Guide for a Synar Sampling Frame Coverage Study

January 2006

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Substance Abuse and Mental Health Services Administration
Center for Substance Abuse Prevention
www.samhsa.gov

Guide for a Synar Sampling Frame Coverage Study

January 2006

U.S. Department of Health and Human Services
Substance Abuse and Mental Health Services Administration
Center for Substance Abuse Prevention

Contents

Guide for a Synar Sampling Frame Coverage Study

Introduction

Under the Substance Abuse Prevention and Treatment (SAPT) block grant Synar requirement, States must conduct annual unannounced random inspections of tobacco retailers to determine the compliance rate with laws prohibiting the sale of tobacco products to persons under the age of 18. To fulfill the Synar requirement, each year the States conduct a Synar survey, which involves choosing a random sample of tobacco retail outlets, inspecting the sampled outlets, and then estimating the overall State retailer violation rate (RVR) based on the results of the inspections the sampled outlets.

Nearly every State chooses the Synar sample from some type of list of tobacco outlets. This list is referred to as a "list sampling frame" or, more simply, a "list frame." The list frame serves as the foundation for the Synar survey; therefore, its quality is very important. A coverage study needs to be conducted to measure one very important aspect of the quality of a list frame. The purpose of this guidance document is to describe what a coverage study is and why it is needed, and to provide guidance on how it can be conducted.

Characteristics of a List Frame

A State's tobacco list frame can be compiled from various sources. Sources include lists of licensed tobacco retailers (in States that require such licensing), other administrative lists, commercially available business lists, and combinations of these.

Two aspects of the list frame are important to have a good quality list frame: coverage and accuracy.

> **Coverage.** Coverage indicates how completely the list contains (*covers*) all of the eligible outlets in the State for the Synar survey. An eligible outlet is a retailer that sells tobacco and is accessible to minors. The coverage rate is the percentage of all eligible outlets in the State that actually appear on the list frame. If 80 percent of the eligible outlets appear on the list frame, the coverage rate is 80 percent. This means that 20 percent of the eligible outlets are missing from the list frame. The coverage rate can be estimated through a coverage study, which is a special type of survey conducted to measure the coverage or incompleteness of the list frame.
>
> **Accuracy.** The accuracy of a list frame reflects the quality of the key information about the outlets contained on the list frame, such as address accuracy and outlet eligibility contained on the list frame. Less than 100 percent accuracy means there are outlets in the list frame which may not sell tobacco, be accessible to youth or have a correct address.

Inaccurate information about outlets that are contained on the frame can be properly accounted for through the survey results. However, a low coverage list frame may bias the estimate of the retailer violation rate, and it is not possible to adjust for this effect after the survey has taken place. Thus, the Substance Abuse and Mental Health Services Administration (SAMHSA) is far more concerned about the quality of the coverage of the list frame than its accuracy.

Why Coverage Matters

A low coverage rate is a potential source of bias[1], since the unlisted outlets might be different from those on the list with respect to their likelihood of selling tobacco to youth.

To get some idea about how a low coverage rate can bias RVR estimates, look at a hypothetical example. Suppose that a list frame includes 70 percent of the total number of eligible tobacco outlets (a coverage rate of 70 percent). This would mean that 30 percent of the total eligible outlets were missing from the list frame. Further suppose that the violation rate for the outlets contained on the list frame is 20 percent but the outlets that are not on the list frame have a violation rate of 35 percent. Thus, the overall violation rate for all outlets in the State is 24.5 percent.[2] However, the Synar survey would be based only on the 70 percent of outlets covered by the list frame and therefore the expected value of an RVR estimate would be 20 percent. The difference between the expected value of 20 percent and the true overall RVR of 24.5 percent results in a bias of 4.5 percentage point due to the low coverage rate of the frame. (See Appendix A for further illustration.)

