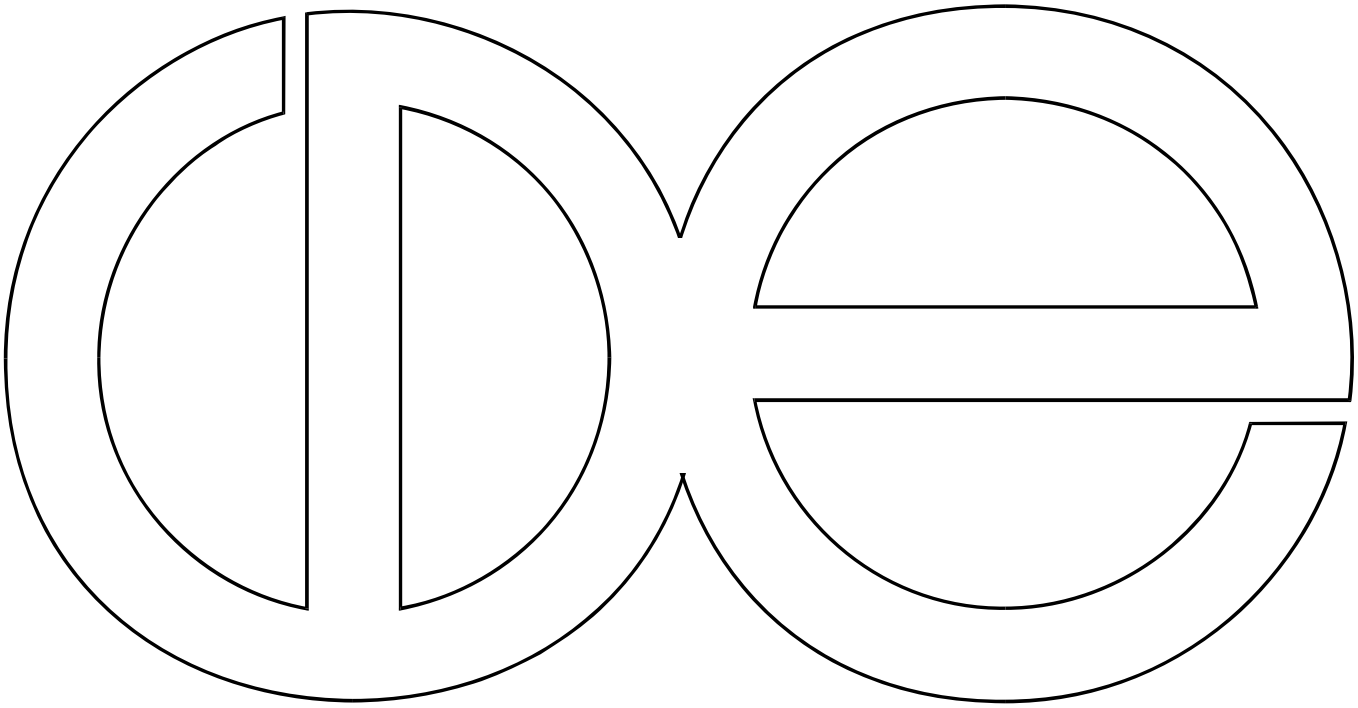
Center for Demography and Ecology

University of Wisconsin-Madison

Coverage Error: The Achilles' Heel of Web-Based Surveys

Salvador Rivas

CDE Working Paper No. 2006-06



Coverage Error: The Achilles' Heel of Web-based Surveys

by

Salvador Rivas, Ph.D.
Center for Demography & Ecology
University of Wisconsin, Madison

April 2006

srivas@ssc.wisc.edu
W: (608) 263-7958
H: (734) 904-0122

Word Count: 6588

Direct all correspondence to Salvador Rivas, Center for Demography and Ecology, 1180 Observatory Drive, 4412 Social Science Building, University of Wisconsin, Madison WI 53705 (srivas@ssc.wisc.edu). A version of this paper was originally prepared and presented at the d3 Conference at the University of Michigan, Ann Arbor. August 2-5, 2001.

*** DRAFT: NOT FOR CITATION OR QUOTATION ***

Science is a cynic, so Karl Popper tells us (cited in Maxim 1999). Its method is the method of elimination and as such it cannot "prove" our assertions, but its logic can show us where we err (Maxim 1999). The recent increase in the diffusion of computer and Internet technology in the United States and other industrialized countries has brought about an interest in this technology as a medium for survey research (Couper 2000; Dillman 2000; Witte 2004). Despite the allure and hyperbole surrounding this new technology, the reality is that its diffusion pattern in the United States is highly stratified by race, income, education, age, and geography (National Telecommunications and Information Administration and Economics and Statistics Administration 2004; Rivas 2004) and cognitive skill and personality (Freese and Rivas 2005). Consequently, the uneven diffusion of computer and Internet technology seriously hinders the viability of Internet/Web-based survey research. As of now, coverage error makes the Internet unsuitable for social science research that aims to be representative of any population that extends beyond the very select group of people making up the online population. If we cannot eliminate the possibility that the unobserved individuals are importantly different from those observed, then we cannot and should not rely on those data to generalize to anyone beyond the collected sample.

Survey methodology, like any other mode of social science research, has its strengths and weaknesses. A well-designed survey that makes use of a carefully selected probability sample in combination with a standardized questionnaire can offer valid and reliable claims about its target population. On the other hand, a haphazardly designed survey that makes use of a convenience sample without regard to representativeness or

question design can result in observations that are at best a snapshot of its own idiosyncratic sample. The prospect for Internet/Web-based surveys, as of now, is more akin to the latter than the former. If an appropriate sampling frame can be assembled for a select population like university professors or company-x employees (where all employees of interest have an email address), then conducting an Internet/Web-survey to assess an aspect of it might be advisable (depending on the topic and types of questions being considered). If, however, the goal of a survey is to evaluate some parameter of the general population the researcher is best advised to follow phone, mail, or face-to-face data collection protocols.

This article has two aims: 1) to argue that as of now the internet is not a viable medium for socially representative survey research and 2) that it need not be so. First, however, I provide a brief history of survey sampling that leads into the recent interest in the Internet/Web as a medium for survey research. I review the four basic survey errors as defined by survey methodologists to introduce the idea that coverage error is critically important when thinking about the viability of nationally representative web surveys. I describe some of the main reasons for the great appeal of using the Internet for social science research, but temper that enthusiasm by reviewing the key reasons why the Internet is not viable as of now or in the immediate future. Despite this gloomy forecast, however, I conclude with reasons to be hopeful as new technology is developed and adopted more widely than current Internet technology.

