COVERAGE ERROR IN ESTABLISHMENT SURVEYS

Carl A. Konschnik
U.S. Bureau of the Census[1]

I. Definition of Coverage Error

Coverage error which includes both undercoverage and overcoverage, is defined as "the error in an estimate that results from (1) failure to include all units belonging to the defined population or failure to include specified units in the conduct of the survey (undercoverage), and (2) inclusion of some units erroneously either because of a defective frame or because of inclusion of unspecified units or inclusion of specified units more than once in the actual survey (overcoverage)" (Office of Federal Statistical Policy and Standards, 1978). Coverage errors are closely related to but clearly distinct from content errors, which are defined as the "errors of observation or objective measurement, of recording, of imputation, or of other processing which results in associating a wrong value of the characteristic with a specified unit" (Office of Federal Statistical Policy and Standards, 1978). Thus, an interviewer's failure to properly identify and hence to record data for what should be a selected unit is a coverage error. On the other hand, failure to pick up data for a properly selected unit (which results in an imputed value being assigned to the unit) is a content error. Content errors include response and nonresponse errors. However, content errors as well as other nonsampling error types will not be discussed in this paper apart from contrasting them to coverage error.

II. Sources of Coverage Error

While the definition divides coverage error into two major components--undercoverage and overcoverage--another important duality is implied within each of these. Coverage error shows up (1) in defective sampling frames and (2) as a result of defective processes associated with the selected sample. (Sampling frame, or stated simply, frame is used here to mean the collection of sampling units, either given explicitly as a list or implicitly in terms of well-defined procedures.)

Thus coverage error results either because the frame does not properly represent the sampled population, or because the sample does not properly represent the frame. Note that, using the definitions of Cochran (1977), we are making a distinction between the sampled population, defined as the population to be sampled, and the target population, defined as the population about which information is wanted (if possible). Ideally, the sampled and target populations should coincide. However, cost or other practical considerations sometimes result in a lack of coincidence between the two. Consequently, the target population is usually modified to coincide with a workable sampled population.

Any difference between the sampled and target populations can contribute importantly to coverage error, especially where excessive compromise in the survey planning stage results in a sampled population which is too far removed from the target population. Since estimates based on data drawn from the sampled population apply properly only to the sampled population, interest in the target population dictates that the sampled population be as close as practicable to the target population. Nevertheless, in the following discussion of the sources, measurement, and control of coverage error, only deficiencies relative to the sampled population are included. Thus, when speaking of defective frames, only those deficiencies are discussed which arise when the population which is sampled differs from the population intended to be sampled (the sampled population).

Coverage Error Source Categories

We will now look briefly at the two categories of coverage error--defective frames and defective processes associated with the selected sample.

Defective Frames--Defective frames are characterized by (1) deficiencies in meeting the requirement that every element of the sampled population belongs to one and only one sampling unit, (2) erroneous inclusion of units (including the wrong units or having duplicates of units which belong in the frame), or (3) erroneous exclusion of sampling units. These problems can result from vague or unworkable definitions of the sampling units relative to the sampled population; improper procedures or processing in establishing and maintaining the frame; timing, which affects the updatedness (agreement with the proper reference period) of the frame; or miscoding of sampling units. Erroneous inclusion (overcoverage) results from including duplicates and out-of-scope or out-of-business units. Erroneous exclusion of sampling units (undercoverage) results from failure to include the proper units or failure to account for birth (new) units. Misclassification of units, such as for Standard Industrial Classification (SIC), geography, size class, or company structure can lead either to undercoverage or overcoverage.

Some frame problems cannot be overcome without expending significant resources. For example, most frames suffer from some degree of outdatedness. A monthly survey in which the frame and sample are updated quarterly, such as the Census Bureau's Monthly Wholesale Trade Survey (MWTS), does not have an up-to-date frame for at least two out of every three months--and this is over and above the lag time in getting new units on the list frame. This time lag itself can be as much as 12 to 18 months after a business starts up. For example, the Social Security Administration (SSA) lists of Employer Identification (EI) numbers newly assigned by Internal Revenue Service (IRS) are given to the Census Bureau after SSA receives the EI application forms from IRS and codes them. Each proc-

essing step contributes to the lag. Because the cost and processing difficulties preclude correcting for this frame error, the Census Bureau accounts for new units in its estimates by an imputation technique. The overall objective is to correct errors which can be corrected within resource limitations and thereby keep coverage error as low as is feasible.

