Non-Response Bias

Nathan Berg

University of Texas at Dallas

Outline

I. Motivation for Analyzing Non-Response Bias

II. Classifying Types of Error and Bias

A. Sampling Error

B. Non-Representative Samples

C. Dealing with Non-Representativeness Before or After Data are Collected: the Sample Design Stage and the Data-Analysis Stage

D. Appreciating the Similarity among Different "Biases" with Different Labels

E. Mis-reporting Versus Non-response

III. Analysis of Survey Data with Missing Responses

A. Item versus Unit Non-Response

B. Little and Rubin's Missing Data Framework

1. Imputation

2. Weighting

3. The Maximum-Likelihood Approach

4. Missing-at-Random, Missing-Completely-at-Random, Mixture Modeling, and Multiple Imputation

C. Other Perspectives on Correcting for Non-Response Bias

IV. Measuring Non-Response Bias

A. Validation

B. Designing Surveys so that Non-Response Bias Can Be Estimated

1. Randomized Response

2. A Budget Constraint Means Trading Off Sampling Error for Bias Reduction

C. Parsing the Meaning of the "Don't Know" Response

D. Panel Data and Attrition

V. Summary

Glossary

Bias: the expected difference between an estimated characteristic of a population and that population's true characteristic.

Non-response: a survey response that falls outside the range of responses that the survey designers consider to be valid.

Item non-response: non-response to a particular survey item accompanied by at least one valid measurement for the same respondent, e.g., leaving just one item on a questionnaire blank, or responding to some questions by saying, "I don't know," while providing a valid response to other questions.

Unit non-response: complete non-participation on the part of someone who survey designers intended to include in the survey.

Unit: one observation, i.e., a single vector of measurements, usually corresponding to a particular individual at a given point in time, many of which comprise a sample.

Definition Statement

NON-RESPONSE BIAS refers to the mistake one expects to make in estimating a population characteristic based on a sample of survey data in which, due to non-response, certain types of survey respondents are under-represented.

Text

I. Motivation for Analyzing Non-Response Bias

To illustrate and underscore the importance of analyzing non-response bias, consider the following scenario. A researcher working for a marketing

firm desires to estimate the average age of New Yorkers who own a telephone. In order to do this, the researcher attempts to conduct a phone survey of 1000 individuals drawn from the population of phone-owning New Yorkers by dialing randomly chosen residential phone numbers. After 1000 attempts, however, the researcher is in possession of only 746 valid responses, because 254 individuals never answered the phone and therefore could not be reached. At this point, the researcher averages the ages of the 746 respondents for whom an age was obtained and ponders whether this average is likely to be too high or too low. Should one expect the 254 non-responders to be about the same age as those who answered their phones, or are they likely to be older, or younger? After thinking it over, the researcher concludes that the average age of the 746 responders is a biased estimate, because the surveys were conducted during business hours when workers were likely to be at work rather than at home, implying that the 746 valid responses probably contain a higher fraction of retirees than would be found among non-responders. In this case, the difference between the expected value of the estimated average age, which is too high, and the true, but unknown, average age is precisely non-response bias.

Social scientists often attempt to make inferences about a population by drawing a random sample and studying relationships among the measurements contained in the sample. When individuals from a special subset of

the population are systematically omitted from a particular sample, however, the sample cannot be said to be "random," in the sense that every member of the population is equally likely to be included in the sample. It is important to acknowledge that any patterns uncovered in analyzing a non-random sample do not provide valid grounds for generalizing about a population in the same way that patterns present in a random sample do. The mismatch between the average characteristics of respondents in a non-random sample and the average characteristics of the population can lead to serious problems in understanding the causes of social phenomena and may lead to misdirected policy action. Therefore, considerable attention has been given to the problem of non-response bias, both at the stages of data collection and data analysis.

II. Classifying Types of Error and Bias

A. Sampling Error

Anytime one generalizes about a population based on a sample, as opposed to conducting a complete census of the population, there is an unavoidable possibility of mistaken inference. As such, sampling error arises even under the best of circumstances simply because, due to chance, averages of variables in a random sample are not identical to the corresponding averages in the population. Fortunately, sampling error typically disappears as the sample size increases. More importantly, sampling error does not lead to

bias, since population characteristics can be estimated in such a way that the probability-weighted average of possible over-estimates and under-estimates is precisely zero.

