PART B OF THE SUPPORTING STATEMENT (FOR STATISTICAL SURVEYS)

INTRODUCTION TO PART B

The Environmental Protection Agency (EPA) proposes to conduct the following type of statistical survey for the 2007 Drinking Water Infrastructure Needs Survey and Assessment (DWINSA). EPA proposes a mail assessment of community water systems (CWSs) serving populations of more than 3,300. EPA will send site visitors to collect data from CWSs serving 3,300 or fewer people. EPA is proposing the same methodology for collecting data for CWSs serving more than 3,300 people, as was used in the 2003 DWINSA. EPA is also proposing the same approach used in 1999 to collect data from CWSs serving 3,300 or fewer people.

This page intentionally left blank.

B.1 SURVEY OBJECTIVES, KEY VARIABLES AND OTHER PRELIMINARIES

B.1.a Survey Objectives

The primary objective of the 2007 DWINSA is to collect information from CWSs on the infrastructure they need to continue to provide safe drinking water to consumers. These data are used to produce a national estimate as well as state-specific estimates of water systems' 20-year need. EPA has established policies to ensure that the overarching goals of the survey are met. These polices are provided to the states and help EPA:

- Estimate the total national 20-year need.
- Estimate the total 20-year need for each participating state.
- Provide complete and accurate data to Congress.
- Provide a tool to fairly distribute DWSRF capitalization funds to states.
- Maintain the credibility of the DWINSA findings.

EPA proposes to collect information on the cost of systems' infrastructure needs; if cost data are not available from systems, EPA proposes to collect information that will enable the Agency to model costs. In the data collection instrument, the respondent will identify needs on a project-by-project basis and list the "type(s) of need" that the project will meet. The "types of need" includes raw water source, transmission, source water treatment, storage, distribution, pumping stations, and other needs.

EPA will use the information from the DWINSA to project capital investment requirements of drinking water systems. The information will be used to allot Drinking Water State Revolving Fund (DWSRF) monies among states and as part of an allotment formula for the American Indian and Alaska Native DWSRF set-aside program.

EPA is proposing the same methodology as used in previous DWINSAs. Two changes were made for the 2007 DWINSA from the approach used in 2003. Data will be collected from systems serving 3,300 or fewer people. Data was not collected from these systems in the 2003 DWINSA. Data will be collected from a random sample of systems serving 50,001 – 100,000 people. These systems were selected with certainty in 2003. The sampling design will be discussed in detail below.

B.1.b Key Variables

Several key variables are available from the Safe Drinking Water Information System (SDWIS). To ensure accuracy, the 2007 DWINSA will verify these data by asking respondents to confirm existing information (pre-printed on the data collection instrument), or correct it. These variables include population served, total design capacity, number of service connections, primary source of supply, ownership type (private or public), and whether the system purchases water from, or sells water to, another public water system (PWS).

Information on capital needs will be collected from respondents on a project-by-project basis. For each project, respondents will be asked to provide the following types of information: type of need; documentation of need and cost (if necessary); if the project is a new project or rehabilitation of existing infrastructure; if the project is needed now to protect public health or if it is needed over the next 20 years to continue to provide safe drinking water; the federal regulation or state requirement if the project is to meet a current regulation or state requirement; design capacity of source, storage, and treatment projects; cost of the project; and date of the cost estimate. For most of these variables, respondents will choose the appropriate "documentation," "type," or "regulation or requirement" from a lists of codes.

The principal variable of interest is total projected capital needed for each CWS in the 2007 DWINSA for the time period 2007 – 2026. The total capital need for all systems in each state (to be derived from the statistical sample of systems) is the key variable that decision-makers at EPA use to allocate funds to states based on need.

The method of data collection has been designed to minimize burden on respondents while ensuring that information is collected in a consistent manner. Collecting information on a project-by-project basis, for example, will be particularly helpful in reducing burden since most respondents develop Capital Improvement Plans (CIPs) in this manner.