BRIEF HISTORY OF SURVEY SAMPLING

Contemporary survey methodology dates back to the mid-1930s. Sampling, as a research method, gained general acceptance in 1935 when George Gallup established the American Institute of Public Opinion for the purposes of conducting weekly polls on national political and consumer issues for private and public sector clients. Gallup developed a quota sampling method based on age, sex, and geographic region that normally selected between 1,500 to 3,000 respondents (Rea and Parker 1992). George Gallup and his methodology gained notoriety in opposition to the Literary Digest Presidential Election Poll of 1936. Using a much smaller sample, Gallup correctly predicted the outcome of the Alf London vs. Franklin Delano Roosevelt election. Up until then, the Literary Digest, a popular news magazine published between 1890 and 1938, had correctly predicted American presidential elections from 1920 to 1932 by mailing out postcards to a large number of telephone subscribers and registered automobile owners, then basing its prediction on the returned postcards. The Digest's prediction failed because telephone subscribers and registered automobile owners were disproportionately affluent and Republican, while most of the citizens who voted in that election did not have a telephone nor were they Republican. From this point on, polling organizations realized that it *did* matter who was selected to participate in a survey and that simply collecting information from a large number of respondents did not matter as much as the sampling plan. According to Alain Desrosieres (1998), it was from this event that we gained a strong notion of "representativeness."

Quota sampling, as originally used by Gallup, was improved upon by area probability sampling, which eliminated interviewer influences on the sample selection and thus made it possible to accurately assess the probability of selection for any unit in the population.[1] By the 1960s, telephone surveys became more widely used than personal interviews for two main reasons: 1) by then more than 90% of the U.S. population had access to telephones and 2) door-to-door efforts were becoming too expensive due to the rapid increase of women working outside the home, thus making it more difficult to find respondents at home (Sudman and Blair 1999).

The use of the telephone for survey purposes prompted the refinement and expansion of sample selection by developing random digit dialing (RDD) methodology. Random digit dialing has evolved from using telephone directories as the sampling frame,[2] which was often very inefficient due to problems with unlisted numbers, to a much more efficient strategy of grouping telephone numbers by geographical "exchanges" and then randomly dialing the last 2 digits of a number (Groves 1990). Further improvements have been made to the random selection of individual household members as well. For example, once contact has been established with the selected household, interviewers can ask for the household member with the most recent birthday, thus randomizing the selection process (Sudman and Blair 1999:271).

---

[1] Quota sampling attempts to reproduce a sample that looks very much like the population in general. Interviewers are asked to select so many men and so many women, or so many young and so many older persons for participation, in accordance with the census proportions. In this sense, the interviewer is free to select the sample as long as it fulfills the quota requirements. For more details, see Raj, Des. 1972. *The Design of Sample Surveys*. New York, NY: McGraw-Hill Publications.

[2] Sampling frame is the complete list of names of a target population from which a sample is drawn.

Seymour Sudman and Edward Blair (1999) also document the improvements in mail surveys and the development of other less reliable sampling methods like mall surveys, focus group studies, and other self-selected opinion polls readily available for news organizations. For the purposes of this paper, however, I do not focus on these types of developments because my purpose is to discuss the viability of the Internet as a medium for representative survey research on household populations.

Overall, survey methodology has evolved from simply gathering a large number of responses to more sophisticated efforts that attempt to measure the target population through careful random selection of participants. Survey researchers have also improved survey question design and other related methods (Czaja and Blair 1996; Groves 1989). More recently, computer technology, according to Couper and Nicholls, has had a significant impact on telephone, in-person, and self-administered data collection methods (1998). While the exact effects are still being assessed, it is generally believed that computer assisted data collection is an improvement over previous paper and pencil methods and a must for the current pace of information gathering. In the remainder of the paper, I describe the reasons why the Internet is not viable for nationally representative survey research, but why there are reasons to be hopeful.

SURVEY ERRORS

Survey errors are defined as “deviations of obtained survey results from those that are *true* reflections of the population” (italics added Groves 1989:6). Survey errors are therefore any survey results that do not match up with the theorized, presumably known,

results from the larger target population. These errors, as I briefly review below, occur for various reasons. Some are easy to assess and account for because they are part of the sampling design (e.g., sampling error) while others are much harder to do anything about (e.g., nonresponse error). Given an inexhaustible source of money and interviewing resources, many of these errors can be kept at a minimum, but as Robert Groves aptly puts it, "the art of survey design consists of judging the importance of unmeasurable (or not cheaply measurable) sources of error relative to the measured" (1989: 34). Thus, designing and administering any survey often implies many compromises depending on available resources.

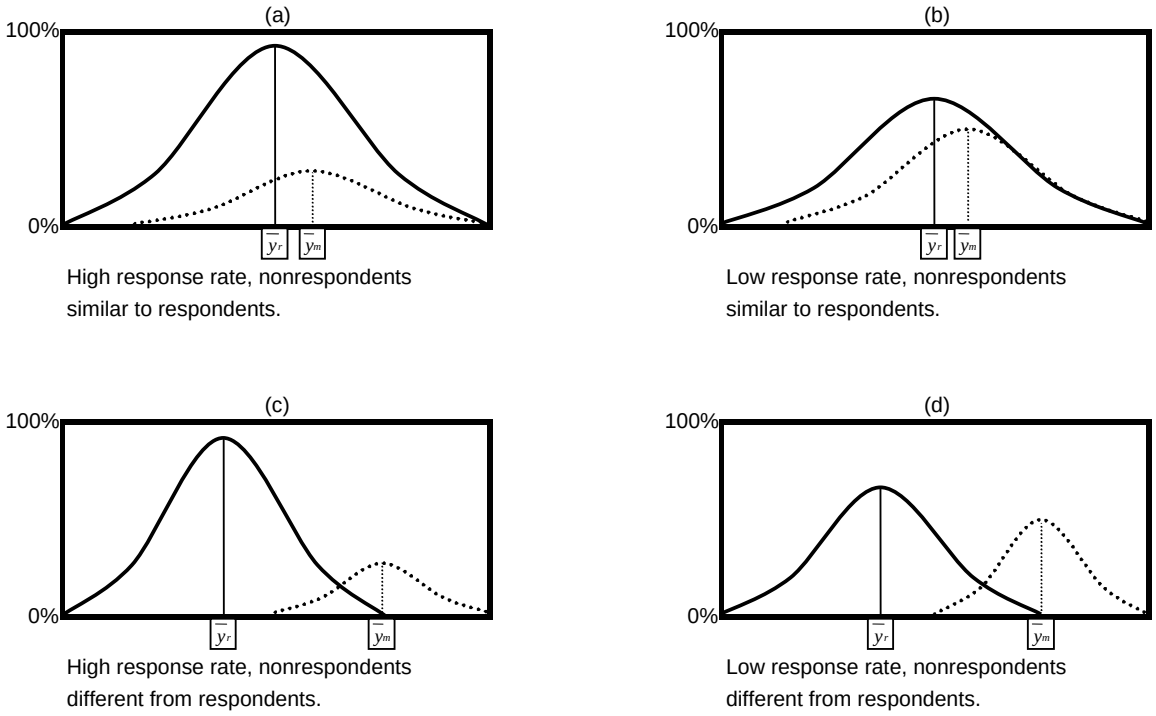
Generally, the consensus is that there are four main sources of error in surveys – coverage error, sampling error, nonresponse error, and measurement error (Groves 1989; Groves and Couper 1998). I briefly review each in following section, but I expand on coverage error in greater detail later in the paper.