SAMHSA Coverage Requirement

The overall quality of the list frame—as reflected by its coverage and accuracy—is vital for conducting the Synar survey successfully. In 2000, the Government Accountability Office (GAO) examined how SAMHSA has implemented the Synar requirements. The GAO found that States were using inaccurate and incomplete lists to select random samples of tobacco outlets to inspect. The GAO made a recommendation to SAMHSA to work more closely with the States to increase the completeness (coverage) and accuracy of tobacco retailer lists used for the Synar surveys (GAO, 2001[3]).

In recent years, most States using a list frame have improved the accuracy of the lists used to draw the Synar sample by using various methods to clean their lists, including contacting listed outlets, confirming their location, and determining whether they sell tobacco, etc. However, many States have not regularly evaluated and reported on their frames' coverage rate.

[1] Bias refers to the difference between the "true" RVR and the expected value of an RVR estimate.

[2] Calculated as $20\cdot 0.7 + 35\cdot 0.3$.

[3] U.S. General Accounting Office (2001). Synar Amendment Implementation: Quality of State Data on Reducing Youth Access (GAO-02-74). U.S. General Accounting Office.

SAMHSA Requirement

Synar sampling frames must conform to the minimum coverage requirement of 80 percent

A well-maintained list frame should have a high coverage rate. Therefore, SAMHSA recommends a 90 percent coverage rate. However, recognizing that it is difficult to achieve this level of coverage by a list frame created from commercial sources, SAMHSA requires a coverage rate of at least 80 percent for the Synar survey.

SAMHSA Requirement

A frame coverage study should be repeated every three years. If a State provides compelling evidence to SAMHSA that the coverage rate is = 90 percent and the list frame used is comprehensive and stable, the State may be allowed to implement the coverage study at five year intervals.

When a sampling frame is well established and the procedures for developing and updating the list are consistently followed year after year, its coverage does not change much because it is dependent mainly on these maintenance procedures. However, a list frame also is susceptible to potentially unnoticed effects of changing regulatory, economic, or social factors that affect the emergence and disappearance of outlets in ways that the established procedures were not designed to capture. Sampling frame coverage deterioration can result when such small changes accumulate over time. Thus, a coverage study should be repeated every three years unless there is strong evidence to suggest that this requirement should be waived.

How to Measure Coverage

The true coverage of a sampling frame is the number of eligible outlets in the State's sampling frame divided by the number of *all* eligible outlets in the State. To get a perfect measure of the true coverage, all outlets in the State would need to be identified and their eligibility assessed. To do this would involve physical canvassing of the entire State to look for retail tobacco outlets, in the same way that the population census canvasses every street and structure. This is termed a census approach.

However, a sample survey that costs a fraction of what it would cost to implement the census approach can be used to produce a reasonably reliable estimate of the coverage. Such a survey still relies on physical field canvassing, but limits the canvassing to specific sampled areas. To conduct such a formal coverage study based on a sample survey, one has to use an area frame, sound survey sampling principles, and practical survey field procedures.

The following guidelines are provided to assist the States in designing and implementing a coverage study. These are only general concepts; the specific procedures require considerably more detail and will vary depending on each State's circumstances, resources, and expertise. A State may propose an alternative method; however, the State must show that its coverage study meets SAMHSA's requirements and the methodology must be approved by CSAP before conducting the coverage survey.

Sample Design for a Coverage Study

An area frame is a list of easily identifiable and manageable geographically defined areas that represent the full geographic extent of the State, without gaps or overlaps. SAMHSA recommends using census tracts[4] as the sampling area for a coverage study. Alternatives include ZIP code areas and census blocks. However, some ZIP code areas are too big in terms of the number of outlets, especially in urban areas; some are too small and without any outlets. Large size variation is another serious problem of ZIP code areas. Census blocks are usually too small - many of them may not have any outlets at all. Therefore, if these alternative areas are used, combining small ones and splitting large ones should be done to create an adequate area frame for a coverage study. As a rule of thumb, the average area size should be such that it is expected to contain not less than 7 but not greater than 20 outlets on average. Please note that this rule of thumb is a rough guide, not a precise rule. If a State can use a different average size outside of this range more effectively, this rule does not preclude that. Furthermore, it is stated in terms of average size. For example, if a State creates an area frame for a coverage study that has an average size of 10, some areas will have more than 10 outlets and some have less; some even may not have any at all. However, it is desirable to create areas of similar size as much as possible. Among readily identifiable geographic areas, census tracts have the most desirable features in this sense. It should also be noted that the area size is an approximation and its actual size will not be known until it is thoroughly canvassed and all eligible outlets in it are correctly counted.