Information on type of need will be used to disaggregate total capital needs for EPA's Report to Congress. Information on documentation of need will be used to verify the public health benefit of the need. Information on the date of the cost estimate will be used to provide a consistent basis for cost estimates across systems. Information on a regulation or requirement will be used to determine the reported project costs related to Federal regulations or state requirements.

If a system cannot provide cost estimates, additional data are necessary so that the Agency can impute costs. Each of these variables will be described in greater detail later in this document.

B.1.c Statistical Approach

The 2007 DWINSA is being designed to achieve a desired level of precision for state-level estimates of total capital needs for medium and large CWSs. It also is being designed to estimate the total capital needs of small systems for the nation as a whole. EPA proposes a survey of a statistical sample to estimate total capital needs for CWSs serving populations of more than 3,300. This statistical approach minimizes burden while achieving the desired level of precision.

Medium and Large CWSs

The 2007 DWINSA design divides CWSs serving populations of more than 3,300 into two groups: large CWSs (serving populations of more than 50,000), and medium-sized systems (serving populations of 3,301 – 50,000). EPA proposes to sample with certainty systems serving more than 100,000 people. These systems have the largest capital needs, and they have the staff to respond efficiently to the 2007 DWINSA. EPA proposes to draw a random sample of medium systems (serving 3,301 – 100,000 people). This methodology can reduce burden and still achieve the DWINSA data quality objectives. To meet the state-level precision targets, EPA will first determine the total sample size for each state to meet the target level of precision. EPA will then allocate the sample to strata in order to maximize the efficiency of their design.

Small CWSs

The objective of the 2007 DWINSA is to develop state-level estimates of total capital needs for CWSs. For large and medium systems, as explained above, this objective is achieved by selecting samples that are allocated across various strata in the population to achieve an overall precision level for each state. Several barriers prevent us from developing state-level estimates for systems serving populations 3,300 and fewer:

- First, a mail survey is not an effective approach to collection of data from these small CWSs. State experience with mail surveys for small CWSs suggests that total non-response and item non-response would be very high with a mail survey. Also, states believe that the absence of knowledgeable respondents at small CWSs limits the general reliability of the responses. Therefore, the 2007 DWINSA workgroup recommended, and EPA agrees, that the best way to gather information from small CWSs is through site visits made by EPA contractors. This will minimize total non-response, eliminate item non-response, and significantly improve the reliability of data collected.

- Second, if EPA assumes that all data collected from small CWSs will require site visits, then the number of such visits is constrained by the budget allocated for the 2007 DWINSA. EPA's current budget provides for approximately 600 site visits. Including small CWSs in the state-level design proposed for the medium and large systems, however, would require approximately 22,000 site visits. Thus, the statistical design for medium and large systems cannot be applied to small CWSs.

- Given this dilemma, the EPA workgroup committee recommended that EPA adopt a different approach for small CWSs, one that focused on national-level estimates. The direct sample estimates of total capital needs at the national level will be used to infer the total capital needs for small CWSs in each state. The workgroup preferred this approach.

EPA is designing and conducting the 2007 DWINSA with the assistance of a contractor:

Contractor	Contractor Roles
The Cadmus Group, Inc. **57 Water Street** **Watertown, MA 02472** **(617) 673-7000**	**– Technical oversight for all contractor activities** **– Oversight of data collection instrument design and testing.** **– Oversight of statistical sample design** **– Training** **– Mailings; logistics** **– Technical support for respondents and states** **– Model development** **– Data processing** **– Statistical sample design**

B.1.d Feasibility

The 2007 DWINSA data collection instrument has been designed with the capabilities of the typical respondent in mind. To fully assess feasibility, the Agency undertook the following steps. First, EPA convened a workgroup (see Section A.5.b) to comment on the proposed data collection and its feasibility. Second, EPA met with individual CWS operators and discussed the proposed survey. System operators were asked to comment on all proposed data elements and the feasibility of collecting information by a mail survey. The data collection instrument was pre-tested, as described in Section B.3.a.

