Polimetrix

Polimetrix, Inc.l
364 University Avenue
Palo Alto, CA 94301

Phone (650) 462-8000
Fax (650) 462-8422
www.polimetrix.com

Sample Matching

Representative Sampling from Internet Panels

Douglas Rivers, Ph.D.
Founder, President and CEO, Polimetrix, Inc.
Professor of Political Science, Stanford University

Section
1

Introduction

Sample matching is a new methodology for selection of study samples from pools of opt-in respondents. The methodology addresses the primary substantive and technical issues of how large, but unrepresentative, panels can be used to construct representative study samples for particular target populations. The procedure uses a listing or enumeration of the population that can be obtained from large scale consumer and voter databases that have been developed in recent years. The existence of such data has not been exploited in previous Internet research. We show how the procedure performed in predicting the outcome of the 2005 California special election. On both a theoretical and a practical level, this approach substantially improves upon existing weighting procedures, which are also reviewed.

The Web Sampling Problem

Most samples today, whether for phone or the Internet, do not even roughly approximate random samples. In the case of phone surveys, where random digit dialing (RDD) or random selection from a list is used to select respondents, typical response rates for media polls or market research surveys are in the range of 20 percent. As a result, sample selection is primarily determined by whom *chooses* to respond, not the random selection mechanism.

In the case of Web surveys, most Internet panels do not even pretend to be randomly selected. Panel members are recruited by a variety of means (banner ads, email lists, promotions, and offers) and those who “opt-in” become the pool of respondents available for sample selection.

There are a few Internet panels, such as NetRatings and Knowledge Networks, that do use random selection. NetRatings uses RDD to recruit a panel of Internet users who allow their Web traffic to be monitored. Knowledge Networks uses RDD to recruit a panel of both existing Internet users and non-users. Those without home Internet access are provided with an inexpensive device that allows them to be interviewed on the Internet. However, both NetRatings and Knowledge Networks have struggled with low response rates, high costs, and limitations imposed by small panel size.

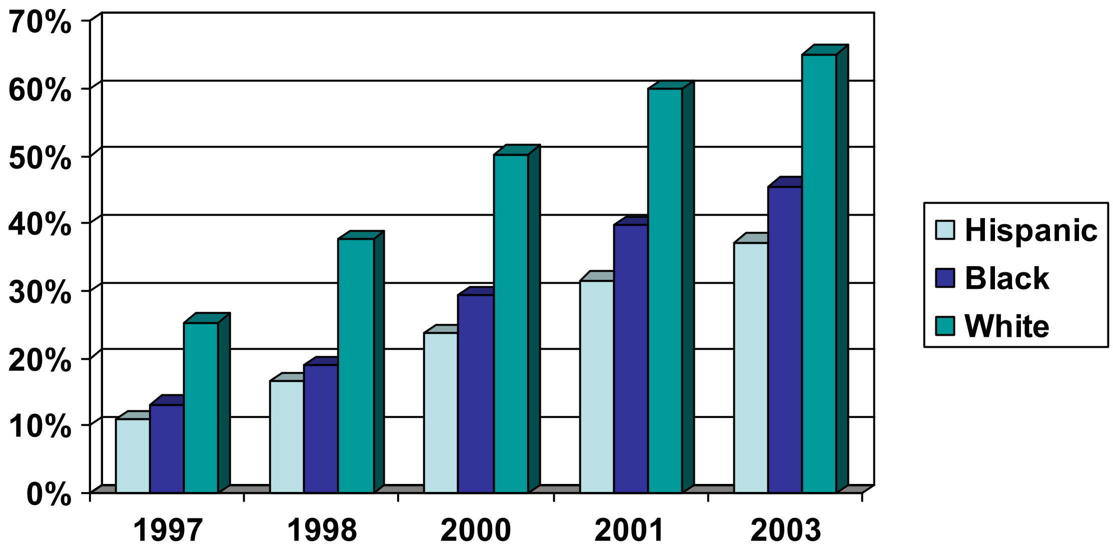
Sample quality is largely a function of two factors: *population coverage* and *selection bias.* Population coverage refers to the proportion of the target population that is reachable, while selection bias refers to the willingness of reachable respondents to complete an interview. It would be nonsensical, for example, to use an opt-in Internet panel for a study of non-internet users, since the panel lacks coverage of that population. On the other hand, even if a population can be reached by, say, RDD, sample quality will still be poor if patterns of respondent cooperation cause selection bias.