Defective Processes Associated with the Selected Sample--Coverage errors in which the selected sample does not correctly represent the frame may be the result of selected cases being inadvertently dropped from the sample or non selected cases being added to the sample erroneously. Also, errors may be made in selecting the sample. Errors of this type are likely to occur when the sample is determined by interviewers in the field. In business area samples where the sampling units are geographic land segments, failure to properly identify the population units (business establishments of a particular type) is a common form of coverage error. Such errors may result from inadequate definitions or inadequately specified field or office procedures, outdated or otherwise incorrect maps of selected area sample units, or misapplication of the sampling or canvassing rules by the interviewer. Failure to sample from an updated frame on a timely basis also results in a sample that is not representative of the frame, and hence of the sampled population. For other papers which discuss coverage concepts and issues, see Garrett, et al. (1986) and United Nations (1982).

It is worth noting here that even where coverage of a total population is fairly good, serious problems may exist for certain subpopulations. For example, national estimates might be good, while estimates covering smaller geographic areas may be inadequate because of defective geographic coding at the lower (state, county, etc.) level.

Specific Error Sources

As we have seen, errors of undercoverage or overcoverage can be the result of defective frames or of faulty sampling processes. Moreover, the same sources of error can affect both the frame and the selected sample and can lead to either undercoverage or overcoverage. Following are some specific sources of coverage error that are observable and measurable:

Coding Errors--Miscoding of industry or Standard Industrial Classification (SIC) coding, geographic coding, size coding, or company structure assignment results in frame errors. Such errors lead either to undercoverage or overcoverage depending on whether the correct units are excluded from the frame or incorrect units included in the frame. Including out-of-scope units (units which should not be included in the sampling frame based on the nature of their business or industrial activity) in the frame results from errors in industry coding and causes overcoverage. By the same token, the exclusion of units of the proper industry results in undercoverage. Similarly, if address, geographic codes, size, or any other attribute is a determinant for the sampling frame, errors in coding will cause overcoverage or undercoverage of the frame.

Two prevalent forms of miscoding are (1) completely unclassified units (especially for SIC) and (2) units which do not have sufficient coding detail for survey purposes. Unclassified units lead to undercoverage since units belonging in the frame cannot be identified. Insufficient coding detail--for example, when four-digit SIC detail is needed and only two- or three-digit detail is available--can lead to either undercoverage or overcoverage for surveys requiring finer levels of industry coding.

Some causes of miscoding are (1) inadequate information on which to base a code; (2) poorly trained coders; and (3) faulty procedures or processes, such as miskeying.

Errors of Timeliness--Errors of timeliness result when the frame or sample is not updated to the same reference period as that of the survey. For example, units no longer in business that remain in the frame or sample may lead to overcoverage. Lack of timely updating for new units may lead to undercoverage. For a list frame in which the presence of nonzero payroll is used as an indicator of "activeness," seasonal businesses may be erroneously deleted during their off season. Here again we see the dichotomous nature of coverage error: in surveys which are carried out over time, it is possible to have timely updating of the sampling frame, but unless the sample, in turn, is updated to reflect these changes, significant coverage error can result. In some survey designs it is impossible to completely eliminate coverage error due to the timing of frame or sample updates. This is especially true for list sample designs. However, use of an area sample to supplement the list sample, such as the Census Bureau uses in its Monthly Retail Trade Survey (MRTS), can theoretically reduce coverage error due to timing to zero.

Structural, organizational, or activity changes not reflected in the frame or sample may occur because of the lack of timeliness in updating. Often SIC changes occur which are not reflected in the frame or sample. Similarly, failure to update for other characteristic changes, such as company reorganizations, acquisitions, and divestments or mergers, results in coverage error.