Coverage Error: A key requirement for any well-executed survey is a well-defined target population. If the target population is not easily identifiable and listed, drawing a random sample from it becomes a problem. Failure to adequately define and list the target population results in coverage error. More specifically, coverage error occurs when some members of a target population do not have a *known nonzero* probability of selection. That is, when an individual has *no* probability or an *unknown* probability of being included in the sample, we get coverage error. In practice, this occurs when some people are not part of the list or "sampling frame" used to identify

members of the population. A clear example of this occurs when using telephone numbers to sample the U.S. household population because not all households have telephones and thus these households have a zero probability of being selected for the sample. For our purposes, a similar thing would happen when using email addresses to construct a sampling frame from which to draw a nationally representative sample of the United States.

Nonresponse Error: This type of error, akin to coverage error, occurs when selected respondents cannot be located or refuse to participate. Nonresponse error has a sinister effect in so far as those who do not participate are importantly different from those who do. In other words, nonresponse error is really a function of both the nonresponse-rate in combination with any substantive, but unmeasured and unknown difference between respondents and nonrespondents.

Figure 1. Hypothetical Frequency Distributions of Respondents and Nonrespondents



Based on Figure 1.1 in Groves, Robert M. and Mick Couper. 1998. Nonresponse in Household Interview Surveys. New York: Wiley. (p. 4-5).

Figure 1 depicts four possible nonrespondent-effects. Graphs A and B are of little consequence, although neither is desired if it can be avoided. Graphs C and D are worse case scenarios, particularly Graph D. Graph D shows not only a low response rate, but also a large and significant difference between respondents and nonrespondents. The combination of these two factors can be a disaster for the purposes of representativeness and accuracy of results. For example, if the nonrespondents in this case were to be included in the sample as they should be, it is very likely that the observed respondent mean would be substantively and statistically different. Perhaps the most unsettling

aspect of this type of error is that often there is no clear way to detect or check how respondents are different from those who chose to participate. Even more frustrating, however, is that several of the surveys that I list below made no clear effort to mention nonresponse error in their results. Because there are commonly substantial differences between respondents and nonrespondents, nonresponse error is not to be taken lightly (Couper 2000; Groves, Cialdini, and Couper 1992; Groves and Couper 1998; Raj 1972).

Sampling Error: This type of error is not an error as in the sense of making a mistake, but rather a consequence of using samples to predict population characteristics. More precisely, it is a byproduct of calculating a statistic based on a subset or a sample of the population. That is to say, every time one draws a sample from a target population, the sample mean will vary slightly from one time to the next, but over an infinite number of samples of the same size, the mean of the sample distribution equals the true population mean. This type of error is by far the most widely reported error found within the selected digital divide studies below, because it is relatively easy to estimate.

Measurement Error: Unlike the aforementioned errors, which deal with issues of nonobservation, that is, effects of not adequately capturing a representative sample, measurement error is one that occurs when answers by respondents deviate from their true intentions or beliefs. These types of errors can be a result of several things ranging from misinterpretation or miscomprehension of the questions to lack of motivation by the

respondent, the characteristics of the interviewers, social desirability, order of questions, or any other questionnaire design issue.

In a similar fashion to nonresponse error, measurement error can be hard to assess unless specific efforts are made to pretest questionnaires. Other fruitful efforts to minimize measurement error include racially matching interviewers to respondents and properly training interviewers on how to probe without enticing or biasing the responses.

SUMMARY

These four types of errors, according to the survey statistics field, can be summed into the total error of a survey – the "mean square error" (Groves 1989:8). This formulation is important because it allows us to establish a benchmark by which to gauge reported errors. As mentioned earlier, however, designing and implementing a large survey often means plenty of compromises, especially in relation to what is an acceptable degree of error. Groves (1989) notes that most survey designs attempt to reduce only one type of error - sampling error - because it is the easiest to measure and control and thus to justify in terms of cost. This choice, however, is misleading because it assumes that a reduction of this type of error will also reduce the other three.