Once the sampling areas have been determined and defined, the list of all the sampling areas forms the area frame for a coverage study. To select a sample of areas from the area frame, a State must decide how many areas to select and how to select them. This means that the State must choose a sample design, preferably a simple design such as a simple random sample (SRS) of areas, if operationally feasible. States may use a more complex sample design with substantial benefit, depending on their specific situations. For example, if coverage rates and/or canvassing cost factors are known (or can be predicted reasonably) to be substantially different for certain types of areas (such as urban versus rural areas), a stratified design can reduce survey cost and improve efficiency. (See Appendix C for detailed discussion on a

[4] Census tracts are small, relatively permanent statistical subdivisions of a county used by the Bureau of the Census for conducting a decennial census. Census tract boundaries normally follow visible features, but may follow governmental unit boundaries and other non-visible features in some instances; they always nest within one county. There are about 65,443 census tracts in United States, excluding the territories. Based on the 2000 Census, the average human population size of the census tract is about 4,300 inhabitants (this roughly translates into 7 to 10 outlets in most States). The Census Bureau website provides all details, including their maps. http://www.census.gov/geo/www/tiger. The census tracks can be downloaded from this website free of charge.

stratified design). For a large State, it would be beneficial to use a multi-stage design. (Appendix D provides an outline of a multi-stage design.)

SAMHSA recommends a sample of areas that are expected to contain a total of 130 to 200 outlets, depending on the average area size. If the areas chosen contain 7 to 20 outlets on average, this means approximately 10 to 19 areas will need to be selected. (See Appendix B for details.)

It is important to select a coverage study sample and implement the sample as close as possible to the time when the Synar survey is conducted because the outlet population keeps changing over time.

Coverage Survey Field Procedures

The following field procedures are provided to guide implementation of the study sample. These are only general concepts; the specific procedures require considerably more detail and will vary depending on each State's circumstances, resources, and expertise.

Within each selected area, field staff members need to canvass the area completely and make an exhaustive list of all tobacco establishments that are accessible by minors. It is extremely important that this list be thorough and detailed. Otherwise, the result will be a faulty coverage estimate. It is also important that this listing be carried out without the assistance of the existing list used for the Synar survey.

The following general field procedures are provided to guide implementation of the survey for the SRS area approach:

1. Obtain accurate maps that include the boundaries and streets of the sampled geographic areas.

2. For each sample area, map out a route for the field worker to follow so that he or she covers every street or other location (e.g., parks and recreational facilities) necessary to ensure finding all outlets operating in the sampled area.

3. Create canvassing sheets on which the field staff will record the existence and necessary identifying information for each eligible outlet encountered during the canvassing. The sheets should include places for the canvassers to record the name and address/location of each outlet and any other relevant information (e.g., telephone number if possible) that will support the process of determining whether the outlet appears on the list frame. The more complete such information is, the less time will be spent for cross-matching with the list frame and the fewer false negatives there will be.

4. The field worker should follow the route and list on the canvassing sheet all eligible outlets in each sample area. The field worker should carefully check the eligibility of each identified outlet by determining whether tobacco products are

sold and whether the outlet is accessible to youth under the legal age (e.g., 18). In high density areas with large buildings or other types of complexes, it may be necessary to check throughout the entire complex or speak with the management to identify tobacco outlets operating within the complex. This could also apply to malls, recreational areas, campuses, etc.

5. When the canvassing is completed, the field worker should make a final check that all routes and locations on the map have been covered and all necessary information about the identified outlets has been gathered and prepared for transmission, and then transmit the field forms to the central office for use in measuring the coverage of the Synar survey list frame.