The Agency recognizes that some medium CWSs (and a few large CWSs) may not have cost data or documentation of costs for some projects. In those cases, the 2007 DWINSA data collection instrument requests other readily-available information that EPA can use to model costs. EPA will make it very clear to respondents that they are not expected to develop cost estimates for the purposes of the 2007 DWINSA. In addition, EPA (or states) will provide large and medium CWSs with a helpline to assist them complete the data collection instrument.

Unlike the medium and large systems, the DWINSA will not be self-administered by small CWSs; rather, EPA contractors, accompanied by state personnel if state personnel participate in this portion of the 2007 DWINSA, will visit the small CWSs. Prior to the visit, the contractors will have access to all state records on the system (e.g., the results of recent sanitary surveys and inspections). The contractors will spend approximately 3 hours with the system owner or operator, requesting information that will be helpful in estimating system infrastructure needs. The contractor will then conduct a physical inspection of the system to confirm information provided by the owner or operator.

The EPA contractor will focus attention on the capital needs associated with treatment of source water, transmission, storage, and distribution. Capital needs associated with treatment will be modeled using methods similar to those currently used by EPA in the development of economic analyses. (In these analyses, data on occurrence of contaminants and cost estimates for treatment of source water to remove contaminants yield the cost of compliance with regulations that require the removal of contaminants from finished water.)

Reliance on site visits to small CWSs was strongly recommended by the EPA workgroup to avoid problems that have faced every state survey of small CWS infrastructure needs:

- *Total non-response.* Since many systems have not clearly identified responsible parties, and since responsible parties often are unwilling to respond to data collection instruments, it is difficult to use a mail survey to obtain the necessary information. Working with participating state regulatory agencies and representatives of small CWSs should minimize non-response problems.

- *Item non-response.* System owners and operators often are not knowledgeable about the capital needs of their systems. Unlike larger systems, who may maintain CIPs, small CWSs lack information to answer questions. Since EPA contractor engineers will conduct site visits to gather data, item non-response should be eliminated.

- *Reliability.* State drinking water regulators are suspicious of information provided directly from owners or operators of small CWSs. Unlike larger systems, small CWSs usually do not have professional, certified operators. Instead, one is likely to meet mobile home park owners, volunteers from homeowners associations, and others who are not water supply professionals. State drinking water administrators clearly prefer the judgments of EPA contractor engineers, accompanied by their own staff, for reliable information on capital needs.

Finally, employing site visits will substantially reduce the burden on small CWSs. Total burden on the systems, on average, will be about 1 hour. Instead of completing a data collection instrument, the system owner or operator can answer questions asked by the visiting engineer. The approach was discussed with knowledgeable state drinking water regulators, as well as representatives of small CWSs, and all parties agreed that it was the best approach to achieve the desired results of the 2007 DWINSA.

Sufficient contract funds have been identified to complete the 2007 DWINSA.

The time frame for the 2007 DWINSA is acceptable to the users of data within the Office of Ground Water and Drinking Water (OGWDW) and sufficient to complete a report to Congress by its anticipated due date of early 2009. The schedule also is acceptable to other users of the data.

B.2 SURVEY DESIGN

This section contains a detailed description of the statistical survey design including a description of the sampling frame, sample identification, precision requirements, data collection instrument, pre-test, collection methods, and follow-up procedures.

Medium and Large CWSs

The sample design for the DWINSA is stratified random sampling within each state. Stratification increases the precision of estimates compared with a simple random sample of the target population of systems. In stratified samples, the target population is divided into non-overlapping groups, known as strata, from which separate samples are drawn. The goal of stratified sampling is to choose sample sizes within each stratum in a manner designed to obtain maximum precision in the overall estimate for the population. Stratification variables for this study include: population size (populations of: 3,301 – 10,000; 10,001 – 25,000; 25,001 – 50,000; 50,001 – 100,000; and populations of more than 100,000), and primary source of supply (surface and ground). Systems serving more than 100,000 people are selected with certainty. The size of each state's sample of systems serving populations of 3,301 – 100,000 is set to meet the DWINSA's data quality objectives.