Population Coverage

In the early days of Internet surveys, the primary sampling problem was the "Digital Divide." Internet usage was concentrated in more affluent and better educated segments of the population, while racial minorities, the elderly, and women were substantially underrepresented among Internet users. Today, nearly three quarters of the adult population has access to the Internet, either at home, work, or school, so that most of the population is, at least in principle, reachable by the Internet. Usage rates are lower for African Americans, Latinos, persons with a high school education or less, and the elderly, but none of these groups is excluded altogether.

Figure 1 provides data on Internet access by race, as measured by the Current Population Survey. Internet usage has grown at about the same rate in all racial groups. The effect of this growth, however, has been to substantially reduce (though not eliminate) the degree to which minority groups are underrepresented among Internet users. In 1997, for example, whites were more than twice as likely to have Internet access as blacks and Hispanics. By 2003, whites were only about a third more likely to have Internet access as Blacks. Similar patterns can be found in other groups.

Figure 1: Race and Internet Access



The Digital Divide has diminished substantially and will largely disappear in the next decade, as the Internet becomes the vehicle for the delivery of home entertainment and phone services. Even today, Internet coverage is adequate for most types of research. The problem is not coverage—who can be reached on the internet—but sample selection.

Selection Bias

Most Internet surveys are not conducted using a random sample of Internet users. Instead, "access panels" have been developed from which samples are selected for individual studies. The properties of these panels vary depending upon how they were recruited. In this section, we compare selection biases in Internet surveys with selection biases in phone surveys.

Different types of people have different propensities for participation in survey research. These propensities lead to underrepresentation of certain groups in both Internet panels and RDD phone samples. In fact, the degree of underrepresentation of these groups (except for the elderly, discussed in more detail below) is not much different in an opt-in Internet panel, than in an unweighted RDD phone sample. Tables 1 shows the proportion of several difficult to reach groups in national media polls conducted by one of the national television networks during 2004. Table 2 shows the proportion of an opt-in panel belonging to each of a similar set of groups.

Table 1: Unweighted Sample Composition of National Media Poll

	Census	Average of 11 Surveys	Implied Weight
Blacks	11.0%	7.9%	1.4
Hispanics	12.4%	4.8%	2.6
Aged 18-24	12.3%	6.4%	1.9
HS or less	46.6%	32.7%	1.4
Postgraduate	8.7%	17.2%	0.5
Never married	23.8%	16.2%	1.5

Table 2: Composition of Opt-in Web Panel

	Web Panel	Internet Users	Web Panel
Blacks	4.3%	9.3%	11.0%
Hispanic	3.3%	7.2%	12.4%
Postgraduate degree	23.3%	14.7%	8.7%
Aged 18-24	8.7%	16.0%	12.3%
Male	58.8%	48.7%	48.9%
Married	60.4%	55.3%	54.3%

The conclusion to be drawn from these data is not that opt-in Web panels are representative of any particular population. This is demonstrably false—people who opt-in for taking Web surveys have different demographics than either the population of all Internet users or the population of all adults. But the same is true for RDD telephone samples. In both cases, an appropriate methodology is required to produce usable samples for individual studies. We will discuss various solutions to this problem in Sections 2 and 3 below.

The Elderly on the Internet

The Internet is often viewed as a venue for the young. Among the elderly, there tend to be fewer Internet users and a larger proportion who express no interest in having Internet access. While both statements are true, a lesser known fact is that elderly Internet users are much more likely to participate in Web surveys so most Internet research panels have an excess of elderly participants, not a shortage.

Of course, the relevant question is not whether a panel has too many or too few elderly, but whether its elderly participants are representative or atypical of the elderly population.

The evidence suggests that elderly Web survey participants are somewhat different—more affluent and knowledgeable about technology—but, after controlling for these factors, similar to elderly phone respondents.