Estimating Frame Coverage from the Coverage Survey

Once the list of outlets developed by the field canvassing is finalized, central office staff need to compare it with the list sampling frame that is being used for the Synar survey. If an outlet found in the field cannot find its match in the list frame, it is treated as a missing outlet.

The general matching process is typically done in two steps. The first step is to identify each outlet on the canvassing list that is clearly and easily confirmed as one appearing on the list frame. Then the more difficult step is to assess whether any remaining outlets on the canvassing list truly do not appear on the frame. In some cases discrepancies may be found between the Synar list frame and the coverage survey field results due to one or more of the following:

- A new owner changes the name of the outlet
- A street is known by several names (e.g., both a name and route number or it changes names along its length)
- The address can be defined by either a street number or a location (e.g., “2415 State Highway” is the same geographical location as “Mountainview Mall”)
- There are several entrances to the outlet on different streets
- Recording error (transposed numbers, misspellings, partial omissions)

Any unmatched case on the canvassing list should be carefully examined to confirm that there is indeed no match in the list frame. When the name and address information on either list is ambiguous or incomplete, it also may be necessary to check back with the field workers or contact the relevant outlets to clarify and confirm the situation.

After all canvassed outlets have been checked against the list frame, any outlet that is not found in the list frame is classified as missing from the list frame. Count how many outlets are finally matched.

Then the coverage rate is calculated as the ratio of the total number (b) of matched outlets on the frame divided by the total number (n) of outlets found by the coverage survey (i.e., coverage rate = $100 \times b / n$). This simple formula is based on the assumption that the coverage sample was selected by an equal probability sampling method such as SRS of areas. Otherwise, the sampling weights should be used to obtain a valid estimate.

The information about the coverage study and its result are to be reported by answering Question 8 and filling out Appendix D of the Annual Synar Report.

Corrective Actions for Low Coverage

A list frame is the most convenient frame to use for the Synar survey. However, if the list frame has a coverage rate below 80 percent, then the State should either (1) improve the frame to reach the minimum coverage prior to drawing the Synar sample, or (2) use an area frame for the Synar survey itself. See the Synar sample design guidance[5] for further instructions on using a pure area frame or a list-assisted area frame. *If an area frame is used for the Synar survey, a coverage study is not needed as long as the area frame covers the whole State.*

If the State continues to use a list frame, new outlets found during the coverage survey should be added to the list frame. There may be some hesitance in doing so because of concern that partial improvement using the coverage study outcomes may bias the Synar survey. However, the Synar surveys use a random sampling method that eliminates such bias. Therefore, any improvement to the frame is beneficial.

Technical Assistance

This guide is one tool that CSAP provides to help States address this important issue. Please contact your assigned CSAP State Project Officer for further technical assistance if deemed necessary in conducting a coverage survey or to improve your frame's coverage.

[5] SAMHSA/CSAP, *Synar Regulation: Sample Design Guidance*, April or May 2003.

Appendix A

Impact of Bias on the Confidence Interval

The impact of the bias of the example on page 2 on a statistic such as the 95 percent confidence interval of the RVR estimate is shown to illustrate the seriousness of a large bias; the actual confidence level is compared with the claimed confidence.[6] If the Synar survey produces an RVR estimate of R and a standard error estimate of S, then the 95 percent right-sided confidence interval is given as [0, A], where $A = R + 1.645 \times S$. This interval is supposed to contain the true violation rate 95 percent of the time when the survey is repeated many times. Note that A, R, and S are random variables before we get their specific estimates from a particular sample. The expected values of these random variables are the averages of such estimates obtained from all possible samples under the same sample design. This repeated sampling generates a sampling distribution, and the confidence statement that the 95 percent confidence interval contains the true RVR with 95 percent of probability (or 95 percent of the time) is made under this sampling distribution.

