently include readership information from the online access panel so we cannot compare the effect of the self selection adjustment on these key variables.

Table 14 – Self selection bias statistics for national radio stations

	Self Selection Bias					
Radio station	Unadjusted %	Relative %	Rim weight Adjusted %	Relative %	PS Adjusted %	Relative %
BBC Radio 1	-1.9	6.2	-2.0	6.6	-2.2	7.5
BBC Radio 2	7.2	27.4	5.3	20.0	6.3	20.6
BBC Radio 3	1.3	36.8	2.0	57.8	1.9	37.9
BBC Radio 4	0.8	4.2	1.8	9.8	0.9	3.5
BBC Radio 5 Live	2.4	19.7	2.7	22.2	4.0	28.5
Classic FM	5.1	47.7	5.6	51.9	6.3	45.4
Talk Sport	0.5	8.4	0.6	10.1	0.7	12.8
None of These	-2.2	14.3	-2.7	17.5	-1.9	16.1
Average station		21.5		25.5		22.3

The average self selection bias is much smaller than the coverage bias shown in table 12. The different adjustments do not show any improvements in the bias. In fact, the unadjusted estimate looks the most promising of the three compared here.

Similar patterns of difference are observed when looking at technology ownership and financial product usage that are asked on both datasets.

There are three possible reasons for these weak results where the pre and post adjusted Self Selection bias are of the same magnitude.

(1) In many cases the variables described in this document involved at least some level of recoding to allow the merging of datasets. This recoding may have resulted in a confounding of errors, where differences between Panel based statistics and NRS Online respondents based statistics are actually due to recoding difficulties - or data processing error.
(2) The question and other external information available for this part of the modelling is much more limited than that available for the coverage error estimation. Perhaps other variables covering the softer personality traits relevant to joining a panel may yield better results.
(3) It is possible that the cause of these differences is due to classic mode effect between a CAPI face to face NRS survey and an online recruited panel

Combining the two error models

Table 15 shows the results for radio listening when the two error models are combined. Also shown are the population estimates for these variables derived from NRS.

The first Access Panel estimate is derived simply from a common rim weighting solution of Age, Gender, Social Grade and Standard Region. Profile targets are taken from the NRS sample[4]. The second is created using the three phased model as described in this paper.

Whilst some radio stations (for example BBC Radio 1) show a larger difference with our new weighted estimated, most show a considerable improvement with an average relative difference of 29% with the three phased weighting compared to 51% with the simple rim weighting.

[4] We have tested also the addition of Internet frequency and Broadband penetration to the Rim weighting with almost identical results

Table 15 – Comparison results of normal rim weighting and three phase weighting for radio listening – NRS versus online access panel

		Normal Rim weighting			Three phase weighting		
Radio station	NRS combined online-offline estimate	Access panel Estimate %	Difference %	Relative Difference	Access panel Estimate %	Difference %	Relative Difference
BBC Radio 1	23.9	23.6	0.3	1.3	22.0	1.9	7.9
BBC Radio 2	24.4	33.6	-9.2	37.6	28.8	-4.4	18.0
BBC Radio 3	3.3	7.0	-3.7	113.3	5.1	-1.8	54.5
BBC Radio 4	16.6	22.8	-6.2	37.5	17.9	-1.3	7.8
BBC Radio 5 Live	10.1	14.2	-4.1	41.1	14.6	-4.5	44.6
Classic FM	10.1	19.9	-9.8	96.7	16.4	-6.3	62.4
Talk Sport	5.1	6.5	-1.4	27.2	5.4	-0.3	5.9
None of These	19.6	13.1	6.5	33.0	14.2	5.4	27.6
Average radio station				50.7			28.7

Conclusion and further work

We believe that to improve the quality of access panel based survey inference a structured, theoretical sound and transparent approach needs to be taken. This work is an early contribution of a long period of analysis being undertaken by Ipsos MORI to attempt to understand the nature of online access panels and how to compensate for the known bias in estimates when compared to the general population. By constructing a unique dataset we have accepted our initial hypothesis that the coverage and self-selection biases are often different and sometimes work counter to each other when looking at population characteristics of the offline, online and panel participating populations. Our work has shown that it is possible to identify the scale of these errors and that, over most of the attributes measured, the coverage bias is of higher magnitude compared to that of self selection. We have also presented our initial attempts to reduce the total bias by applying a phased propensity score weighting procedure that looks to offer some improvement compared to more conventional approaches.

The coverage error, while post adjustment is still sizable over some measurements is, we feel, well understood and a focus in future work will be to improve on the theoretical model, the model specification and consequently the operational aspects to reduce this error further. Improvement has not been seen for the self-selection error and there remains ambiguity regarding the causes of the underperformance in this model.

One drawback of using the Total Survey Error framework is that, the more accurate and realistic you are theoretically, the more complex the empirical work of applying such models involves. The online access panel bias correction problem covers many dimensions and so much work still lies ahead.

