Worldwide Readership Research Symposium 2005

USING DATABASE OVERLAYS TO DETECT NON-RESPONSE SURVEY BIAS

Roland Soong, Lindsey Draves & Hugh White, KMR/MARS

Synopsis

Using a database overlay on a sample from the MARS OTC/DTC Pharmaceutical Study, we document how the presence of people who like to participate in surveys in general and are interested in the survey topic may bias overall survey results.

Background

Most readership studies are based upon surveys. Such surveys usually begin with a pre-designated sample which is contacted to obtain usable data. As much as we would like usable data from the entire pre-designated sample, the social reality is that only a portion will cooperate. The reasons include lack of time, annoyance with the process, lack of topic interest, concern for privacy, cultural habits and so on. Thus, the pre-designated sample is divided into the responders and the non-responders.

If the percentage of non-responders is small (say, 5%), then their absence is unlikely to affect the survey estimates by large amounts. These days, in many surveys around the world, the percentage of non-responders may be over 50% or 60%. By itself, a low response rate is not fatal. Bias occurs when there are differential response rates among groups which also differ with respect to the survey variables (e.g. print readership, product usage, etc).

Bias Correction by Weighting

Professional survey researchers are aware of the potential for survey biases. The difficulty is that, by definition, they draw little or no information from the non-responders themselves. Therefore, they have to use indirect approaches. The most common approach is geo-demographic weighting. For the survey universe, there are known universe estimates drawn from census sources with respect to the geo-demographic variables (such as geography, age, sex, income, education and so on). The collective experience is that these geo-demographic variables are correlated with readership survey variables (such as print readership, product usage, etc). Hence, the sample of responders is weighted to these geo-demographic universe estimates (see Sharot (1986) and Kish (1992)). This is a standard tool for the professional survey researcher.

However, geo-demographic weighting only corrects the immediately apparent biases, because of the existence of a reliable set of universe estimates. In that sense, geo-demographic weighting serves a cosmetic, albeit demonstrably useful, fix on the obvious. Meanwhile, there may be other blemishes hidden from sight.

Propensity for Survey Participation

There are entire books written about responders vs. non-responders (see, for example, Groves and Couper (1998)), and we cannot cover all those issues. We will treat an obvious one that is at least feasible to deal with.

We start off by recognizing that we are conducting a survey. It is automatic to hypothesize that there may be a class of people who generally like to participate in surveys. When we get our sample of responders, such people will be over-represented. To the extent that they differ in the survey variables, there is a potential for survey bias.

But how do we identify these people? Is there a set of reliable universe estimates? The answer is no – there is no commonly accepted definition for this; and even if there is one, there would not be a set of universe estimates.

However, Cable, Jennings and Appel (1999) had come up with an operationalization for the propensity to participate in surveys. Their approach is to use a database overlay for a pre-designated sample. Among the database elements were some variables that were related to survey participation.

For example, one data element was a check list of interest and activities taken from warranty card returns. The response rates for those who were marked "yes" to any activity were between 15%-19% higher than for those who were not marked in the three studies conducted by Cable Jennings and Appel (1999).

As another example, the data element was about whether the person had previously responded to a direct mail solicitation. The response rates for those who were marked "yes" were between 12%-18% higher than for those who were not marked in the same three studies.

Survey Topic Saliency

Another factor that might impact survey participation is topic saliency. This was personally brought home to the first author when one day his aunt showed him a television diary and said, "I don't watch much television, so I don't think that they'll be interested in me." In truth, of course, 'they' are very interested.

In Groves, Presser and Dipko (2004) (see also Groves, Singer and Corning (2000)), there was a series of special topic surveys. Each survey was delivered to a general population control group and a target group matched to the topic. The results were that the response rate for teachers was 32% higher than the control group in a survey about education and schools; new parents responded higher by 15% in a survey about children and parents; persons 65 or older responded higher by 10% in a survey about medical care and health; and political contributors responded higher by 28% in a survey about voting and elections.

Study Design

Consider a particular survey on a special topic. For example, this could be a survey about print readership, or pharmaceutical product usage, or television viewing, or radio listening, or automobile ownership, or information technology, or personal finances, or social attitudes and opinions. There are well-established surveys in the United States for all of the above topics.

Based upon the preceding two sections, we can easily imagine that there may be an interaction between those two factors. We hypothesize that people who regularly participate in surveys <u>and</u> who are interested in the survey topic are more likely to participate in this survey. We will use an actual database and test out this hypothesis.

Our database is the 2005 MARS OTC/DTC Pharmaceutical Study. This is a mail survey of adults in the United States. We are interested in the national sample portion of the study. From Acxiom Corporation, we ordered 24,500 names/addresses drawn as a random sample from their master file of about 130 million records.

We did two mailings, the first with a US$5 incentive. The survey instrument is a twenty-page, 4-color questionnaire. This questionnaire has six pages on print readership for 100 magazines and 4 national newspapers, plus much more on healthcare and pharmaceutical product usage including insurance, ailments, treatments, brands, attitudes and opinions.