It is misleading to assume that the true RVR falls within the 95 percent confidence interval as calculated from the Synar survey with the stated confidence if the actual confidence is very different from the claimed confidence of 95 percent. This can happen if the RVR estimate has a large bias. To see this more clearly, suppose that the true standard error of the RVR estimate is 1.82 percent (i.e., the expected value of S is 1.82), the maximum allowable value for the Synar survey. The right limit of the confidence interval is on average 23 percent (i.e., the expected value of A equals 23 percent). In this case, the actual confidence is only 20.5 percent, not 95 percent.[7] This means that the confidence interval will contain the true RVR of 24.5 percent only 20.5 percent of the time in repeated sampling. If the standard error is smaller than 1.82 percent, the actual confidence will be even lower than 20.5 percent, and so the discrepancy between the actual confidence level and the claimed 95% confidence level will be even more dramatic. A discrepancy between claimed and actual confidence levels of 5 percentage point or more is considered serious, and so the 74.5 percentage point discrepancy (95 percent minus 20.5 percent) is unacceptable. Sampling literature advises that the ratio of the bias to the standard error should be less than 0.1 to have a minimal impact on the confidence level due to bias.[8] For our example, this would require that the bias should not be larger than 0.182 percentage point. However, the example has a bias of 4.5 percentage point and its bias-to-standard-error ratio is 2.47 (= 4.5/1.82). This is 24.7 times larger than the allowable ratio of 0.1! This is a dramatic example but one should be aware that a seemingly negligible bias of 0.182 percentage point can affect unacceptably the actual confidence level.

[6] It is claimed by a 95 percent confidence interval that the true RVR is contained in the interval with 95 percent confidence, which means that the interval contains the true RVR 95 percent of time in repeated sampling. If the RVR estimate is biased, the actual confidence is lower than the claimed confidence.

[7] The probability of 20.5 percent is obtained as probability of $P(24.5 \leq A)$, where A is a normal random variable with mean 23 and standard error of 1.82.

[8] William G. Cochran (1977), *Sampling Techniques*, 3rd ed., New York: John Wiley & Sons, p. 14.

Appendix B

Sample Size Determination for a Coverage Study

The sample size of a coverage study is determined by making a compromise between the desire to obtain a more precise estimate of the coverage rate (which calls for a larger sample size) and the desire to limit the study cost (which calls for a smaller sample size). As a reasonable compromise, SAMHSA recommends a sample of areas that are expected to contain 130 to 200 outlets. For a given sample size of outlets, the precision of the coverage estimate depends on the true (unknown) coverage rate. If the coverage rate is 80 percent, this sample size amounts to roughly a standard error of 4 percentage points, provided the average area size falls in the recommended range (i.e., 7 to 20 outlets).

Determining the targeted sample size in terms of outlets for a coverage study is not simple because we do not know exactly how many outlets will actually be in the sample areas. This will be known only when the field work is completed. However, for the purpose of the coverage survey, reasonably good approximate values are sufficient. When the area frame for the coverage study is ready, the average area size can be approximated by dividing the total number of outlets in the list frame for the Synar survey by the total number of areas in the area frame for the coverage study. Let this average size be denoted by *m*. It should not be too small for cost-effective reasons or too large for design-effective reasons; the rule of thumb that the average area size should be between 7 and 20 outlets is based on this consideration. However, this rule is rough guide not a law. A State may use a different average size outside of this range if deemed more effective. After determining *m*, use the following chart to determine the number of areas to be selected (denoted as *k*) and the overall outlet sample size (denoted as *n*).