EPA's precision target is to be 95 percent confident that the true need lies within an interval, the upper and lower bounds of which do not exceed 10 percent of the sample mean (or estimated need). Once the total size of the sample of systems serving more than 3,300 people has been determined for each state, the number of samples to be taken in each stratum within each state will be allocated in a manner that minimizes the variance of the estimated total capital costs. EPA will use a Neyman allocation to determine the number of systems to select from each stratum. The Neyman allocation is described in detail in Section B.2.b.ii.

Small CWSs

The 2007 DWINSA design for small CWSs, like that for medium and large systems, is stratified random sampling. The stratification variables for small CWSs are the same as those for other systems: size of population served and primary source of supply.

Unlike the medium and large systems, the design for small CWSs is driven by a significant budgetary constraints: EPA cannot afford to complete more than approximately 600 site visits. EPA's objective in sampling is to achieve the maximum level of precision on a national basis without exceeding that budgetary constraint. Precision targets will be discussed in Section B.2.c, below.

B.2.a Target Population and Coverage

The target population is CWSs in the nation. A CWS is a PWS that serves at least 15 service connections used by year-round residents or regularly serves at least 25 year-round residents (40 CFR 141.2). The DWINSA is being designed to produce estimates of the capital need of medium and large systems for each state. It is being designed to produce estimates of the capital need of small systems for the nation as a whole.

B.2.b Sample Design

This section describes the sample design. It includes a description of the sampling frame, target sample size, stratification variables, and sampling method. The sampling design employed is a stratified random sample of CWSs. The strata employed in the design are discussed in Section B.2.b.iii. Neyman allocation is used to efficiently allocate the sample of water systems among the strata.

B.2.b.i Sampling Frame

The sampling frame is developed from SDWIS. SDWIS is a centralized database for information on PWSs, including their compliance with monitoring requirements, maximum contaminant levels (MCLs), and other requirements of the Safe Drinking Water Act (SDWA) Amendments of 1996. The following information will be extracted from SDWIS for the statistical survey and verified by participating states:

- Name of system
- Contact person
- Address of system
- Population served
- Total design capacity
- Number of connections
- Primary source (surface water or ground water)
- PWS identification number (PWSID)
- Ownership type
- Consecutive system (i.e., does system purchase or sell water)

From these data, EPA will develop the frame from which EPA will (1) calculate summary statistics (e.g., number of systems per state in pre-defined strata) for use in calculating sample size, and (2) randomly choose systems within the design strata to take part in the 2007 DWINSA.

Justification for the Use of SDWIS

The following criteria are often used in assessing a proposed sampling frame:

- It fully covers the target population.
- It contains no duplication.
- It contains no foreign elements (i.e., elements that are not members of the population).
- It contains information for identifying and contacting the units selected in the sample.
- It contains other information that will improve the efficiency of the sample design.

The units of observation for this medium and large system survey are CWSs, a subset of PWSs. SDWIS is the ideal choice for a sample frame because of its inclusive coverage of all units of observation for the 2007 DWINSA. In addition, SDWIS has two other advantages: it contains information that will facilitate contacting the respondents, and it contains other information that is useful in stratifying the sample, thereby improving the efficiency of the sample design.

In previous surveys where SDWIS was used as a sample frame, there have been criticisms of its utility. Since 1989, EPA has conducted audits of the quality of SDWIS data. As a result, EPA is aware of the problems with SDWIS. The audits, however, show that errors in classification of systems by strata proposed for the 2007 DWINSA are rare. The audits show that systems are misclassified by population or source in less than 1 percent of all cases.