Duplication Errors--Duplicate units on a frame can occur when, for example, a partnership business appears twice, once under each of the partners' identifiers, or when the predecessor and successor establishments both show up as active on the frame, as in the case of a business takeover. This same predecessor/successor situation can affect the sample if one of the units involved is a selected sampling unit. In addition, both a parent firm and its subsidiary could appear as separate sampling units on a frame if the association were not indicated. This would lead to overcoverage if a parent firm and all its subsidiaries are intended to be one sampling unit. Thus, processing or procedural errors can result in duplication error.

Duplication error may also occur when the sampling frame is composed of various lists, which must then be unduplicated. Any error in

this process can result in duplicate units being overlooked. This is often a problem where the primary identifiers on the component lists either don't match or are incomplete. Duplication problems also show up in dual frame surveys. For example, in the Census Bureau's Monthly Retail Trade Survey (MRTS), business establishments interviewed by personal enumeration in the area sample must be unduplicated from the list sample frame. When the employer identification (EI) number, which is the primary identifier, is incorrect or missing, the potential for duplication error is particularly great. Here again, while duplicate units cause overcoverage, problems in proper unduplication can also result in a case being incorrectly deleted.

Deficiencies in administrative record systems, censuses, or surveys on which the frame is based--Lack of or delays in reporting in the administrative systems, censuses, or surveys can cause coverage error. For example, although firms are asked to submit a separate report form for each of their establishments in the economic censuses of the Census Bureau, some firms invariably provide combined reports on one form. This results in both a deficiency in the frame of multiunit establishments and also in an undercount of the number of business establishments.

Nonlocatable units--Sometimes units selected into the sample are not contacted because they cannot be found. In area sample surveys, for example, certain types of businesses, such as service nonemployer establishments may not be locatable. Noncontact can also occur where street addresses (for personal interview surveys) or mailing addresses are erroneous or incomplete.

Interviewer errors--Errors made by an interviewer in the field can result in the sample being improperly identified. Interviewer "curbstoning" (that is, the interviewer filling out the survey forms without ever properly identifying the establishment or conducting the requisite interviews) and careless canvassing can also lead to an improperly selected sample, loss of population units, or inclusion of erroneous units.

Processing errors--Computer programming errors can cause a portion of the selected sample to be omitted from the survey or can result in a deficient frame from which to draw the sample. Units not included due to the processing error can also result from poor field procedures or inadequate or incorrect sample maps or materials. Improper identification of the sample at the central sampling facility due to computer or procedural problems can also result in undercoverage. Processing errors (including errors in drawing the sample at the central sample facility) can lead either to undercoverage or overcoverage.

III. Control of Coverage Error

Coverage error can be controlled by many different means. One principle often followed is to identify those areas where coverage error is most serious and assign resources to reduce the error there. Some specific and frequently used techniques which reduce miscoding, lack of timeliness, duplication of units, omission of units, and other errors resulting in incorrect coverage of the sampled population follow:

Sampling from multiple frames--Using an area sample to supplement and complete coverage for a list sample is sometimes necessary to obtain complete coverage of the sampled population.

Integration of multiple lists for frame development--Integrating and unduplicating several lists to construct a single frame is frequently done since most lists are composites of various sources.

Conducting special frame improvement surveys--The Company Organization Survey and SIC classification card mailings for the Census Bureau's Standard Statistical Establishment List (SSEL) are examples of these types of surveys. The economic censuses themselves constitute a frame improvement mechanism for all surveys drawn subsequently from the SSEL.

Use of two-phase sampling--This is done in the Census Bureau's business birth sampling program. A first-phase sample is selected based on SIC (including unclassified or insufficiently classified units) and payroll or employment size. A survey is conducted on this sample to produce better coding and to obtain sales data which are used as the measure of size for second-phase sampling.

Updating for births--Timely updating of the frame and sample for births and deaths.

Updating for structural changes--Timely updating of the frame and sample for structural and organization changes of the sampling units.

Sample validation--Producing a proof of sample tabulation whereby sample estimates are compared to universe totals for the same characteristic. This provides verification that the sample properly represents the frame.

Enlarging the scope of the survey--Often, in order to capture all of the units relevant to the survey, it is necessary to include possible or marginally possible units. During editing, the out-of-scope units can be dropped. Care must be taken to properly drop all the out-of-scope units so that overcoverage does not occur.