Table 1: Sample Size Determination Chart

Average area size (*m*)	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Number of areas (*k*)	19	17	16	15	14	14	13	13	12	12	11	11	10	10
Total outlet sample size (*n*)	133	136	144	150	154	168	169	182	180	192	187	198	190	200

The 14 pairs of (*m*, *n*) have roughly the same precision in terms of the coverage rate estimate that will result from the survey.[9] Note that when the *m* is smaller, the *n* tends to be smaller as

[9] Assuming a moderate intra-class correlation of 0.05, the effective sample size is about 100, and so the standard error of a coverage rate estimate is expected to be 4 percent if the true coverage rate is 80 percent. If the true coverage rate is higher than 80 percent, then the precision level gets higher too (i.e., the standard error gets smaller). The effective sample size is the sample size under a simple random sample of outlets that is equivalent, in terms of the precision, to the actual sample size for the cluster sample design employed. Because of the design effect, the effective sample size of the cluster sample will likely be smaller than its actual sample size (also see following footnote).

well. This is related to the design effect[10] of the sample design and is due to the fact that the smaller the *m*, the smaller the design effect becomes.

The sample designer should keep in mind that the average area size and total outlet sample size are the starting targets. Depending on which areas are actually selected for the sample, the average and total will likely be different from the targets. However, the specific number of outlets in each sampled area will be distributed around the average.

[10] Design effect is the ratio of variance of a cluster sample design (or any other design) to that of a simple random sample of outlets with the same outlet sample size. A design effect grater than 1 means that the cluster design is less efficient in terms of variance than the simple random sample of outlets (this is usually the case for a cluster design). The effective sample size is obtained by dividing the outlet sample size (*n*) by the design effect. For example, if n = 150 and the design effect is 1.45, then the effective sample size is 103. In other words, the actual sample size of 150 outlets in the area sample would be needed to achieve the same effect on statistical precision as a smaller sample of 103 would achieve with a simple random sample.

Appendix C

Stratified Design for the Coverage Study

If stratification is desired, it is recommended to stratify the sampling areas by rural and urban, as surveying rural areas is generally more costly, due to remoteness, sparseness, and distance factors. Stratification makes it possible to treat rural and urban areas differently in sampling, especially by undersampling rural areas to reduce the costs and effort associated with rural areas. To avoid making the sample design any more complex, simple random sampling of areas within strata may be used.

To exploit the urban/rural stratification more fully, it is recommended to use "optimum allocation" of the areas rather than other simpler allocation. Optimum allocation prescribes allocation of the total area sample proportionately to the stratum number of outlets and the stratum standard error but inversely to the square root of stratum unit cost (the per-unit cost of conducting the coverage survey for each outlet found in that stratum). The unit cost does not need to in actual dollar figures but can be in relative term. To introduce the allocation formula, let us define some notation. In the following, all sizes (number of outlets) are assumed to be available from a list frame or some other sources. It is again reminded that any size in this discussion is approximate because the real size is unknown.

N : The total frame size

N_u : The urban stratum size

N_r : The rural stratum size

P_u : Coverage rate for the urban stratum

$Q_u := 1 - P_u$: Undercoverage rate for the urban stratum

P_r : Coverage rate for the rural stratum

$Q_r := 1 - P_r$: Undercoverage rate for the rural stratum

$S_u = \sqrt{P_u Q_u}$: The standard deviation in the urban stratum

$S_r = \sqrt{P_r Q_r}$: The standard deviation in the rural stratum

c_u : The unit cost to examine an outlet in the urban stratum

c_r : The unit cost to examine an outlet in the rural stratum

a : Cost ratio, that is, c_u / c_r

n : The total outlet sample size for the coverage study

n_u : The outlet sample size for the urban stratum

n_r : The outlet sample size for the rural stratum

Then the optimum allocation determines the stratum outlet sample sizes by:

$$n_u = n\frac{N_u S_u}{N_u S_u + N_r S_r \big/ \sqrt{a}}.$$

The overall sample size n is set to be a number between 130 and 200 depending on the average area size. To determine the overall sample size (*n*) and the number of areas to select (*k*), refer to Table 1 in Appendix B.

The knowledge of stratum level coverage rates and unit costs is required to use the above formula. Since they are unknown for the current study being planned, the figures for the last coverage study or reasonable guess values for both the coverage rates and unit costs may be used.