To mitigate any potential problems with the sample frame, the 2007 DWINSA design anticipates substantial state involvement in the 2007 DWINSA process. Participating states, for example, will be checking the sample frame of systems that will be used to determine the final sample. In EPA's experience, states often have in-house data systems with very accurate data, particularly on medium- and large-sized CWSs. Even if these data are not transmitted to SDWIS, they are available to states and can be used by states to check the sample frame.

B.2.b.ii Sample Size

Medium and Large CWSs

Exhibit B-2-1 at the end of this subsection shows the preliminary sample sizes for the 2007 DWINSA. As shown on this exhibit, the sampling design will be implemented to achieve state-level precision targets for medium and large CWSs. Precision targets are discussed in Section B.2.c.

The task of determining the sample size for each stratum requires two steps. The first step determines the sample size for each state that achieves the precision targets for that state. The second step allocates the sample across the relevant strata in the state. The strata are described in section B.2.b.iii.

The first step in determining the sample size is calculating the total number of samples required at the state level to meet the precision requirements. The sample size is given by:

$$n_{0g} = \frac{\left(\sum_{h=1}^{H} N_{gh} s_{gh} \right)^2}{V_g}$$

Where: n_{0g} = the sample size (prior to the finite population correction)

N_{gh} = the total number of systems in the h^th^ stratum in the g^th^ state (taken from SDWIS)

s_{gh} = the standard deviation of the variable of interest for the h^th^ stratum in the g^th^ state (estimated using data from the 2003 DWINSA)

H = the number of strata defined in the sample design for the g^th^ state

V_g = the desired sampling variance for the total medium and large system capital needs estimate for state g.

The desired error in the sample is expressed as a relative error. In the above equation, V_g = (d/Z_α * $\widehat{Y}_g$)2. $\widehat{Y}_g$ is an estimate of the total capital needs for a given state. $\widehat{Y}_g$ is computed for each state by calculating the mean total capital needs for stratum h (from the 2003 DWINSA) and multiplying by the actual number of systems in each stratum for that state (N_{gh}). Summing across strata provides an estimate of $\widehat{Y}_g$. d is the half-width of the desired confidence interval (0.10 for the Assessment). Z_α is the value of a standard normal distribution for a confidence level of 1- α, (1.96 for the Assessment).

Because the number of water systems is known and finite, the following population correction is applied:

$$n_g = \frac{n_{0g}}{1 + \frac{1}{V_g} \sum_{h=1}^{H} N_{gh} s_{gh}^2}$$

The second step allocates the total sample to each of the strata EPA will randomly draw this number of samples from each of these strata. The Neyman allocation formula is used for the allocation:[1]

$$n_{gh} = n_g \left(\frac{N_{gh} s_{gh}}{\sum_{h=1}^{H} N_{gh} s_{gh}} \right)$$

(Because systems serving populations more than 100,000 are to be sampled with certainty, H is reduced by the number of large-system strata in the sample design for the large and medium systems.)

In order to implement these sample size and sample allocation equations, EPA needs estimates for V_g, N_{gh}, s_{gh}, and mean total capital needs by stratum. Information on mean total capital needs by stratum and s_{gh} were estimated using data from the 2003 DWINSA.

Exhibit B-2-1 State Sample Sizes

STATE	TOTAL NUMBER OF MEDIUM AND LARGE SYSTEMS	ESTIMATED SAMPLE SIZE FOR MEDIUM AND LARGE SYSTEMS
Alaska	21	16
Alabama	354	137
Arkansas	174	82
American Samoa	1	1
Arizona	127	28
California	676	194
Colorado	159	39
Connecticut	60	46
District of Columbia	1	1
Delaware *	26	3

[1] J. Neyman, “On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection,” *Journal of the Royal Statistical Society*, Vol. 97 (1934), pp. 558-606; as cited in William G. Cochran, *Sampling Techniques* (New York: John Wiley & Sons), 1977.