B. Non-Representative Samples

To be distinguished from sampling error is an entire family of non-sampling errors that arise when a sample is selected from a population in such a way that some members of the population are less likely to be included than others. In such cases, the sample is said to be non-random, or non-representative, with respect to the population one intends to study. In contrast to sampling error, a non-representative sample generally leads to biased estimation.

A number of factors may cause a sample to be non-representative. One possibility is that, because of a flawed survey design, the survey simply fails to reach certain segments of the population. As in the example described in section I, a daytime phone survey tends to under-represent people who work, just as a survey of Kansans would tend to under-represent urban Americans, or a survey of car owners would tend to under-represent those who use public transportation.

Another possible cause of a non-representative sample is mistakes made by surveyors in coding survey responses. The key question is whether such mistakes are correlated with the type of individual being surveyed. For instance, a surveyor who, in the course of interviewing survey respondents,

sometimes gets carried away discussing sports and forgets to record the respondent's last few responses will end up with a sample in which sports fans are under-represented among the complete survey responses.

Perhaps the most common reason for non-representative samples, however, is the behavior of survey respondents themselves. Often times, the very fact of being a non-responder correlates with other characteristics of interest. When it does, non-response inevitably leads to non-random sampling and creates the potential for biased estimation of the characteristics under study. Researchers working with survey data must always consider the possibility that certain types of individuals are more likely to refuse to respond. This problem is acute when one of the key variables of interest determines, in part, who is more likely to select themselves out of a sample by not answering a survey question.

It is often suspected, for example, that individuals with high incomes are less likely to voluntarily disclose their income, biasing survey-based estimates of income downward. Similarly, those engaged in illicit drug activity, fearing the consequences of divulging that sensitive information, are probably less likely to participate in a survey about drug use, leading, again, to the potential for systematic underestimation. A slightly more subtle example is the case of estimating the percentage of a population that supports one of two political candidates. Apathetic voters are often thought to be the least

likely to cooperate with political pollsters, even though many of them will in fact vote. Basing election forecasts on a sample of only those who agree to answer the poll can be misleading, because the opinions of apathetic voters are under-represented in pollsters' samples.

C. Dealing with Non-Representativeness Before or After Data are Collected: the Sample Design Stage and the Data-Analysis Stage

In dealing with non-representative samples in general and non-response bias in particular, it is helpful to distinguish two broad stages in a social scientist's research project, namely, data collection and data analysis. Some researchers conduct surveys themselves and therefore have direct control over the details of data collection. Others work with data sets originally collected by someone else, in which case the researcher exerts no direct control over the data collection stage.

For those who have a say about how the data are to be collected, it is crucial to try foreseeing potential flaws in order to reduce the likelihood that differing incentives of different types of survey respondents will ultimately lead to bias. A vast literature exists on the topic of survey design, covering everything from the wording of survey questions to the issue of how many times those who do not answer the phone on a phone survey ought to be called back. Sometimes surveys can be designed in such a way − e.g., by obtaining some information from face-to-face interviews and the rest by phone − so

as to provide a means of estimating the non-response bias associated with a particular data collection technique.

Many researchers in the social sciences, rather than collecting new data themselves, study data that have been collected by others, e.g., the U.S. Census, the Current Population Survey, or the General Social Survey. At this secondary stage of data analysis, the researcher must decide what to do about survey respondents who failed to answer particular questions, the so-called "missing data problem." An additional issue is what to do about the target respondents who did not participate in the survey at all.

D. Appreciating the Similarity Among Different "Biases" with Different Labels

One finds many different labels for biases that are, in fact, instances of one common problem, i.e., trying to learn about a population based on a non-representative sample. It is helpful to see the underlying similarity among biases that arise from non-representative samples, because a successful approach to dealing with bias in one particular context often can be applied directly in new settings. In particular, survey data with missing responses can frequently be analyzed using techniques from the statistical and econometric literature under the heading, "measurement error." Terms such as "non-completion bias" or "volunteer bias," referring to the non-representative sample problem that arises when only special kinds of respondents actually

"complete" a survey questionnaire, or to situations where the subpopulation of "volunteers" is substantively different from the rest of the population, should be viewed as essentially the same as non-response bias.