The problem of sampling the elderly using an opt-in Internet panel provides a good illustration of the issues that a valid sample selection procedure must deal with. There are usually some characteristics associated with sample selection that need to be identified to correct sample biases. In many years of experience with phone surveys, these factors have, for the most part, been identified and reasonably satisfactory measures developed for handling them.

Problems with Phone Samples

The quality of phone samples, however, has been deteriorating for a variety of reasons. First, cell phones have replaced landlines, especially among younger age groups. (Over 25 percent of those between the ages of 18 and 29 are not reachable on landlines.) Because of regulations on outbound calls to cell phones, this population is no longer reachable in a RDD phone sample. Phone coverage, which as recently as five years ago was in excess of 96 percent of the adult population, now appears to be under 90 percent and will continue to fall.

Caller ID and answering machines now make it harder to contact respondents. In a short field period, it is practically impossible to contact more than half of the working numbers in a RDD sample. This pushes overall response rates to well under 50 percent.

Finally, declining cooperation for all types of surveys (including in-person interviews) has reduced the completion rate among contacted respondents. The overall response rates are so low that few survey organizations publish them for phone studies. To some degree, willingness to consider using opt-in Internet samples just reflects a realization that most phone samples are mostly opt-in samples too.

Section
2

Current Practice for Selection and Weighting

Quota Sampling

By far the most common method for sample selection in consumer market research is quota sampling. In quota sampling, one defines a set of groups (*e.g.,* men, women, 18-29 year olds, 30-64 year olds, 65+, *etc.*) and specifies how many respondents should be recruited for each group. Recruitment is then done on an *ad hoc* basis and any respondents in excess of the specified quota are turned away.

Needless to say, quota sampling has no basis in sampling theory, since the surveyor has almost complete discretion in selection of respondents with the "cells." In practice, the hard-to-fill quotas are the last to be filled and often end up being highly unrepresentative.

For example, many phone surveys use explicit or implicit quotas for gender, since men are more difficult to reach by phone than women. Different devices—always ask for a man first and, if none is available, then accept a woman—are employed to "balance" phone samples. The resulting samples are often very unrepresentative of men, since the available men are less likely to be employed and often older. Some media organizations have tried to address this problem by asking first for the *youngest* male at home and, if unavailable, then to ask for the oldest female. These procedures do not produce accurate age distributions within gender groups.

Quota sampling is a relic of the 1930's and should not be employed in the twenty-first century. It is, unfortunately, the standard sampling procedure for most modern Web surveys.

Raking

For samples that have already been selected, the most popular method of weighting is the method of *raking,* also known as *rim-weighting,* first proposed by W. E. Deming during the 1940's. In raking, the sample marginals are forced to match the known population marginals (from a census or other source) by an iterative procedure. The primary advantage of raking is that it does not require the joint distribution of the variables to be known. It has a number of serious disadvantages. First, if the population marginals are skewed the iterative weighting procedure often does not converge. Second, it generally does not find the correct weighting for combinations of variables. It can be shown that the implied joint distribution maximizes the entropy over a certain class of distributions. Since the weighting variables are often expected to be highly intercorrelated (*e.g.,* race, education, and income), this is undesirable behavior. Third, and perhaps most important, raking yields unstable and unreliable estimates when the number of variables used to weight the sample is large. Which variables are used for weighting can often have serious implications for survey estimates. The reliability of these estimates then becomes a subjective judgment about which variables to use in weighting.

Cell Weighting

An alternative to raking is cell weighting, where the population is divided into a set of mutually exclusive and exhaustive categories (or "cells"). The sample is then weighted by the ratio of the population fraction in each cell to the corresponding sample fraction. This is sometimes called post-stratification. It differs from the usual type of stratification in that the sample observations in each cell are not a sample from the corresponding subpopulation, because of non-response. The procedure is valid if an ignorability assumption, similar to that described below, holds—the survey measurements need to be conditionally independent of non-response given the variables used for post-stratification.

There are two primary deficiencies of cell weighting. First, if the weights are large, the estimates can be highly inefficient and unstable. It is common practice to trim the weights (so, for example, weights are constrained to lie between, say, ½ and 2), but with current phone and Internet samples, larger weights are often needed to deal with differential non-response. Second, usually the cross-classification of only a few variables is available, so cell weighting is only applicable with a small number of variables and categories. This means that the range of non-response problems that can be remedied with cell weighting is limited.