In the first instance we wish to understand the results of the self selection adjustment. Its ineffectiveness may be a result of either (1) questionnaire wording differences we have failed to control, (2) lack of variables that can explain the relationship between the probability of joining an access panel and the different media measurements examined or (3) model specification. We are in the process of launching a new survey that will allow us to identify whether it is questionnaire wording that is the problem and additional information collected in future will give us a wider range of variables to understand the causes of respondents to join the panel. A final possibility of a straight forward mode effect between offline interviewer administered and online self-completion surveys that results in variation is very possible and one that would be most difficult to untangle.

In addition we are keen on examining additional aspects of the new model that have not been explored empirically. First, we have only superficially attempted to take advantage of the modelling y directly by regression estimation (which is equivalent to Calibration). The new model also provides a closed form variance estimator to be used in the calculation of confidence intervals, we shall examine through simulation how well these confidence intervals behave, specifically whether a 95% calculated confidence interval in fact covers the real readership figure 95% of the times. Lastly we shall compare more carefully the benefit of our phased approach to a single stage propensity score adjustment procedure.

References

Bethlehem, J. . (2008) . How accurate are self-selection web surveys? . Statistics Netherlands, 2008

Bethlehem, J. . (2009) . Applied Survey Methods: A Statistical Perspective (Wiley Series in Survey Methodology) . Wiley - Interscience: Haboken

Dever, J. A., Rafferty, A., Valliant, R. L. . (2008) . Internet Surveys: Can Statistical Adjustments Eliminate Coverage Bias? . Survey Research Methods, 2, 2, pp. 47-62

Duffy, C., Smith, K., Terhanian, G., Bremer, J. . (2005) . Comparing data from online and face-to-face surveys . International Journal of Market Research, 47, 6, pp. 615

Carroll, Collins, Farrer (2009) “Presenting Publication in Online Print Audience Questionnaires” presented in WRRS, Valencia 2009.

Chang, LinChiat, and Jon A. Krosnick. In press. “National Surveys Via RDD Telephone Interviewing vs. the Internet: Comparing Sample Representativeness and Response Quality.” Public Opinion Quarterly.

Dayan, Yehuda, Carina Schofield and A.J. Johnson 2007 “Responding to sensitive Questions in Surveys: A Comparison of Results from Online Panels, Face-to-Face, and Self-Completion Interviews “WAPOR 60th Annual Conference “Public Opinion and the Challenges of the 21st Century” Berlin, September 19—21.

Groves, R.M. (1989). Survey Errors and Survey Costs. New York: John Wiley and Sons.

Groves, R. and R. M. Couper (1998). Nonresponse in Household Interview Surveys. New York, NY: Wiley.

Groves, R.M., F.J. Fowler, M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau. (2004). Survey Methodology. Hoboken, NJ: John Wiley and Sons. [ISBN 0-471-48348-6]

Lee, S. (2004). Statistical Estimation Methods in Volunteer Panel Web Surveys. Unpublished Doctoral Dissertation. University of Maryland, Joint Program in Survey Methodology.

Lee, S. (2006). Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys. Journal of Official Statistics, Vol. 22, No. 2, pp. 329–349.

Lerner, Jennifer S., Roxana M. Gonzalez, Deborah A. Small, and Baruch Fischhoff. 2003. “Effects of Fear and Anger on Perceived Risks of Terrorism: A National Field Experiment.” Psychological Science 14: 144-150.

Little, R. J. & Vartivarian, S. (2005), ‘Does weighting for nonresponse increase the variance of survey means?’, Survey Methodology 31(2), 161–168.

Moskalenko, Sophia, and Clark McCauley. 2009. “Measuring Political Mobilization: The Distinction Between Activism and Radicalism.” Terrorism and Political Violence 21: 239-260.

Malhotra, Neil, and Alexander G. Kuo (2008). “Attributing Blame: The Public's Response to Hurricane Katrina.” The Journal of Politics 70: 120-135.

Neyman, J. (1938). Contribution to the theory of sampling human populations. Journal of the American Statistical Association 33, 101{116.

Rivers, Douglas (2005) “Sample Matching”. YouGov Polimetrix white paper. Can be found at http://www.polimetrix.com/.../Polimetrix_Whitepaper_Sample_Matching.pdf

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, 387, pp. 516-524

Sarndal, C.-E. and B. Swensson (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. International Statistical Review 55 (3), 279-294.

Skitka, Linda J., and Christopher W. Bauman. 2008. "Moral Conviction and Political Engagement." Political Psychology 29: 29-54.

Smith, Tom W. 2003. "An Experimental Comparison of Knowledge Networks and the GSS." International Journal of Public Opinion Research, 15: 167-179.

Stead, Farrer (2009) "The Mobile Generation – Reaching Respondents on the phone" Ipsos MORI research publication. Can be found at http://www.ipsos-mori.com/researchpublications/publications/publication.aspx?oItemId=1200

Schonlau, M., R. Fricker, and M. Elliott (2002). Conducting research surveys via e-mail and the web. Monograph 1480, RAND Corporation. Accessed from http://www.rand.org/pubs/monographreports/MR1480.

Yeager et al. 2009 "Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples"; with Krosnick, J. A, Chang, LinChiat, Javitz, Harold S., Levendusky, Matthew S. Simpser, Alberto and Wang Rui. Stanford University August, 2009. http://comm.stanford.edu/faculty/krosnick/Mode%2004.pdf