The connection between non-response bias and selction bias warrants special mention. Non-response is clearly a special kind of selection problem of the type analyzed in the work of James Heckman. Thus, "selection bias," when referring to the mechanism by which some survey respondents choose not to answer survey questions (thereby selecting themselves out of the sample), overlaps with what was defined earlier as "non-response bias." Heckman, in turn, interpreted the selection problem more generally as a kind of econometric misspecification. For illustration, it is useful to consider a regression model, used frequently by labor economists, in which expected wage depends on a number of demographic variables as well as other factors thought to influence workplace productivity. If no account is taken of the mechanism by which only special kinds of individuals choose to become workers and therefore wind up included in the sample (implying that regressors are correlated with the error term in the regression model), then the econometric model is, in Heckman's words, "misspecified," leading to so-called "misspecification bias."

E. Mis-reporting Versus Non-response

When those collecting data ask respondents to report on their own be-

havior in connection with activities such as cheating, personal finance, sex, or alcohol and drug use, some respondents, instead of refusing to answer, will mis-report. When interpreted at face value, a sample in which certain kinds of individuals tend to misreport themselves does not accurately represent the population under study. As with non-representative samples caused by non-response, mis-reporting usually leads to bias, referred to with labels such as "mis-classification bias," "mis-reporting bias," "contaminated data bias," or simply "response bias." The task of the researcher is to consider how such mis-reporting will influence the estimates of key population characteristics. An important reference for anyone attempting to estimate mis-reporting bias in a discrete-response setting is the 1998 article of Hausman, Abrevaya, and Scott-Morton in *Journal of Econometrics.*

II. Analysis of Survey Data with Missing Responses

A. Item versus Unit Non-Response

An important distinction to make regarding non-response is "item" versus "unit" non-response, a distinction that turns on whether there is at least one survey item for which a valid response was obtained or whether the entire unit is missing. When entire units are missing from a sample, no test or correction for bias is available without obtaining additional data that include information about the targeted respondents who did not respond at all to the initial survey. In contrast, item non-response does not doom es-

timation to be biased, since techniques are available for using the partially completed responses returned from item non-responders to control for differences across responders and item non-responders. The following sections discuss techniques for computing unbiased estimates with samples that feature item non-response.

B. Little and Rubin's Missing Data Framework

Roderick Little and Donald Rubin, individually and in joint work, have written a number of frequently cited articles on the subject of analyzing data with missing values. Their approach is quite general and applies directly to most situations applied researchers working with survey data are likely to face.

1. Imputation

One possible approach to dealing with missing survey responses is to somehow "fill in" the missing values, "imputing" good guesses in place of missing survey entries. Some researchers, for instance, may replace missing measurements with the average value across the complete cases. A more sophisticated approach involves replacing missing values with estimates based on prediction equations that are fitted with the complete cases and subsequently used to predict missing values using the partial responses of item non-responders. After imputing values to fill in the missing data, data analysis proceeds using traditional estimation techniques.

A serious drawback to this technique is that the precision of estimates computed using the data set with imputed values will be overstated, for two reasons. First, imputed values generally are computed by averaging over other observations and, therefore, will be more tightly clustered about the mean than a fresh collection of *bona fide* observations would be. And second, the use of traditional statistical techniques after imputing values for missing entries in one's data matrix will be based on an overstated sample size, since a sample of N observations, some of which have been imputed, will contain less than N independent pieces of information. This means that the computed standard errors will be too small, and that the nominal size of significance tests will be inflated. Those interested in using imputation and weighting schemes taking such potential pitfalls into account should consult Rubin's *Multiple Imputation for Nonresponse in Surveys.*

2. Weighting

Another approach to working with incomplete data involves discarding partial observations and assigning a weight to each complete observation so that the weighted sample better represents the average characteristics of the population. For instance, if one were working with a sample of 68 men and 32 women in which women appear to be under-represented, one might consider placing additional weight on the female units in the sample, perhaps based on the gender ratio from the U.S. Census, in order to reduce bias.

In principle, weighting should work well to correct for bias that arises from estimation based on non-representative samples. A severe complication, however, is knowing how to compute standard errors that accurately account for the imprecision in the weights themselves. Doing so is notoriously difficult. Therefore, many authors, including Little and Rubin, recommend against using this technique. Those authors also point out that the most common approach to non-response is simply to discard incomplete responses, effectively giving each of the complete sample units the same weight. Except for the unusually lucky case where the complete-only sub-sample is a truly random sample of the population, this technique, although simple-to-use and widely practiced, leads to biased estimates.