Section
3

Sample Selection by Matching

Description of Sample Matching Methodology

Sample matching is a newly developed methodology for selection of "representative" samples from non-randomly selected pools of respondents. It is ideally suited for Web access panels, but could also be used for other types of surveys, such as phone surveys.

Sample matching starts with an enumeration of the *target population.* In other contexts, this is known as the *sampling frame*, though, unlike conventional sampling, the sample is *not* drawn from the frame. For a study of registered voters, the target population is the set of registered voters, who are enumerated (with some exceptions) in the registered voter list. For general population studies, the target population is all adults, as enumerated (again with some exceptions) in consumer databases maintained by commercial vendors such as Acxiom, Experian, and InfoUSA. The development of comprehensive consumer and voter databases is a relatively recent phenomenon that has important implications for survey sampling.

Sample selection using the matching methodology is a two-stage process. First, a random sample is drawn from the target population. We call this sample the *target sample.* Details on how the target sample is drawn are provided below, but the essential idea is that this sample is a true probability sample and thus representative of the frame from which it was drawn.

Ideally, we would interview the respondents in the target sample and conventional sampling theory would describe the properties of the sample. However, we have no economical way of contacting most members of the target sample: they have not provided their email addresses to us, many do not have listed phone numbers, and those who do have listed numbers would not agree to be interviewed. *We do not attempt to interview members of the target sample.*

Instead, for each member of the target sample, we select one or more *matching* members from our pool of opt-in respondents. This is called the *matched sample.* Matching is accomplished using a large set of variables that are available in consumer and voter databases for both the target population and the opt-in panel. Details of matching are provided below.

The purpose of matching is to find an available respondent who is as similar as possible to the selected member of the target sample. The result is a sample of respondents who have the same measured characteristics as the target sample. Under certain conditions, described below, the matched sample will have similar properties to a true random sample. That is, the matched sample mimics the characteristics of the target sample. It is, as far as we can tell, "representative" of the target population (because it is similar to the target sample).

Selection of the Target Sample

In explaining the sample matching methodology, it may be helpful to think of the target sample as a simple random sample (SRS) from the target population. However, the efficiency of the procedure can be improved by using stratified sampling in place of simple random sampling. SRS is generally less efficient than stratified sampling because the size of population subgroups varies in the target sample.

With stratified sampling, we partition the population into a set of categories that are believed to be more homogeneous than the overall population. These categories are called *strata.* For example, we might divide the population into race, age, and gender categories. The cross-classification of these three attributes divides the overall population into a set of mutually exclusive and exhaustive groups or strata. Then a SRS is drawn from each category and the combined set of respondents constitutes a stratified sample. If the number of respondents selected in each strata is proportional to their frequency in the target population, then the sample is self-representing and requires no additional weighting.

At Polimetrix, we usually stratify on race, gender, and age. If it is a political study, we also stratify on party registration and region. For other types of studies, custom strata can be developed.

The Distance Function

When choosing the matched sample, it is necessary to find the closest matching respondent in the panel of opt-ins to each member of the target sample. Various types of matching could be employed: exact matching, propensity score matching, and proximity matching. Exact matching is impossible if the set of characteristics used for matching is large and, even for a small set of characteristics, requires a very large panel (to find an exact match). Propensity score matching has the disadvantage of requiring estimation of the propensity score. Either a propensity score needs to be estimated for each individual study, so the procedure is automatic, or a single propensity score must be estimated for all studies. If large numbers of variables are used the estimated propensity scores can become unstable and lead to poor samples.

At Polimetrix, we employ an proximity matching method. For each variable used for matching, we define a *distance function*, d(x,y), which describes how “close” the values x and y are on a particular attribute. For numerical characteristics, such as age, years of schooling, latitude, longitude, income, etc., the distance function is usually just the absolute value of the difference |x – y|, though, occasionally, we use the square of the distance to penalize large discrepancies.

The overall distance between a member of the target sample and a member of the panel is a weighted sum of the individual distance functions on each attribute. The weights can be adjusted for each study based upon which variables are thought to be important for that study, though, for the most part, we have not found the matching procedure to be sensitive to small adjustments of the weights. A large weight, on the other hand, forces the algorithm toward an exact match on that dimension.