In fact, according to Groves (1989), it is often the case that researchers attempt to weight their samples by using the Current Population Survey or U.S. Census demographic characteristics to create a more "representative sample." While this technique does provide a multiplier by which to adjust raw number counts in the sample, it *does not* and *cannot* adjust for how the missing nonrespondents may have truly

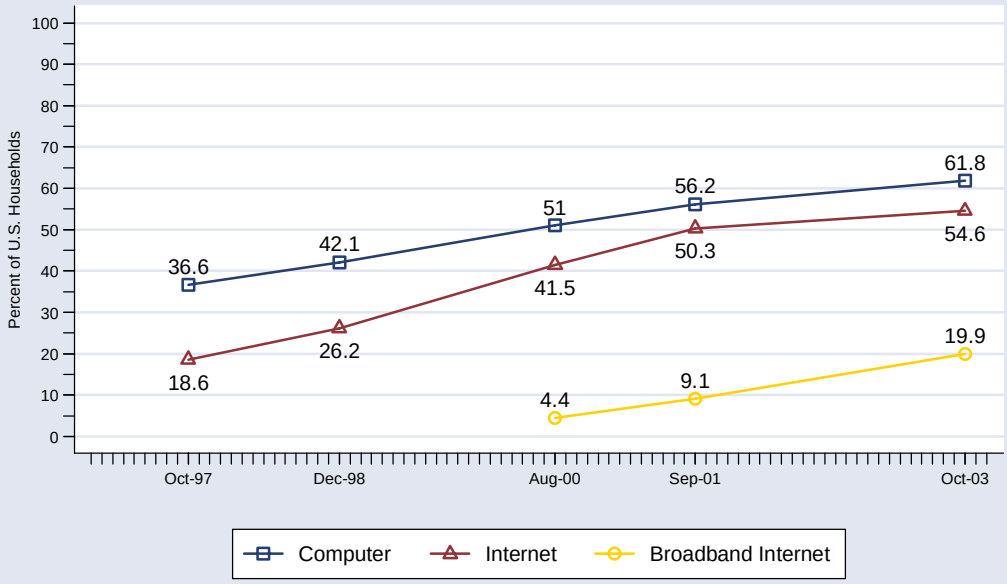
answered had they been included in the sample. That is, it only can account for response biases to the extent they are correlated with the demographic variables used in adjustment. However, nonresponse may be associated with many other variables other than demographic characteristics. Therefore, it is very problematic to assume that by adjusting marginal counts one will also appropriately adjust responses to attitude questions about computer and Internet technologies or any other socially sensitive question. More succinctly, this tactic assumes no difference between respondents and nonrespondents within the categories used to adjust response distributions.

PROSPECTS OF INTERNET/WEB SURVEYS

The recent diffusion of personal computers and the subsequent adoption of various forms of Internet communication programs and devices have brought about the next step in the evolution of survey research – Internet/Web survey methodology (Best, Krueger, Hubbard, and Smith 2001; Couper 2000; Couper 2001; Crawford, Couper, and Lamias 2001; Dillman 2000).[3]

[3] For an updated list of Web Survey literature, visit http://www.ris.org/indexuk html.

Figure 2. Percent of Households with Computer and Internet Connections, Selected Years, 1997-2003



Source: NTIA (2004) 'A Nation Online: Entering the Broadband Age.' U.S. Department of Commerce

Figure 2 shows the percent of households with a Computer, an Internet connection, or Broadband Internet access from 1997 to 2003. Using Current Population Survey data, the percentage of households that report having a computer has grown from 36.6 in 1997 to 61.8 percent in 2003. The percent of households with an Internet connection has also grown from 18.6 in 1997 to 54.6 in 2003. Starting in August 2000, the Current Population Survey began asking about broadband internet access and found that only 4.4 percent of households had such access but that percentage grew to 19.9 by 2003. These figures have paved the way for interest in the web/Internet not only for commercial purposes (e.g., book retailers like Amazon.com or Barnes & Noble) but also data collection via web-surveys.

The growth in computer and Internet access has increased interest in web/Internet survey methodology. As reported by Vicki Pineau (2005) overall spending for online research increased by 53% between 2000 and 2001, another 61% in 2002, and this growth rate is expected to continue for the time being. Samuel Best and his colleagues (2001) note that the Internet presents an unprecedented opportunity for data collection since it makes possible the implementation of complex survey instruments that can deliver video and audio in addition to traditional text. Moreover, the Internet also offers quick and efficient capabilities to collect, store, and manage survey data (Dillman 2000). These capabilities fuel the enthusiasm for this medium as the possibilities seem limited only by the creativity of the researcher. Consequently research institutions, mainly market research firms, have increased their efforts to collect surveys via the web (e.g., Forrester, HarrisInteractive).

While computer and Internet access has increased significantly over the last decade, the diffusion pattern is not equal across age group, education level, income, region, urbanicity, or racial group. Using the October 2003 cross-section of Current Population Survey data on computer and Internet usage, Table 1 shows the distribution of reported Internet access for those ages 18 and over for each of the factors noted above by racial group. The table shows that only 35.8% of Hispanics report having access to the Internet, either at home, work, school, library, or a friend's home. That value is followed by 45.9% of individuals that identify as Black, and 52.2% of Native Americans/Others. Finally, 65.4% of Whites, and 65.6% of Asians report having access to the Internet as of October 2003. Simply based on this cross-classification we find that anywhere from

34.4% to 64.2% of Americans did not have access to the Internet at that point in time, suggesting that at best nearly 35% of U.S. adults would be excluded from a web/Internet-based survey whose goal was to capture a nationally representative sample.

When we consider Internet access by age, level of education, income, region, or urbanicity, we also find inequality in the diffusion pattern. Internet access is highest among the younger age groups, the highly educated, the affluent, those living in suburban areas, and those in the Western region of the United States. This particular pattern does not bode well for a web/Internet-based research study whose aim is to understand attitudes among the poor, low-educated, elderly living in southern rural areas. In general then, Table 1 clearly shows that as of October 2003, Internet access was not dispersed equally enough among the adult, non-institutionalized, civilian population to be an appropriate medium for collecting population level data.

In previous work, I show that even after controlling for important sociodemographic and socioeconomic factors Black and Hispanic households are significantly less likely than White and Asian households to report having a computer at home (Rivas 2004). For our purposes, this finding is particularly important because it suggests a significant difference by race that expands beyond the usual SES suspects, which as indicated earlier are often used to weight or adjust known sampling biases.

Such error should be of particular concern to students of the digital divide because differences in access to computers and the Internet have been closely associated with issues of poverty and low education. Unfortunately, these factors are also associated with those who *do not* have telephones, thus making this portion of the population much less

likely to be captured by random digit dialing telephone surveys. The consequences of this drawback have not yet been fully appreciated in recent research on the digital divide.

MEASURING THE EFFECT OF COVERAGE ERROR

As noted above, a key requirement for any well-executed survey is a well-defined target population. If the target population is not easily identifiable and listed, drawing a random sample from it becomes a problem – resulting in coverage error. In other words, whenever any or some members of a target population do not have a *known nonzero* probability of selection or said differently, when an individual has *no* probability or an *unknown* probability of being included in the sample, we have coverage error.