3. The Maximum-Likelihood Approach

The maximum-likelihood approach is, far and away, the preferred approach to correcting for non-response bias, and the one advocated by Little and Rubin. The maximum-likelihood approach begins by writing down a probability distribution that defines the likelihood of observing the sample, as a function of population and distribution parameters θ. If x_1 and x_2 represent responses to two different survey questions by a single individual, the likelihood associated with a complete response may be expressed as $f(x_1, x_2; \theta)$, where f is the joint probability density function of x_1 and x_2. For individuals who only report x_1, the likelihood associated with x_1 is $\int_{-\infty}^{\infty} f(x_1, x_2; \theta) dx_2$,

which can, under the assumption of joint normality, be simplified to a more convenient form. In this way, a likelihood function is specified that includes terms corresponding to each observation, whether completely or only partially observed. The likelihood objective is then maximized with respect to θ, which produces estimates of the desired characteristics, enjoying all the well-known properties of maximum-likelihood estimation.

Most important among those properties, the maximum-likelihood estimate of θ converges to the true θ under the assumption that the probability distribution is correctly specified. Maximum likelihood estimates are also asymptotically normal and asymptotically efficient, meaning that the maximum-likelihood estimate of θ is approximately normal and is the best use of the information contained in the sample, given a sufficiently large number of observations. In addition to these advantages, the maximum-likelihood approach makes it possible to estimate fairly elaborate multi-equation models in which the probability that an individual fails to respond depends on other observable variables. Within such a framework, it is often possible to construct a quantitative test of the "missing at random" hypothesis, implemented as a straightforward significance test of an appropriate parameter restriction. The main drawback to maximum likelihood estimation is that the researcher must make strong assumptions about the probability distribution generating the random survey responses. Still, the advantages of this approach usually

are thought to outweigh the drawbacks, making it the approach of choice for many quantitative researchers.

4. Missing-at-Random, Missing-Completely-at-Random, Mixture Modeling, and Multiple Imputation

A frequently-mentioned distinction in the missing-data literature involves the two terms, "missing-at-random" and "missing-completely-at-random." If the probability of non-response for a variable Y is the same for every unit of observation in the population, then Y is said to be missing-completely-at-random. If, on the other hand, the probability of non-response systematically relates to other variables in the model, but not to the value of Y itself, then Y is said to be missing-at-random. Defining the random variable $R = 0$ if Y is missing, and $R = 1$ otherwise, another important distinction can be expressed: so-called *selection* models require the user to observe the conditional distribution $Y|R = 1$ and hypothesize the probability $R = 1|Y = y$, whereas *mixture* models require observing $Y|R = 1$ and hypothesizing $Y|R = 0$. The technique of multiple imputation, which has been used to advise policy-making entities such as the U.S. Department of Commerce in analyzing survey data, can be understood as a mixture model in which a range of distributions $Y|R = 0$ is hypothesized. There are a number of connections, including some surprising technical results, that relate selection and mixture models to one another.

D. Other Perspectives on Correcting for Non-Response Bias

Lawrence Marsh and his co-authors have proposed a number of interesting models of non-response, and developed the associated maximum-likelihood estimators, which appear to work well in practice. Marsh's work, in addition to providing straightforward maximum-likelihood estimators of non-response bias, compares the performance of maximum-likelihood-based corrections for non-response bias against those associated with alternative techniques of estimation, such as maximum entropy, finding consistent support for the maximum-likelihood approach. These results rest on the existence of auxiliary relations that determine the missing response mechanism. In the absence of auxiliary relations, Lien and Rearden's 1988 article in *Economics Letters* shows that, when the missing observation is the dependent variable in a limited dependent variable model, nothing is gained by applying maximum likelihood-based corrections. Thus, special caution is warranted when estimating a model in which the dependent variable is frequently missing.

III. Measuring Non-Response Bias

A. Validation

Validation is a general approach to testing for non-response bias that almost always involves comparing two different samples drawn from the same population. The technique of validation permits one to measure non-response bias, to test the hypothesis of no bias, and to identify which variables, if any,

are correlated with non-response. This approach is only feasible, however, if one is lucky enough to have two samples drawn from the same population.

Given a pair of samples, it is usually clear, either from the number of missing entries or from descriptive notes attached to the data, which data set has a lower non-response rate. The general philosophy of validation assumes that the sample with the lower non-response rate is, for all practical purposes, the "reliable" one. Accepting this view, significant departures among the observations in the unreliable sample relative to the average characteristics in the reliable sample can then be attributed to non-response bias, providing a qualitative measure (too high versus too low) along with a quantitative measure of the severity of the problem.