At the worst, we may assume that there is no difference between the urban and rural strata in terms of coverage rate and unit cost. Then the formula is reduced to the proportional allocation formula, namely:

$$n_u = n\frac{N_u}{N_u + N_r}.$$

However, in this case it defeats the very purpose of stratification, and so stratification is not necessary unless there are other reasons rather than study efficiency.

If only the cost differences are reflected, then it becomes:

$$n_u = n\frac{N_u}{N_u + N_r \big/ \sqrt{a}}.$$

The above formulae give the urban stratum outlet sample size. The rural stratum outlet sample size is then the balance of the total outlet sample size, that is, $n_r = n - n_u$.

After allocating the sample to the two strata, the number of areas to be selected from each stratum can be determined by dividing the stratum outlet sample size (n_u or n_r) by the average area size, *m*. Apply rounding to get a whole number of areas to be selected from the stratum. Let the numbers of areas to select be k_u and k_r from the urban and rural strata, respectively, and then they should sum to *k*, the total number of areas to be selected, that is, $k = k_u + k_r$. Because the stratum sample sizes of areas must be whole numbers, the actual outlet sample sizes would be generally different from the allocated sample sizes (n_u and n_r). The revised sample sizes will become $n'_u = k_u m$ for the urban stratum, $n'_r = k_r m$ for the rural stratum, and the total outlet sample size is given by $n' = n'_u + n'_r$ for the State. These outlet sample sizes (n', n'_r, and n'_u) refer to the expected or targeted number; the real sample sizes will depend on the actually selected areas that have varying sizes and the realized number of outlets found in the field through canvassing.

Coverage Estimation for the Stratified Design

Suppose that k_u areas were selected from K_u areas for the urban stratum, and k_r areas from K_r areas for the rural stratum. If simple random sampling of areas was used within strata, then the probability of area selection is k_u/K_u and the sampling weight is K_u/k_u for all sampled areas in the urban stratum. The probability and sampling weight for the rural stratum are similarly defined. The steps involved in the calculation of the coverage rate are as follows:

1. Count the number of outlets found from the field canvassing for which a match was found in the frame files for each sample area, and let it be denoted by b_i for i-th sample area.

2. Count the number of outlets found from the field canvassing for each sample area, and let it be denoted by n_i for i-th sample area

3. Determine the sampling weight (i.e., inverse of the selection probability) of each sample area, and let it be w_i for i-th sample area.

4. Assuming that all k selected areas were canvassed, the percent coverage rate (C) is estimated by

$$C = 100 \frac{\sum_{i=1}^{k} w_i b_i}{\sum_{i=1}^{k} w_i n_i} .$$

If not all selected areas were canvassed, this fact should be reflected in all computations (weight calculation and count of the canvassed areas).

Appendix D

Two -Stage Design for the Coverage Study

Although more complex, this design would be appealing to large States (with thousands of census tracts) for cost and operational reasons since it would allow efficient geographical clustering of the sample areas, rather than sampling many single census tracts scattered across the entire State. Clustering of the sample areas is less costly and allows for a more efficient canvassing of the areas.

Counties can serve as the first stage sampling unit and census tracts as the second stage sampling unit. Further, it would be more efficient to select the counties by the probability-proportional-to-size (PPS) sampling method, with the size measure of the county being defined by the number of census tracts. To get an equal probability sample, it is recommended to select an equal number of census tracts from each selected county. For example, if all together 16 census tracts are to be selected, first select 8 counties by the PPS methods, and then select 2 census tracts from each selected county. To determine the sample size, Table 1 in Appendix B can be used as a guide.

If desirable, urban/rural stratification can be added, and the result is then a stratified two-stage design. In this case, counties should be classified first as either rural or urban to create the strata, and then two-stage cluster sampling is carried out within each stratum. This complex design may offer the State very high operational and statistical efficiency. Due to its complexity, this option is not further explored in this guide. If a State wants to use this design but lacks the needed expertise, technical assistance can be obtained from CSAP.

A Life in the Community for Everyone

SAMHSA

Substance Abuse and Mental Health Services Administration

U.S. Department of Health and Human Services