Using independent control counts--These counts are often needed to verify the correctness or completeness of the frame. The source of the counts could come from those for the frame for an earlier period as well as other sources.

Internal consistency checks for frame content--This involves performing internal consistency checks on the frame data fields, especially in record identification fields and fields which determine whether the unit is in or out of scope.

Internal consistency checks for duplicate records--This procedure involves performing internal consistency checks to identify duplicate records on the frame.

Include as inscope units with out-of-scope address, geography, industry, size--The practice of considering as inscope units those which are truly out of scope due to updates or changes in address, geographic, industry or size code is sometimes used in an effort to represent true inscope units which are not picked up because they are thought to be out of scope. This amounts to adjusting for coverage error.

Include units closed for the season--Retaining units closed for a season rather than dropping them and losing their contribution when they become active again is usually necessary to maintain a frame because of the lack of timeliness in reinstating the units.

Having correct, clear, and manageable sample control and frame maintenance procedures--All aspects of sample control and frame construction and maintenance must be well thought out and clearly specified.

Setting up adequate checks on processing--This is necessary to ensure correct processing of all types: interviewer, clerical, and computer.

Improving field materials--Improving field procedures and materials, such as addresses, maps, and other interviewer materials helps to reduce coverage error.

Interviewer selection and training--Carefully selecting and training interviewers and coders can have a substantial impact on reducing coverage error. This includes having well-trained supervisors oversee the survey operations.

Instituting a public relations campaign--This involves notifying the survey population of the survey or census in advance in an attempt to elicit their participation.

Reinterviewing procedures--These serve as a quality check on coverage error, especially for area sample surveys.

For an example of the procedures which are followed for maintaining frame and sample coverage for a large, ongoing retail trade survey, see Konschnik, et al. (1985).

IV. Measurement of Coverage Error

The measurement of coverage error is necessary in surveys if one is to have some idea of its extent as well as to identify sources most in need of improvement. While the focus of coverage is on the inclusion or exclusion of the proper sampling units in the frame and sample, the measurement of coverage error frequently centers on its effects on the published estimates of the survey. For example, it may be determined that a published estimate for retail sales of establishments in a certain SIC failed to include estimates for a signficant number of nonemployer establishments, but that including these nonemployers would only very slightly influence the survey results. The measure of undercoverage would be deemed small despite the number of sampling units excluded.

Indirect Techniques

Coverage error can often be ascertained by comparing current survey data with results from earlier surveys or from external sources. Coverage error may be indicated if the existing sample shows certain changes at a significantly higher or lower rate than the comparative data. Such measures as the birth rate, out-of-business rate, out-of-scope rate, unclassified rate, miscoded rate, duplication rate, and sample attrition rate can all be used to identify and measure coverage error.

Birth rate--Birth rates may be reviewed, comparing one period to another in order to indirectly measure coverage error.

Out-of-business rate--The rate at which frame or sample units go out of business, when compared to other measures or other time periods, provides a useful coverage error measurement.

Unclassified rate--A component of coverage error can be estimated by looking at the rate of unclassified units. These when combined with studies of the correct classification of this group provide a measurement of undercoverage.

Misclassified rate--A look at this rate and related studies can provide measurements of the extent of coverage error at all levels of survey tabulation.

Duplication rate--Determination of the number of repeated or duplicated units in a frame or sample gives useful information on coverage problems.

Sample attrition rate--The sample attrition rates, or the rates at which the units in the sample go out of business, when contrasted to birth rates and independently identified out-of-business rates, provide indications of the extent of coverage error.

Direct Techniques

Direct techniques for measuring coverage error usually entail carefully planned and executed survey procedures designed to provide a reliable estimate of coverage error. The following are examples of these direct techniques:

Post-enumeration surveys--Used here, this is synonomous with a post-audit whereby more extensive methods and procedures are used after the conduct of a survey or census in order to identify and determine the effect of coverage errors and other nonsampling errors.

Matching known population units against frame units--Checking known population units against the frame provides some indication of the quality of coverage. However, a carefully drawn sample of known units is required before accurate estimates of coverage error can be provided.