For instance, it is well accepted that face-to-face interviews typically draw a higher response rate than phone surveys do. Now suppose one draws two samples of measurements on ethnicity, one face-to-face and the other by phone, and discovers that the fraction of Asian-Americans in the phone data is half that of the face-to-face interview data. Taking the estimated racial composition of those who respond to the face-to-face interview as the reliable benchmark, one might plausibly infer that Asian-Americans are twice as likely to non-respond in a phone survey compared to other types of Americans. The qualitative finding that phone survey data may under-represent Asian-Americans is valuable in qualifying further estimates of characteris-

tics on which Asian-Americans are known to be different from other Americans. Beyond this, the magnitude of the difference, in this case a factor of one half, can be used to place additional weight on the phone responses of Asian-Americans in order to correct for the fact that they tend to be underrepresented in phone surveys.

Sex researchers, who must routinely deal with survey data suffering from very high non-response rates, have applied validation to gain a feel for the ways in which the respondents in their data are different from the U.S. population at large. A straightforward approach is to compare, say, the age distribution among sex survey respondents with the age distribution of the population of Americans as measured by the U.S. Census. Sex survey respondents, in fact, appear to be younger than average Americans are.

Validation is virtually the only way to learn about the characteristics of unit non-responders since, by definition, there is no information on unit non-responders in the rounds of data collection in which non-response occurs. One study by Heather Turner in the *Journal of Sex Research* used validation techniques to uncover some surprising distinctions that need be made among those who are typically categorized together as non-responders. She identified two types of non-responders, differentiating those who refused to participate twice from those who could not be contacted after 17 attempts. Using data from other sources and from follow-up interviews, she discovered that those

non-responders who directly refused to participate in the survey tended to be older, attended church more often, and were more skeptical about the confidentiality of interviews.

An important finding rich with policy implications, she produced evidence suggesting that, in contrast to the low-risk lifestyles of those who directly refuse to participate, the difficult-to-reach non-responders tended to have significantly more sexual partners and higher frequencies of risk factors for AIDS. This demonstrates how difficult it can be to generalize about non-responders and make reliable guesses as to whether non-response bias skews estimates up or down.

Measuring non-response bias in telephone surveys is a frequent concern to polling organizations and those conducting market research by telephone. A fundamental issue confronting anyone attempting to learn about the entire population of Americans based on a phone survey is the fact that not all American households have telephones. Previous attempts to measure the characteristics of non-telephone households indicate considerable differences with respect to phone-owning households across a number of important characteristics such as the propensity to have health insurance.

In a novel approach to measuring non-response bias published in the *Public Opinion Quarterly,* Scott Keeter sought to estimate "telephone non-coverage bias" by conducting a series of phone surveys on the same randomly drawn

sample of phone numbers at several points in time. Among those reached at any given time were, of course, some households who had only recently gained access to a telephone. And among those reached in earlier rounds of phone surveying were some households whose number later became disconnected. Labeling those who gained or lost telephone service at least once as "transient," and comparing the number of transients in his sample with government and industry estimates of how many American households are non-telephone households, Keeter determined that transients make up roughly half of all non-telephone households. Moreover, the demographic characteristics of non-telephone households recorded in other surveys appeared to match those of the transient group in Keeter's study, bolstering confidence in the ability of existing non-response corrected phone survey methodology to produce meaningful insights about the characteristics of American households in general.

Another area of policy research in which non-response bias can play an especially important role is that of valuing natural resources. Developers and government officials often attempt to study the benefits and costs of a proposed building project and must, at some point, put a dollar value on natural resources, including wetlands, endangered animals, and undeveloped green space. Similarly, officials at the Environmental Protection Agency and environmental economists confront the challenge of assessing the value of

parks, wildlife and air quality. Such endeavors must deal with the question of how to reliably elicit valuations that somehow reflect the aggregate preferences of residents. The basic idea is to use samples of citizens to estimate the worth of natural resources in the eyes of an "average" citizen.

It is fairly obvious that the problem of non-representativeness will have a direct effect on such valuations. Suspecting that those who agree to participate in environmental surveys have higher than average subjective assessments of the value of natural resources, researchers in this area worry that non-response bias may lead to overstated valuations. In a 1993 article in *Economics Letters,* John Whitehead and his colleagues employed a combination mail-and-phone survey design in an attempt to produce a bias-corrected valuation of a wetlands preservation project. Using the validation principle, these authors attempted to measure differences between non-responders and responders, both in terms of average demographic characteristics and in terms of willingness to pay for environmental amenities. Validation did, in fact, uncover a disparity between those who initially refused to participate and those who participated without hesitation. Although a non-responder with identical observable characteristics was found to be no less willing to pay than a similar responder, the group of eager respondents included more highly educated individuals and more males. After adjusting for non-response bias, the estimated aggregate willingness to pay fell by 33 percent.