The effect of coverage error on a linear statistic can be expressed as a function of two components: the proportion of the target population that is not covered by the frame and the difference in the survey statistic between those covered and those who are not. This relationship can be expressed in the following form:

$$Y = \frac{N_c}{N} Y_c + \frac{N_{nc}}{N} Y_{nc},$$

where Y = the population parameter being estimated;

N_c = number in the target population covered by the sampling frame;

N_{nc} = number in the target population not covered by the sampling frame;

N = total number in the target population;

Y_c = value of the statistic for those covered by the sampling frame;

Y_{nc} = value of the statistic for those not covered by the sampling frame;

The equation above can be reworked to illustrate the nature of noncoverage error

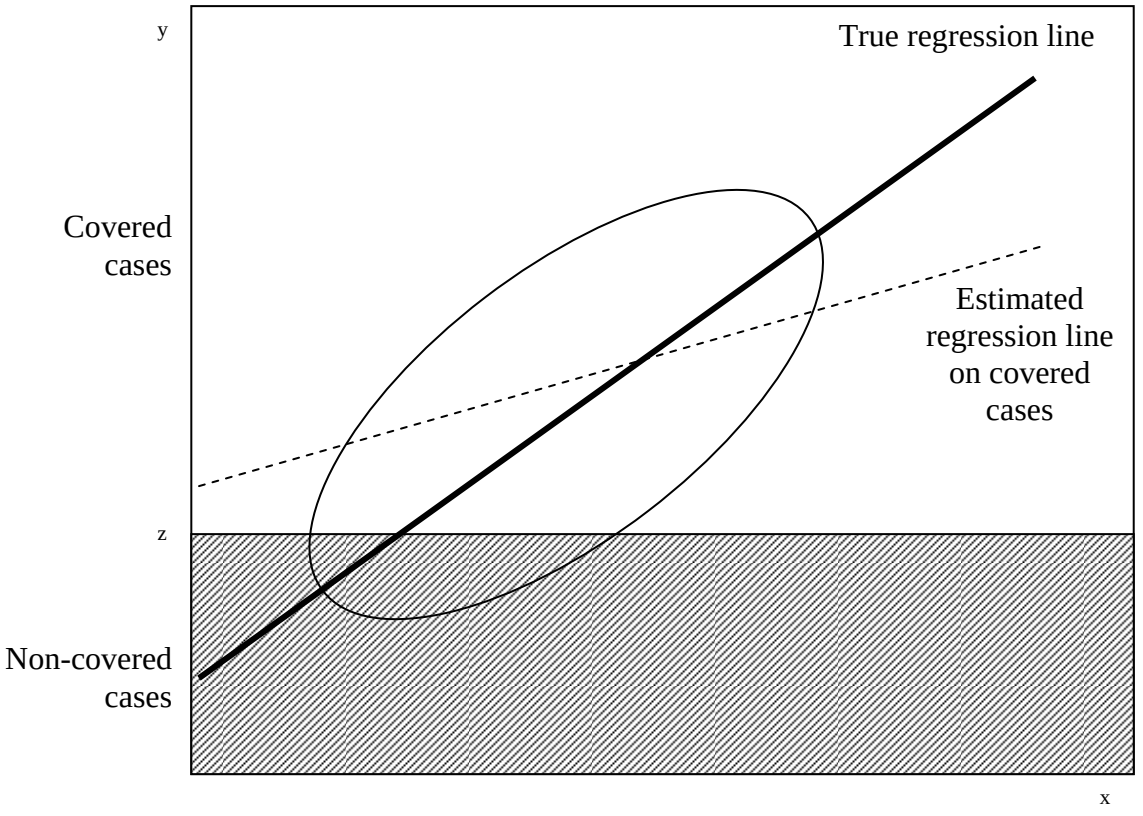
$$Y_c = Y + \frac{N_{nc}}{N}(Y_c - Y_{nc}).$$

That is to say, the value for the covered sample (Yc) is equal to the target population value plus the proportion not covered times the value of those covered minus the value of those not covered. Consequently, this equation suggests that even if the noncovered cases comprise a large fraction of the population frame but that fraction is not importantly different from those cases that were covered, the bias due to noncoverage will be minimal. Conversely, if the fraction not covered is relatively small, but those cases are significantly and importantly different from those covered then a large noncoverage bias may result. Graphically, these scenarios are similar to those depicted in Figure 1 for the nonresponse effects. While the effect of both noncoverage and nonresponse is similar; the mechanisms that produce them are distinct. Noncoverage error arises from incomplete sampling frames while nonresponse error is generally the result of noncompliance by units that have been selected into the sample.

In Figure 3, we see the effect of "truncation" on the variance of the dependent variable Y. The effect of the non-covered cases on the regression line is the result of the reduced (truncated) variance around the dependent variable. The assumption is that the missing, or non-covered, cases would not only add to the variance around Y were they not missing, but that these cases are also substantively different in relation to the

dependent variable. More specifically, Figure 3 shows two regression lines; the dashed line representing the estimated regression line on covered cases only and the darker line representing the "true" regression line that includes all target cases.

Figure 3: Effect on estimated regression line under truncation of the sample on the dependent variable (y); all $y_i < z$ not covered.



Source: Adapted from Groves, Robert M. 1989. *Survey Errors and Survey Costs*. New York: John Wiley & Sons.

Unfortunately, it is often nearly impossible for researchers to know exactly how large the coverage error is for any given statistic. The reason for this is that it is often