Checking the frame against alternative lists--While the selected frame may be the best available list for the survey, checks can be made against other lists (either of greater or lesser quality) to measure coverage error.

Comparing other survey or census data or independent aggregates--Independent aggregate estimates and tabulations covering the same characteristics for all or a part of the population provide a source of comparison for identifying and measuring coverage error.

Rechecking interviewers' field work--Independent rechecks of a sample of inteviewers' work are an excellent way of identifying and measuring coverage error.

Studying components of the frame--This includes assessing the various classifications of units which make up the list.

V. Summary Profile

This section presents some general results compiled from a questionnaire on survey practices which covered 55 major establishment surveys of Federal agencies. For the identification of these surveys, see Office of Management and

Budget (1988). Figures 1 and 2 give a summary of control procedures used in descending order of extent of use. Figures 3 and 4 characterize measurements of coverage error taken for these surveys, in descending order of extent of use, for indirect and direct measures. Note that although the "not applicable" category is included when determining descending order, it is not included in any textual references in this section.

The results in these graphs show that while the majority of these Federal surveys included provisions for controlling coverage error, the measurement of coverage error was less widespread. Moreover, where measurements were taken, only a small percentage was published. Thus, most measurements were for internal use to assess the adequacy of survey estimates.

The most prevalent form of coverage control (96 percent) involved updating the frame for structural changes such as SIC changes, company reorganizations, mergers, etc. Updating of the sample for births was the second most prevalent form of coverage control (87 percent). Other control techniques reported as being used on more than half the surveys were: internal consistency checks for duplicate records on the frame (73 percent); internal consistency checks for frame content (69 percent); including as inscope units with errors or changes in address, geography, industry, or size, rather than dropping them as out of scope (67 percent); sample validation, i.e., comparison of weighted-up sample units to universe totals (67 percent); and integration of multiple lists for frame development (66 percent). Other fairly common control techniques reported were the conducting of special frame improvement surveys (49 percent) and retaining units closed for the season (47 percent).

Typically, little use (9 percent) was reported of two-phase sampling for improving frames and samples although this method can prove beneficial in reducing the variance of estimates caused by frame problems. Also, on the low side in terms of relative use, only about 20 percent of the surveys reported sampling from multiple frames, such as using both a list and area sample.

When looking at the measurement of coverage error, out-of-business and out-of-scope rates are most common with 67 percent and 62 percent of the survey population reported as having these measurements taken, respectively. These measurements also have the highest rate of being published at 13 percent and 9 percent, respectively. A majority (60 percent) of the surveys reported comparing estimates produced in the surveys with estimates based on other independent sources. Measuring the misclassified rate (44 percent), matching known population units against frame units (47 percent), measuring the unclassified rates (38 percent), and measuring the sample attrition rates (36 percent) were also somewhat common.

Least common were the conducting of post-enumeration surveys (20 percent) presumably because of the cost and resources involved; and rechecks on interviewers' listings (16 percent), primarily due to the nonapplicability of interviewers' involvement in listing for many of the surveys.