In addition to its application in studying unit non-response, the logic of validation can also be applied to learn about item non-responders as well. Emil Kupek's 1998 article in *Archives of Sexual Behavior* used a large national sex survey in Britain to study the covariates of item non-response. Kupek partitioned his sample into subsamples based on how reluctant individuals were in answering specific questions about their sexual behavior. Specifying the dependent variable to be a measure of each individual's reluctance to respond, Kupek estimated a model relating other demographic variables to the probability of item non-response. Non-responders in Kupek's sample turned out to be less educated and included relatively more non-whites. Perhaps surprisingly, factors such as gender, declared religious affiliation, age and marital status seemed to have little effect on the probability of non-response. As in this study, simply establishing which variables correlate with non-response can amount to a key step in thinking through the broader consequences of non-response and, in particular, whether one's non-random sample will actually lead to bias in estimating the population characteristics of interest.

B. Designing Surveys so that Non-Response Bias Can Be Estimated

An extensive body of research exists analyzing survey methods, seeking to refine their capacity to overcome potential sources of bias. The results, so far, however, are not reassuring. Survey responses are, without question,

very sensitive to the way in which they are elicited. This phenomenon underlies disparaging remarks one frequently hears directed at survey findings in general, such as: "By changing the wording, anything can be shown with surveys." Although this statement is undoubtedly an exaggeration, the sensitivity of survey results to the fine detail of survey design has been demonstrated in numerous academic studies. Hurd et al's 1998 study in *Frontiers in the Economics of Aging* uses experimental evidence to analyze survey non-response and presents a thorough discussion of survey-response sensitivity in the context of estimating aspects of consumption and savings behavior.

The order of survey questions, the gender of the surveyor, re-wordings such as "10 percent survived" instead of "90 percent died," and a number of other seemingly innocuous differences in the implementation of surveys can sharply affect the average response. Relative to mail surveys, face-to-face interviews are known to produce higher reported rates of activities with a high degree of social approval such as volunteering, going to church, and engaging in safe rather than unprotected sex. Non-response rates can also vary dramatically depending on whether data is collected by phone, mail, or face-to-face interviews.

Complicating the picture is that these sensitivities to survey design are not always uniform across all segments of the population. For instance, it has been demonstrated that response rates for whites in face-to-face versus

mail surveys are about the same, yet differ significantly for African-American respondents. Such findings underscore the delicate nature of survey design while raising important issues of interpretation that demand consideration even at subsequent stages of data analysis.

1. Randomized Response

The method of "randomized response" explicitly aims at reducing non-response and misreporting on survey items that concern sensitive topics. The idea behind randomized response is to introduce random questions or random coding procedures into the construction of response data so that it is impossible for the surveyor to infer the respondent's original response by looking at the data recorded for that individual. A survey question on illegal drug-use might employ the following survey design. With probability $1-q$, respondent i is asked, "Have you ever taken an illegal drug," from which the response datum, $y_i = 1$, is recorded if the answer is "yes," and $y_i = 0$ otherwise. But with probability q, the response datum is coded $y_i = 1$ no matter what i's answer was (or without ever asking i the sensitive question). The advantage of the randomized design is its capacity to convince respondents that it is safe to truthfully disclose private information. Randomization is meant to eliminate the possibility of using the randomized data to infer individual answers to sensitive survey questions. If $y_i = 1$, it may be that i answered "yes," or it may be that i happened to fall in the $q \times 100$ percent of the sample for

whom y_i is automatically coded 1.

From randomized response data, an unbiased estimator of the true frequency of drug use, denoted λ, is easy to compute, assuming that randomization induces perfect compliance, i.e., full response and no misreporting. Because

$$\mathrm{E}y_i = (1-q)\lambda + q, \tag{1}$$

the estimator

$$\hat{\lambda} = (\frac{1}{N}\sum_{i}^{N} y_i - q)/(1-q) \tag{2}$$

is unbiased. The price to be paid for introducing randomization, however, is a reduction in the precision of estimation, as can be seen by examining the variance formula for $\hat{\lambda}$.