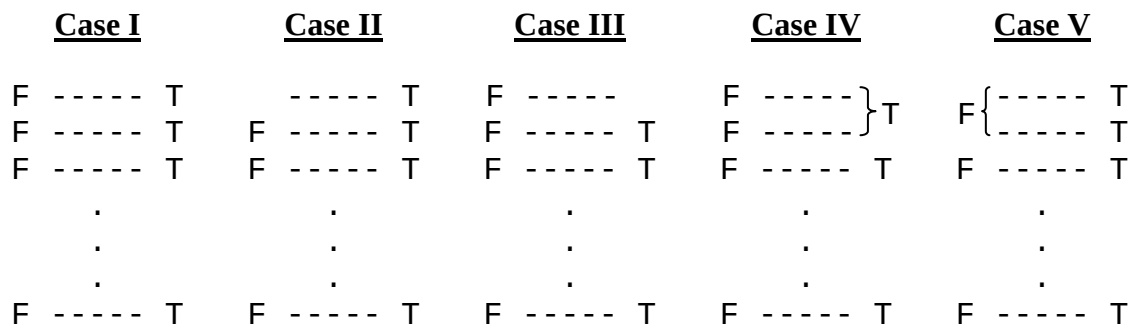
difficult to know how large is the proportion, N_{nc}/N, not being covered and similarly, it is difficult to know how different are the noncovered cases from those that are covered. Moreover, even though N_{nc}/N is constant over the entire sample, it will vary over different subclasses on which statistical estimates might be calculated, for example racial and ethnic categories. The term $(Y_c - Y_{nc})$ can also vary over all estimated statistics. For example, in the United States the omission of Hispanics and Blacks is likely to have a larger effect on our estimates of the proportion of people that are unemployed as opposed to an estimate of the proportion of people that have ever driven a car. In the first example, being Hispanic or Black is significantly associated with being unemployed while the second example is not necessarily so, thus coverage bias would matter in former more so than in the latter. Nevertheless, as Bob Groves (1989:95) states, "Coverage error can affect both simple and complex statistics calculated on sample surveys." For this reason alone coverage error is why the Internet is not viable as the sole medium by which to collect nationally representative data.

TYPES OF COVERAGE ERRORS

Beyond the difficulties noted above, there are also practical problems that can arise related to coverage issues. Figure 4 presents a typology of coverage problems identified by Kish (1965). The "F's" in Figure 4 represent elements of a sampling frame that are listed as being part of the set. The "T's" are members of the target population; cases that should be included in the study. Ideally, a survey would only ever contend

with Case I scenarios, where there is a one-to-one correspondence between the population frame and the target population.

Figure 4: Components of Coverage Error



Source: Adapted from Groves, Robert M. 1989. *Survey Errors and Survey Costs*. New York: John Wiley & Sons.

More frequently, however, surveys encounter Case II coverage problems in which members of the target population do not appear in the population frame, the list of elements from which the sample will be drawn. For our purposes, our target population could be all civilian adults that are not institutionalized, but given the current state of Internet diffusion only a fraction of this population would ever appear on a list of email addresses from which a sample could be drawn.

Case III scenarios are when the frame includes elements that are not part of the target population. In relationship to phone telephone surveys for example, this type of situation would occur when the survey was meant to sample individuals, but the list of phone numbers from which the sample was to be drawn from included businesses and

other organizations. Similarly in the case of email addresses, this scenario would be the case when the list of email addresses included those of businesses or any other non-individual organization email address. Moreover, users often have multiple addresses that are not associated with any particular internet service provider (e.g., yahoo.com, gmail.com). Thus, distinguishing between personal versus work/organization email addresses in some cases might be easy (e.g., *username*@dell.com) but in other cases it might not be (e.g., *username*@heiberg.com). In the former case, we can assume that the address at Dell.com is most likely an employee since Dell is not yet an Internet service provider. In the second example, however, the user could be a small business owner and his/her address at Heiberg.com is the only email address through which he/she communicates on a daily and consistent basis.

Case IV issues are especially likely to be a problem for Internet/Web surveys. In these scenarios, the sampling frame contains multiple entries for the same target population element. Multiple email addresses for one individual, for example, would be a case of this, also known as "overcoverage." A similar problem exists for personal telephone numbers, which is aggravated if mobile telephone numbers are included in the sampling frame. In the cases where it can be verified that one individual is associated with multiple email addresses, then one address can be randomly chosen to represent that individual, but often times the usernames (handles) are not specific names easily associated with one individual versus another (e.g., chevy74@yahoo.com, warxal@gmail.com, etc.). However, even if this were not known, the researcher could

ask the respondent whether he or she has multiple email addresses and weight their response down in proportion to the number of their email addresses.

The final type of coverage error (Case V) identified by Leslie Kish corresponds to scenarios where one frame element corresponds to multiple target population elements. A clear example of this occurs when one household address corresponds to multiple household members. Or in our case, Case V issues would arise when one family decides to use one email address for the entire family when in fact our goal is to give every individual the same probability of selection into the study. In this case, the researcher could select a respondent at random from among those using the same email address and weight their responses upward in proportion to the number of persons using the address. Adjusting for this type of case would not be hard if the survey were able to verify the number of individuals sharing an email address and to choose a respondent at random; without such procedures, however, there would be a serious a flaw in the design.

OVERCOMING THE WEAKNESS

Despite its wonderful, theoretically possible capabilities, the Internet is not *currently* a medium for the collection of unbiased random sampling of large noninstitutional populations. This shortcoming, as Mick Couper notes (2000) can be overcome if the researcher's target population is *only* current Internet users or members of a specific subgroup for whom internet access is universal and the researcher has a complete sampling frame (e.g., college students, company employees). The latter

criterion is much harder to obtain given our current anti-spamming rules and the fact that many people have multiple e-mail addresses.

Coverage error is not the only complication that researchers must surmount to adequately design a web/Internet survey. As noted earlier, unequal access is one of the difficulties in using the Internet for survey research, but only a part. In the abstract, everyone in the United States has access to the Internet either at home, at work, school, or at their local library. Access, then, is not the real problem but rather that a complete list of email addresses for everyone in the United States, for example, does not exist. Therefore, building an adequate sampling frame from which to draw a random sample is impossible as of now.

In addition, Crawford et. al. (2001) suggest that web surveys may be perceived to be more trouble than they are worth. That is, unlike traditional mail surveys in which the respondent can peruse the instrument at their leisure and make a decision about completing it or not based on how much work they believe it to be, web surveys do not allow for this assessment. Therefore, unless special efforts are made to lessen the "perceived burden" respondents will not be likely to participate.

In the case of the United States, building a population frame from which to sample would be difficult and complex for many reasons. One potential approach to building a population frame would be to assign everyone a unique username in conjunction with their social security number, without of course exposing the actual number itself. While this may sound ideal, the administrative and technical workload alone would be a daunting task as the U.S. population pushes beyond 300 million. But

even if the administrative and technical problems were not an issue, issues of privacy and citizenship would be. Who would have access to these unique usernames? How would we include immigrants into the population frame? How would we include undocumented immigrants?

As James Witte (2004) suggests, a possible solution in the short term could be to take a multimethod approach to data collection. Such an approach would stratify and match the population with modes that are likely to minimize nonresponse and coverage errors. This approach, however, would have to assume or guard against potential mode-effects. In general, dual-frame, mixed mode methods can yield fruitful discoveries not only about the population of interest, but also about mixed-mode survey methodology.