Figure 1

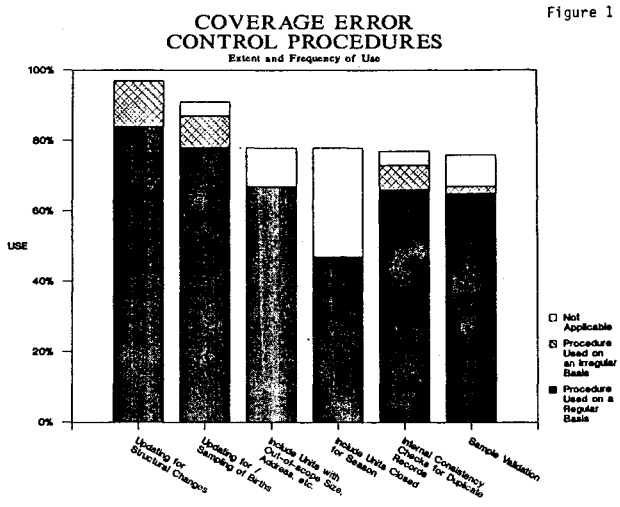


Figure 2

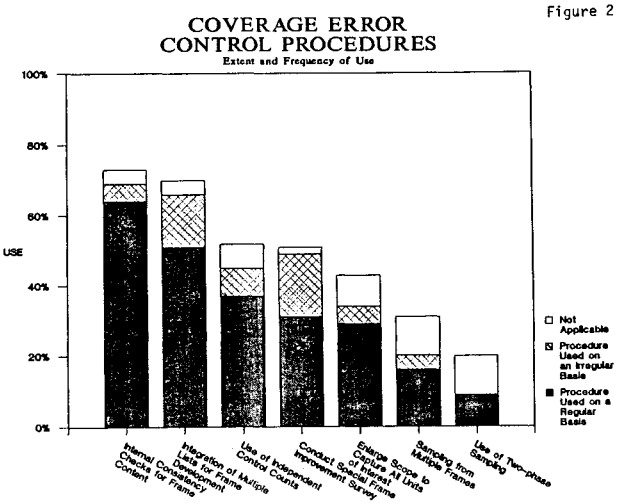


Figure 3

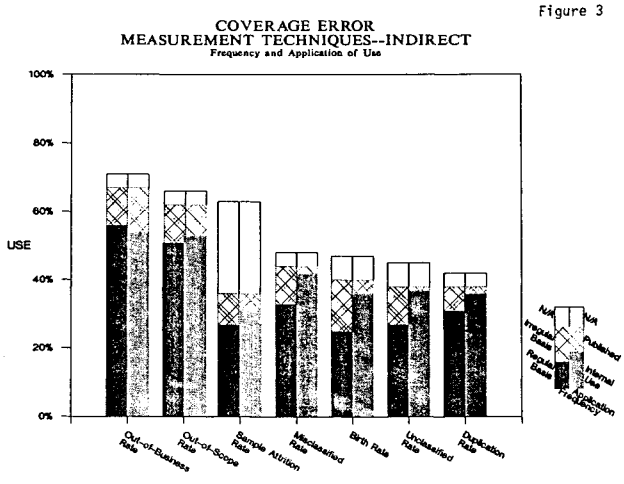
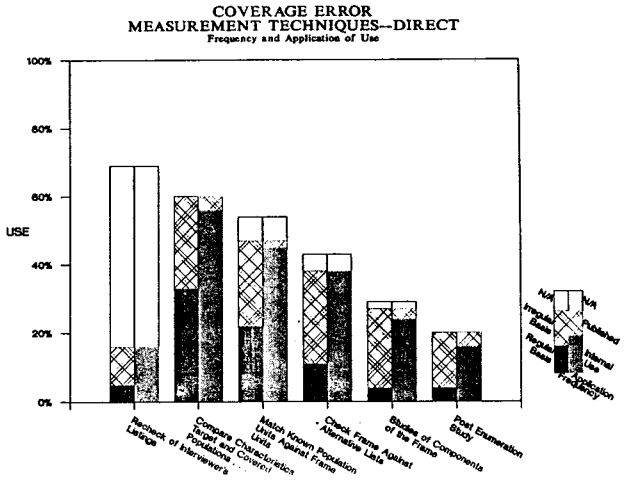


Figure 4



References

Cochran, William G. (1977), Sampling Techniques, 3rd ed., New York: John Wiley and Sons.

Garrett, J., Hogan, H., and Pautler, C. (1986), "Coverage Concepts and Issues in Data Collection and Data Presentation," Proceedings of the Second Annual Research Conference, Washington, D.C.: Bureau of the Census, pp. 329-334.

Konschnik, C., Monsour, N. and Detlefsen, R. (1985), "Constructing and Maintaining Frames and Samples for Business Surveys," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 113-112.

Office of Federal Statistical Policy and Standards (1978), Statistical Policy Working Paper 4, Washington, D.C.: Department of Commerce.

Office of Management and Budget (1988), Quality in Establishment Surveys, Statistical Policy Working Paper 15, Springfield, Va.: National Technical Information Service (PB 88-23294).

[1] This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.