If answering either "yes" or "no" might be perceived by some in the population as leading to negative consequences, a variation on the set-up above can succeed in making it impossible for any inferences about the answers of survey respondents to be made based on randomized data. By asking, "Is it true that you have never taken illegal drugs," with probability q and asking "Is it true that you have taken illegal drugs," with probability $1-q$, a "doubly randomized" variable y_i results, which equals 1 if the answer to the question (whichever question is asked) is affirmative and 0 otherwise. Multivariate versions of randomization are also possible. Fox and Tracy's 1986 monograph, *Randomized Response: A Method for Sensitive Surveys,* provides

further details. The goal of randomization, in all its forms, is to reduce respondents' skepticism about the confidentiality of their responses. Whether randomization accomplishes its goal is open to debate, however, since it is not clear whether respondents understand randomization sufficiently well or trust the survey designers to follow through with an honest implementation.

2. A Budget Constraint Means Trading off Sampling Error for Bias Reduction

Different survey designs have different price tags and, while more data is always desirable, it is not always obvious how to efficiently allocate spending on data collection given a fixed project budget. In designing surveys with the intention of reducing non-response bias in mind, there is often a nontrivial trade-off to consider when selecting a mix of survey techniques. For a given sum of money, an inexpensive mail survey will likely draw a sample with a higher number of units, thereby reducing sampling error. However, a smaller sample collected using face-to-face interviews will probably enjoy the advantage of a lower unit non-response rate. Thus, one is faced with trading off greater precision (increasing the sample size) against a greater chance that non-response bias will contaminate estimation. In this situation, a sound approach generally involves selecting a mix of sampling techniques that will lead to fairly precise estimates while providing reasonably good controls for non-response bias.

C. Parsing the Meaning of the "Don't Know" Response

A problem faced by most applied researchers working with survey data is in interpreting the meaning of those who provide the response, "Don't know," to a survey question. Those involved at the survey design stage often contemplate whether to prompt those who respond "Don't know" to relent and provide a valid answer. Interestingly, there is debate about whether such prompting is a good idea or not. Insofar as prompting induces random guessing, it is not helpful. But when additional prompting succeeds at extracting additional information rather than noise, one's estimation should, in principle, improve.

For example, public opinion researchers have demonstrated that opinions about political candidates elicited from respondents who say they know nothing about those candidates are, in fact, meaningful indicators of future voting behavior rather than random noise. But in other settings, the evidence points in the opposite direction. As a general rule, the responses of reluctant responders that one collects by means of a special technique of elicitation should be interpreted cautiously, with full acknowledgement that they probably contain more noise than the responses of other respondents.

In some contexts it may be useful to try identifying multiple subgroups among item non-responders. The issue at stake is the extent to which one can generalize about non-responders. Qualitative information about non-

response bias is particularly helpful in instances where it can be presumed that non-response bias mitigates against finding a "significant difference," referring here to an estimated characteristic like average income across two groups. In such a case, without doing anything special to correct for bias, discovering a "significant difference" is especially persuasive, in spite of and, in part, because of the bias. But in other settings, rather than helping to converge to a simple conclusion, gathering additional information about non-responders may complicate the analysis, raising additional questions, and revealing the folly in generalizing about non-responders as if they were a homogeneous subset of the population. Often times, they are not.

D. Panel Data and Attrition

A panel data set contains multiple observations on a fixed group of individuals from whom measurements are collected at several points in time. That is, a random list of individuals is initially chosen, and then those same individuals are surveyed multiple times over the course of months or years. Rather than the snapshot view offered by a cross section in which each observation corresponds to a unique individual, a panel contains a time series for a collection of individuals, which allows researchers to study population characteristics through time. A frequent problem with panel data is attrition, meaning that some respondents surveyed in the initial period later drop out. Respondents who *attrit* can be thought of as those who begin as fully coop-

erative responders but then later become non-responders either by choice or circumstance. In this context, "non-response bias" is sometimes referred to as "attrition bias."

Survey panel respondents may be classified as either "full-time," "monotonic attritors" or "non-monotonic attritors," where "nonmonotonic" refers to a respondent who becomes a nonresponder at some point in time and then rejoins the survey. When all three types are present in a panel, a three-category logit or probit analysis can demonstrate relationships between the probability of attrition and variables that do not change with time, such as gender, age, or other variables such as a dummy variable indicating frequent versus infrequent unemployment. Simpler still, researchers sometimes run a sequence of regressions and examine the effect on regression coefficients of including or excluding attritors. By creating dummies for full-time, monotonic, and non-monotonic attritors, and interacting those dummies with the regressors of interest, standard t tests on interaction terms can produce evidence that attrition is causing bias. As an example, Burkam and Lee's 1998 article in *Journal of Human Resources* applied these techniques to a panel of U.S. high school students, discovering that gender significantly affects the probability of attrition, and also that attrition bias leads to an overstatement of black-white disparity on academic achievement tests.