In February of 2005, the city of Philadelphia decided to make free wireless connectivity available to all its citizens who were within the broadcasting range of its wireless system (Dao 2005). Later that year, the city of Alexandria, Virginia also announced a similar plan (Gowen and MacMillan 2005). Developments like these are a step in the right direction: reducing the digital divide and also reducing the amount of potential coverage error associated with web-surveys. These developments, however, still require that individuals have a computer or some sort of device with which to access the free wireless services. Of course, this also assumes that we can assemble a universal email address list as discussed above.

Another recent development still in its infancy is the diffusion of "smartphones." A smartphone is any electronic handheld device that integrates the functionality of a mobile phone, personal digital assistant or other information appliance (Wikipedia). A

key feature of a smartphone is that additional applications can be installed on the device. The applications can be developed by the manufacturer of the handheld device, by the operator or by any other third-party software developer. As of October 2005, the Cellular Telecommunications & Internet Association (CTIA) estimates that there are 194.5 million wireless phone subscribers in the United States with approximately 6% of households being completely Wireless (CTIA 2005). That is to say, about 65% of the U.S. population are wireless subscribers. These numbers, however, do not distinguish between regular cellular telephones and "smart phones." Nevertheless, smartphones, small and highly portable devices that serve many functions, are likely to diffuse wider and faster than personal computers and thus help overcome the computer/Internet divide that currently exists. When that occurs, web-surveys for social science research will be much more viable than at present.

The use of smartphones and other mobile technology assumes that cost is only charged to the caller. Of course, the marginal cost to a potential survey respondent using a hard-wired PC is zero, and the same holds for a fixed-cost wireless connection, but most mobile communication systems still charge recipients as well as initiators of contacts. Thus, the proposed use of smartphones and similar technologies will not be viable until there is not only universal coverage, but universal coverage with only the caller paying as well.

CONCLUSION

In general, Internet/Web surveys are susceptible to coverage error (Couper 2001) since only about 60% of the U.S. population is online. More problematic, however, is that being online is strongly associated with income, education, and race. In so far as the people online are substantively different than those who are not, sampling the current online population will render a biased sample. Any attempt to collect data through this medium with the aim of generalization to the entire population is likely to be skewed. Thus, aside from meaning that certain groups do not have access to the vast amount of information online, the digital divide also has consequences for social scientists who wish to use the web to conduct nationally representative studies. This divide is the main reason why coverage error exists in Web-based surveys, and this source of coverage error makes the Internet unsuitable for social science research that aims at being representative of any population that extends beyond the very select population of users.

If these shortcomings can be overcome, by establishing nearly universal access, as we have done with the telephone, then the Internet will become more viable as a survey medium. Until then, however, unless our interests lie in interviewing specific segments of the population online, our best bet is to rely on traditional survey methods – in-person, mail, or telephone assuming proper and adequate sampling schemes.

Future research on the digital divide should consider, and whenever possible incorporate, appropriate controls and checks for each of the types of errors discussed in this paper. Furthermore, given what we know about telephone penetration rates among

the poor and less educated, it is important that special efforts be made to survey this often neglected subgroup. In addition, with the constant bombardment of media ads for computers and Internet technologies, it might be prudent to watch for interview desirability effects among certain groups in the population; that is, there may be pressure to declare that people have access to a computer or the Internet when in reality they do not. It is possible that a certain level of negative stigma is felt among those who see computer ownership or Internet usage as a status symbol. If we disregard this possibility, we might overestimate computer ownership and Internet access and thus diminish efforts to bridge the digital divide.

In summary, I close with a quote by Des Raj (1972). He states, "It should be conceded…that we do need figures to make the right type of decision. Government, business, and the professions all seek the broadest possible factual basis for decision-making. In the absence of data on the subject, a decision taken is just like leaping into the dark. We need statistics, in fact better and better statistics" (3). This statement may be truer than ever now that we find ourselves in the so called "Information Age." As consumers, producers, and processors of statistical information, we need to learn how to effectively deal with and decipher the large amount of data increasingly available to us. Moreover, as the world population grows, household censuses will be simply too expensive to administer. Consequently, survey methodology will offer the most convenient and least expensive approach for measuring human activity.

As I have noted above, survey research has its weaknesses and Web-based survey research in particular, has its Achilles ' heel, but if done carefully it can yield strong and

reliable data. Overall, we must be cautious not to be too cynical about large numbers or to dismiss them as too general and disconnected from reality – at the grass-roots level. Instead, we must learn as Darrell Huff suggests, "…[to] look a phony statistic in the eye and face it down; and no less important, how to recognize sound and usable data in that wilderness of fraud…" (1983:122). Although perhaps a bit alarmist, Huff's general point is welcome in our age – when data appear to be everywhere, and the pressure to publish and publicize can lead us away from being critical of the source of the data.

REFERENCES

Best, Samuel J., Brian Krueger, Clark Hubbard, and Andrew Smith. 2001. "An Assessment of the Generalizability of Internet Surveys." *Social Science Computer Review* 19:131-145.

Cellular Telecommunications & Internet Association. 2005. "Wireless Quick Facts." Washington, DC.

Couper, Mick P. 2000. "Web Surveys: A Review of Issues and Approaches." *Public Opinion Quarterly* 64:464-494.

—. 2001. "The Promises and Perils of Web Surveys." in *The Challenge of the Internet*, edited by A. Westlake, W. Sykes, T. Manners, and M. Rigg. Chesham, UK: Association for Survey Computing.

Couper, Mick P. and William L. Nicholls II. 1998. "The History and Development of Computer Assisted Survey Information Collection Methods." Pp. 1-21 in *Computer Assisted Survey Information Collection*, edited by M. P. Couper, R. P. Baker, J. Bethlehem, C. Z. F. Clark, J. Martin, W. L. Nicholls II, and J. M. O'Reilly. New York: John Wiley & Sons, Inc.

Crawford, Scott D., Mick P. Couper, and Mark J. Lamias. 2001. "Web Surveys: Perceptions of Burden." *Social Science Computer Review* 19:146-162.

Czaja, Ronald and Johnny Blair. 1996. *Designing Surveys: A Guide To Decisions And Procedures*. Thousand Oaks, Calif: Pine Forge Press.