The final intab was 8,608 respondents. After accounting for non-deliverables, deceased, refusals, media affiliations and data usability, the response rate was 38.9%. At this point, we weighted this sample by geo-demographic variables: geography, age, sex, education, race, marital status, household composition, employment, occupation, personal income and household income.

Our hypothesis concerns regular survey participants who are interested in the subjects of healthcare and pharmaceutical products. Operationally, we use an approach similar to that used by Cable, Jennings and Appel (1999). We took the 24,500 pre-designated sample and we sent it out to two other database suppliers, Equifax and ICOM. Each of these database suppliers have compiled large list of names of people who have certain ailments (allergy, arthritis, asthma, cancer, diabetes, heartburn, high cholesterol, high blood pressure/hypertension, migraine headache, osteoporosis and overactive bladder). That information had come from sources such as warranty cards, web surveys and large-scale mail surveys (such as the famous 'Carol Wright' program). We asked the database suppliers to flag the names on our pre-designated sample with these ailments and send the information back to us.

Assumption Checking

We had worked under the assumption that the Equifax and ICOM names were compiled through survey-like processes, and therefore these people would have a greater propensity to participate in a survey, especially one whose subject is likely to be of interest to them.

Previously, we had reported that the overall response rate was 38.9%. The Equifax sub-sample had a response rate of 63.7% for an index of 164. The ICOM sub-sample had a response rate of 72.7% for an index of 187. So we have established that these sub-samples responded at significantly higher rates to this survey.

Our database supplier had positioned to us that their names are for people who have one or more of the listed ailments. We verify this against the survey responses on ailment incidences in our intab sample:

Table 1. Ailment Incidences in Intab Sample by List Source (percentage/index)

Ailment (past 12 months)	Total	Equifax-list	ICOM-list
Allergy (year round)	18% (100)	23% (127)	23% (125)
Allergy (seasonal)	19% (100)	21% (112)	22% (115)
Arthritis (osteo-arthritis)	17% (100)	19% (114)	30% (179)
Asthma	7% (100)	10% (141)	11% (155)
Cancer	9% (100)	9% (89)	15% (160)
Diabetes	9% (100)	12% (127)	14% (152)
Heartburn/indigestion	17% (100)	19% (108)	23% (129)
High cholesterol	19% (100)	23% (124)	31% (168)
Hypertension	23% (100)	26% (114)	38% (163)
Migraine headache	9% (100)	12% (137)	11% (125)
Osteoporosis	5% (100)	5% (115)	7% (156)
Overactive bladder	5% (100)	6% (108)	11% (219)

Empirical Results: Response Rates

Here is another way of presenting the response rate information.

> Within the pre-designated sample, the incidence of Equifax names was 6.4% while that of ICOM names was 3.3%.
>
> Within the intab sample, the incidence of Equifax was 9.9% while that of ICOM names was 6.7%. So those two lists were over-represented in the intab sample.
>
> With geo-demographic only weighting, the weighted incidence of Equifax was 9.9% while that of ICOM was 5.9%.
>
> With geo-demographic plus list weighting, by definition, the incidence of Equifax was 6.4% while that of ICOM was 3.3%.

In the following, we will compare these three weighting scenarios: no weighting; geo-demographic weighting; and geo-demographic plus list weighting. Given the nature and purpose of the MARS study, we will examine the impact on magazine ratings, ailment incidences and target group ratings (i.e. magazine ratings within ailment sufferers).

Empirical Results: Magazine Ratings

Within the MARS study, one hundred magazines were measured. Readership is determined from a frequency of reading question (number read out of last four issues) after a six-month screen.

In Table 2, we show the summary statistics for the three weighing methods: unweighted; geo-demographic weighting; and geo-demographic plus list weighting. We regard the geo-demographic plus list weighting as the most complete weighting system and we will call this full weighting model. The deviation between a weighted estimate by the full weighting method and the weighted estimate by another method shall be called ‘bias.’ This is not a true bias, because that would require us to know what the true estimate ought to be.

Table 2. Summary Statistics of Unweighted, Geo-demographic Weighted and Full Weighted Estimates of Magazine Audiences Ratings

Statistic	Unweighted vs. Full Method	Geo-demographic vs. Full Method
Index		
Mean	96.8	100.6
Standard Deviation	17.8	1.61
Maximum	140	107
Minimum	57	97
Absolute Deviation		
Mean	0.67	0.05
Standard Deviation	0.63	0.05

To illustrate what goes into Table 2, here is an example. We have an arthritis-themed magazine with the following ratings: unweighted at 1.43, geo-demographic weighted at 1.24 and fully weighted at 1.22.