In another useful example of how to deal with attrition, Fitzgerald et

al's 1998 article, also published in *Journal of Human Resources,* estimated a structural model of attrition, and studied the severity of attrition bias as it relates to a number of standard demographic variables using the Michigan Panel Study on Income Dynamics (PSID). An annual survey panel used frequently by labor economists, the PSID loses roughly 12 percent of the participants each year. More than 20 years after its inception, fewer than 50 percent of the original participants remain. Although the observed characteristics of attritors are noticeably different from full time respondents, coefficient estimates in a variety of models using the PSID, according to Fitzgerald et al, appear to change little when attempts are made to correct for attrition bias. This is good news for researchers attempting to generalize about labor markets in the U.S. based on the PSID.

V. Summary

If one believes that non-responders are different from responders in ways critical to the focus of one's research, then the possibility of non-response bias needs to be taken seriously. Whether designing a survey or analyzing data that have already been collected, a number of interesting techniques may be applied to test for and possibly correct for non-response bias. In the data analysis stage, it is usually best, when feasible, to specify a separate equation for the non-response process and estimate all the parameters simultaneously by maximum likelihood. In particular applications, it can be useful to exploit

other authors' approaches to dealing with the problem of a non-representative sample, even when the problem is not explicitly referred to as "non-response" bias. Rather than attempting to completely "fix" the problems created by non-response, it is often acceptable simply to be sensitive to the potential problems created by non-response, and state to one's readers the likely effect of non-response on the key estimates of interest. Careful attention to the potential problem of non-response is a critical step in conducting high quality research using survey data.

References

Burkam, D. T., and Lee, V. E. (1998). Effects of Monotone and Nonmonotone Attrition on Parameter Estimates in Regression Models with Educational Data: Demographic Effects on Achievement, Aspirations, and Attitudes. *Journal of Human Resources* 33, 555-575.

Fitzgerald, J., Gottschalk, P. and Moffitt, R. (1998). An Analysis of Sample Attrition in Panel Data: the Michigan Panel Study of Income Dynamics. *Journal of Human Resources* 33, 25-74.

Fox, J. A., and Tracy, P. E. (1986). *Randomized Response: A Method for Sensitive Surveys.* Sage Publications, Beverly Hills.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica* 47, 153-161.

Hausman, J. A., Abrevaya, J., and Scott-Morton F. M. (1998). Misclassification of the Dependent Variable in a Discrete-Response Setting. *Journal of Econometrics* 87, 239-269.

Hurd, M. D., McFadden, D., Chand, H., Gan, L., Merrill, A., and Roberts, M. (1998). Consumption and Savings Balances of the Elderly: Experimental Evidence on Survey Response Bias. In *Frontiers in the Economics of Aging* (J. P. Smith, ed.), pp. 387-91. University of Chicago Press, Chicago.

Keeter, S. (1995). Estimating Telephone Noncoverage Bias With a Telephone

Survey. *Public Opinion Quarterly* 59, 196-217.

Kupek, E. (1998). Determinants of Item Nonresponse in a Large National Sex Survey. *Archives of Sexual Behavior* 27, 581-589.

Lee, B. J. and Marsh, L. C. (2000). Sample Selection Bias Correction for Missing Response Observations. *Oxford Bulletin of Economics and Statistics* 62, 305-322.

Lien, D. and Rearden, D. (1988). Missing Measurements in Limited Dependent Variable Models. *Economics Letters* 26, 33-36.

Little, R. J. A. and Rubin, D. B. (1990). The Analysis of Social Science Data with Missing Values . In *Modern Methods of Data Analysis* (J. Fox and J. S. Long, eds.), pp. 374-409. Sage, Newbury Park.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* John Wiley and Sons, New York.

Turner, H. A. (1999). Participation Bias in AIDS-Related Telephone Surveys: Results from the National AIDS Behavioral Survey (NABS) Non-Response Study. *The Journal of Sex Research* 36, 52-66.

Whitehead, J. C., Groothuis, P. A., and Blomquist, G. C. (1993). Testing for Non-Response and Sample Selection Bias in Contingent Valuation: Analysis of a Combination Phone/Mail Survey. *Economics Letters* 41, 215-220.