Dao, James. 2005. "Philadelphia Hopes to Lead the Charge to Wireless Future." in *New York Times*. New York, NY.

Desrosieres, Alain. 1998. "The Politics of Large Numbers: A History of Statistical Reasoning." Cambridge, MA: Harvard University Press.

Dillman, Don A. 2000. *Mail and Internet Surveys: The Tailored Design Method*. New York, NY: J. Wiley.

Freese, Jeremy and Salvador Rivas. 2005. "Internet Adoption Among Late-Midlife Adults." in *Gerontological Society of America*. Orlando, FL.

Gowen, Annie and Robert MacMillan. 2005. "The Web Is in the Air: Alexandria Offers Free Outdoor Access to Wireless Internet." in *Washington Post*. Washington, DC.

Groves, Robert M. 1989. *Survey Errors and Survey Costs*. New York: John Wiley & Sons.

—. 1990. "Theories and Methods of Telephone Surveys." *Annual Review of Sociology* 16:221-240.

Groves, Robert M., Robert B. Cialdini, and Mick P. Couper. 1992. "Understanding The Decision to Participate in a Survey." *Public Opinion Quarterly* 56(4):475-495.

Groves, Robert M. and Mick P. Couper. 1998. *Nonresponse in Household Interview Surveys*. New York: Wiley.

Huff, Darrell. 1983. *How to Lie With Statistics*. New York: W. W. Norton & Company.

Kish, Leslie. 1965. *Survey Sampling*. New York, NY: John Wiley.

Maxim, Paul S. 1999. *Quantitative Research Methods in the Social Sciences*. New York, NY: Oxford University Press.

National Telecommunications and Information Administration and Economics and Statistics Administration. 2004. *A Nation Online: Entering the Broadband Age*. Washington, D.C.: U.S. Department of Commerce.

Pineau, Vicki. 2005. "Web-based Surveys: What Do We Know about Data Quality?" in *Joint Statistical Meetings*. Minneapolis, Minnesota.

Raj, Des. 1972. *The Design of Sample Surveys*. New York, NY: McGraw-Hill Publications.

Rea, Louis M. and Richard A. Parker. 1992. *Designing and Conducting Survey Research: A Comprehensive Guide*. San Francisco: Jossey-Bass Publishers.

Rivas, Salvador. 2004. "The U.S. Digital Divide: Race, SES, and Social Context as Predictors of Computer Ownership, 1997-2001." Dissertation Thesis, Department of Sociology, University of Michigan, Ann Arbor.

Sudman, Seymour and Edward Blair. 1999. "Sampling in the Twenty-First Century." *Journal of the Academy of Marketing Science* 27:269-277.

Wikipedia, The Free Encyclopedia. "Smartphone." (online: http://en.wikipedia.org/w/index.php?title=Smartphone&oldid=37559516)

Witte, James C. 2004. "Prologue: The Case for Multimethod Research." Pp. xv-xxxiv in *Society Online: The Internet in Context*, edited by P. N. Howard and S. Jones. Thousand Oaks, CA: Sage.

Table 1. Reported Internet Access for Age, Education, Household Income, Region, Urbanicity by Race/Ethnicity for Adults (age 18+), CPS 2003 (weighted).

	White	Black	Hispanic	Asian	Native American	TOTAL
Has Internet Access	*65.44*	*45.88*	*35.77*	*65.62*	*52.15*	*59.48*
Age18 to 19	82.02	60.83	50.11	87.83	53.30	73.80
Age 20 to 24	77.72	58.04	45.20	82.71	73.04	69.44
Age 25 to 29	77.82	55.32	36.39	83.20	60.41	66.66
Age 30 to 34	80.57	59.48	39.41	77.73	68.77	70.15
Age 35 to 39	78.33	53.12	36.73	73.48	50.40	68.31
Age 40 to 44	76.92	47.19	38.37	66.61	66.28	68.08
Age 45 to 49	74.89	43.71	37.42	66.05	40.78	66.85
Age 50 to 54	71.11	46.48	33.18	56.18	50.55	64.30
Age 55 to 59	66.31	42.04	29.32	55.00	52.15	60.78
Age 60 to 64	54.96	28.12	19.27	44.10	35.73	49.05
Age 65 to 69	42.34	17.80	13.02	24.24	36.45	37.40
Age 70 to 74	30.97	13.85	10.92	25.79	7.25	27.88
Age 75 to 79	25.11	7.80	9.08	13.29	9.89	22.67
Age 80 +	11.16	4.75	4.49	13.70	1.63	10.38
Less than high school	25.05	16.67	12.53	21.84	22.24	19.59
High school degree	50.81	34.22	35.67	43.35	42.01	46.72
Some college	74.92	60.69	61.36	71.28	74.53	71.82
College degree	87.23	78.25	76.09	81.32	84.49	85.45
More than college	88.14	84.48	82.44	91.87	86.44	88.04
Blank/DK/Refused	53.35	40.34	28.79	54.78	38.07	48.99
Under $25,000	37.93	26.26	21.35	47.58	37.41	32.93
$25,000 - $49,999	61.73	50.67	37.46	59.96	59.77	56.60
$50,000 - $74,999	78.16	70.12	52.66	71.75	72.87	74.74
$75,000 - $99,999	84.72	73.62	68.64	78.15	84.93	82.52
$100,000 or More	90.82	81.68	77.55	88.92	94.78	89.72
Central City	68.16	43.19	31.98	59.44	57.84	55.02
Suburban	69.56	56.74	40.64	70.50	66.90	65.18
Rural	54.17	29.49	30.20	61.71	37.08	50.62
Not Identified	65.97	42.59	36.09	69.99	65.26	60.95
Northeast	64.24	49.91	37.87	62.26	58.07	60.11
Midwest	63.92	44.66	35.55	77.82	43.46	61.00
South	63.28	43.73	35.48	69.68	54.19	56.49
West	72.47	53.60	35.32	62.90	53.36	62.10

Note: Reported Internet access is from anywhere (e.g., home, work, school, library)
Source: Current Population Survey, October 2003

Center for Demography and Ecology
University of Wisconsin
1180 Observatory Drive Rm. 4412
Madison, WI 53706-1393
U.S.A.
608/262-2182
FAX 608/262-8400
comments to: srivas@ssc.wisc.edu
requests to: cdepubs@ssc.wisc.edu