Nonresponse Adjustments

Not all respondents in a matched sample will respond to a survey invitations. At Polimetrix, we use two procedures to deal with nonresponse: *multiple matching* and *rematching.*

Instead of selecting a single match for each member of the target sample, we select multiple matches. The number of matches is based on an estimated response probability (using a hazard model to estimate the probability that a panelist responds by the end of the survey field period). The total number of panelists matched to each member of the target sample is determined by matching panelists until the expected number of responses is greater than or equal to one.

Second, we use a second round of matching when respondents begin an interview. Though the expected number of respondents who arrive for each target sample element is approximately one, randomness in response patterns will mean that some target sample elements are matched more than once and some none at all. The best matching respondent is assigned to the matching target element if that element has not already been matched. Otherwise, the responding panelist is compared to the target sample elements *across all open studies* and assigned to the closest matching respondent using a priority assignment algorithm. This minimizes the number of respondents who are turned away (because a match has already been found) and ensures the most accurate matches possible.

Statistical Theory

The intuition behind sample matching is clear: if respondents who are similar on a large number of characteristics tend to be similar on other items for which we lack data, then substituting one for the other should have little impact upon the sample. Can this intutition be made rigorous? The answer is "yes," as we describe below.

The theoretical conditions the guarantee the validity of sample matching are quite technical, but their content is easily understood. There are three main assumptions:

Assumption 1: Ignorability

Panel participation is assumed to be *ignorable* with respect to the variables measured by survey conditional upon the variables used for matching. What this means is that if we examined panel participants and non-participants who have exactly the same values of the matching variables, then on average there would be no difference between how these sets of respondents answered the survey. This does *not* imply that panel participants and non-participants are identical, but only that the differences are captured by the variables used for matching. Since the set of data used for matching is quite extensive, this is, in most cases, a plausible assumption.

Assumption 2: Smoothness

The expected value of the survey items given the variables used for matching is a "smooth" function. Smoothness is a technical term meaning that the function is continuously differentiable with bounded first derivative. In practice, this means that that the expected value function doesn't have any kinks or jumps.

Assumption 3: Common Support

The variables used for matching need to have a distribution that covers the same range of values for panelists and non-panelists. More precisely, the probability distribution of the matching variables must be bounded away from zero for panelists on the range of values (known as the "support") taken by the non-panelists. In practice, this excludes attempts to match on variables for which there are no possible matches within the panel. For instance, it would be impossible to match on computer usage because there are no panelists without some experience using computers.

Under Assumptions 1-3, it can be shown that if the panel is sufficiently large, then the matched sample provides consistent estimates for survey measurements. The sampling variances will depend upon how close the matches are if the number of variables used for matching is large, but Monte Carlo evidence indicates that these adjustments are usually small. The key issues for an application are whether the variables used for matching are adequate controls for panel participation effects and, if they are, whether the panel is large enough to permit close matches.

Section
4

Validation of Sample Matching

2005 California Special Election

During the 2005 California special election, Polimetrix released survey estimates of the proportion of voters intending to vote for and against seven propositions on the ballot. These estimates were contained in press releases that were published on several Web sites (www.realclearpolitics.com, www.pollingreport.com, and the National Journal's *Hotline*). The outcome of all seven propositions was correctly predicted (a record matched by only one other polling organization) and the root mean square error was 3.0% (only slightly larger than what would be expected from random sampling). The results are shown in Table 3 below.

Table 3: Survey Accuracy in 2005 California Special Election

	Polimetrix Final Survey			**Election Outcome**	
Proposition	Yes	No	Undecided	Outcome	Error
73	43%	54%	2%	47.4%	-3.1%
74	45%	52%	3%	45.1%	1.3%
75	48%	49%	3%	46.7%	2.8%
76	40%	56%	3%	38.0%	3.7%
77	41%	52%	6%	40.6%	3.5%
78	33%	55%	13%	41.5%	-4.0%
79	38%	46%	16%	39.0%	6.2%

While one (or even seven) estimates do not prove that the methodology “works,” these results are very encouraging. In an election which a number of phone and other Internet surveys provided very misleading estimates, sample matching performed very well.