This means that the unweighted-vs-full index is 100 x 1.43 / 1.22 = 117, and the geo-demographic-vs-full index is 100 x 1.24 / 1.22 = 102. An index of 100 would mean that the two ratings are the same. An index that is much greater or smaller than 100 would mean big differences. In this example, the unweighted estimate is significantly higher than the full estimate, whereas the geo-demographic weighted estimate is close to the full estimate.

The absolute difference between unweighted and full estimates is 1.43-1.22 = 0.21, and that between the geo-demographic-weighted and full estimates if 1.24 – 1.22 = 0.02. The same story obtained.

Table 2 contains the summary statistics across all 100 magazines. The arithmetic means are close enough, but the dispersion measures tell a different story. The unweighted estimates look to be very different from the weighted ones, but the geo-demographic weighted estimates are close to the fully weighted ones.

Empirical Results: Ailments

Within the MARS study, the respondents are asked if they have been either professional diagnosed or self-diagnosed with any of 54 different ailment conditions within the past 12 months. In Table 3, we show the incidences under the three weighting methods.

Table 3. Summary Statistics of Unweighted, Geo-demographic Weighted and Full Weighted Estimates of Ailment Incidences

Statistic	Unweighted vs. Full Method	Geo-demographic vs. Full Method
Index		
Mean	99.9	101.0
Standard Deviation	12.3	1.58
Maximum	130	104
Minimum	64	93
Absolute Deviation		
Mean	0.73	0.11
Standard Deviation	0.74	0.10

To illustrate what goes into Table 3, here is an example. Osteoarthritis has the following incidences: unweighted at 16.9, geo-demographic weighted at 14.7 and fully weighted at 14.3.

This means that the unweighted-vs-full index is 100 x 16.9 / 1.43 = 117, and the geo-demographic-vs-full index is 100 x 14.7 / 14.3 = 102. In this example, the unweighted estimate is significantly higher than the full estimate, whereas the geo-demographic weighted estimate is close to the full estimate.

The absolute difference between unweighted and full estimates is 16.9 – 14.3 = 2.6, and that between the geo-demographic-weighted and full estimates if 14.7 – 14.3 = 0.4. The same story obtained.

Table 3 contains the summary statistics across all 54 ailments. The unweighted estimates look to be very different from the weighted ones, but the geo-demographic weighted estimates are close to the fully weighted ones.

Empirical Results: Target Group Ratings

Target group ratings are magazine ratings within groups of people with ailment conditions. In theory, there are (54 target groups) x (100 magazine ratings) = 5,400 target group ratings here. In practice, a small number had to be excluded from the analysis because they were zeros.

Table 4. Summary Statistics of Unweighted, Geo-demographic Weighted and Full Weighted Estimates of Target Group Ratings

Statistic	Unweighted vs. Full Method	Geo-demographic vs. Full Method
Index		
Mean	104.1	101.5
Standard Deviation	40.2	9.0
Maximum	700	259
Minimum	9	85
Absolute Deviation		
Mean	1.00	0.18
Standard Deviation	1.15	0.29

Table 3 contains the summary statistics across the target group ratings. The unweighted estimates look to be very different from the weighted ones, but the geo-demographic weighted estimates are close to the fully weighted ones.

Discussion

For the magazine ratings, ailment incidences and target group ratings, we encountered the same phenomenon. The unweighted estimates are significantly different from the weighted ones. Weighting by geo-demographic variables appeared to be very similar to full weighting.

But we should not interpret this to mean that geo-demographic weighting has fully corrected the bias as a result of the explanatory power of the geo-demographic variables for survey participation and topic saliency. From the section on the empirical results for response rates, we saw that the unweighted incidences for Equifax and ICOM were 9.9% and 6.7%, the geo-demographic weighting brought them to 9.9% and 5.7% whereas the full model weighted them to the correct 6.4% and 3.3% respectively.

The reason that we are not seeing a huge difference between the geo-demographic weighting and full weighing is the relative smallness of these two lists with respect to the universe (6.4% and 3.3%). Bearing in mind that the database compilers work opportunistically by using whatever source that they can get, these two lists do not cover all possible persons who are "frequent survey participants with ailment conditions." By our estimate, the coverage of these two lists combined is between 20%-40%, depending on the ailment. This is why the impact does not show sharply. If we had a fuller listing, then we would see a large impact.

References

Cable, V., Jennings, D. and Appel, V. (1999) Using database overlays to correct survey non-response bias. World Readership Symposium 1999, 189-194.

Groves, R.M. and Couper, M.P. (1998) Nonresponse in household interview surveys. John Wiley & Sons.

Groves, R.M., Presser, S. and Dipko, S. (2004) The role of topic interest in survey participation decisions. Public Opinion Quarterly, 68, 2-31.

Groves, R.M., Singer E. and Corning, A. (2000) Leverage-salience theory of survey participation. Public Opinion Quarterly, 64, 299-308.

Kish, L. (1992) Weighting for unequal p_i. Journal of Official Statistics, 8, 183-200.

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