STATE	TOTAL NUMBER OF MEDIUM AND LARGE SYSTEMS	ESTIMATED SAMPLE SIZE FOR MEDIUM AND LARGE SYSTEMS
Florida	382	108
Georgia	224	57
Guam	3	3
Hawaii *	30	2
Iowa	135	40
Idaho *	45	1
Illinois	462	83
Indiana	213	100
Kansas	120	41
Kentucky	266	102
Louisiana	224	87
Massachusetts	247	57
Maryland	55	20
Maine *	35	1
Michigan	306	55
Minnesota	160	56
Missouri	207	82
Northern Mariana Islands	2	2
Mississippi	197	148
Montana *	34	1
North Carolina	272	47
North Dakota **	31	-
Nebraska	44	21
New Hampshire *	37	2
New Jersey	227	57
New Mexico *	60	1
Nevada	35	11
New York	355	56
Ohio	314	79
Oklahoma	160	53
Oregon	109	43
Pennsylvania	341	75
Puerto Rico	122	56
Rhode Island *	28	2
South Carolina	167	46
South Dakota *	42	1
Tennessee	282	114

STATE	TOTAL NUMBER OF MEDIUM AND LARGE SYSTEMS	ESTIMATED SAMPLE SIZE FOR MEDIUM AND LARGE SYSTEMS
Texas	968	109
Utah *	105	7
Virginia	164	45
Virgin Islands	3	3
Vermont	34	12
Washington	199	50
Wisconsin	177	51
West Virginia *	110	3
Wyoming **	27	-
Total	9,359	2,537

*Eleven states have chosen not to participate in the statistical portion of the survey (i.e., collecting data from systems serving 3,301 – 100,000 people). They will however participate in the census portion of the survey (i.e., collecting date from systems serving more than 100,000 people). The number in the “Estimated Sample Size for Medium and Large Systems” represents the total number of systems in the state that serve more than 100,000 people.

** Two states have chosen not to participate in the statistical portion of the survey (i.e., collecting data from systems serving 3,301 – 100,000 people). In addition, these states do not have any systems that serve more than 100,000 people.

Small CWSs

The total small system sample is set at 600 by available resources. EPA will allocate the sample among six strata to produce the most efficient estimate of small sample need, given this sample size. Section B.2.b.iii discusses the how the sample will be stratified. The sample for systems serving 3,300 or fewer people is allocated among source water and population-served strata using a Neyman allocation. Within each ground water stratum, the sample is divided proportionately between systems in states with and without a substantial occurrence of arsenic. This method was chosen because the data on the variance in system need by arsenic occurrence are not available.

B.2.b.iii Stratification Variables

The objective of stratification is to increase the efficiency of the sampling design (thereby reducing the number of samples required at any level of precision) by the creation of independent strata. Stratified sampling may produce a gain in precision in the estimates of the characteristics of the target population as compared to simple random sampling. In stratified sampling, the target population (i.e., CWSs) is divided into non-overlapping strata that are internally homogeneous, in that the measurements vary little from one unit to another (i.e., the within strata variance is minimized). If the within-stratum variance is relatively small, then a precise estimate of the variable of interest can be obtained with a relatively small number of samples. Each of the strata estimates can be combined to obtain a precise estimate for the target population. If the strata are constructed correctly, the target population estimate can be achieved with greater precision and with fewer samples than the estimate obtained from simple random sampling.

EPA’s drinking water programs have historically evaluated CWSs based on (1) size (number of persons served), and (2) primary source (ground water and surface water).[2] Using total capital need information obtained from the 2003 DWINSA, EPA evaluated several classification schemes. This analysis showed that the stratification scheme selected for the 2007 DWINSA medium and large system sample (10 strata based on size and source) was reasonable. Some states may have a different number of strata; this accommodated using their data as it is currently organized. Varying strata will be permitted only when the 2007 DWINSA’s overall precision is not reduced. The proposed strata for medium and large systems are as follows:

Size of Population Served	Source	Sample Methodologies
3,301 – 10,000	Ground	Random sample.
3,301 – 10,000	Surface	
10,001 – 25,000	Ground	Random sample. In some states the number of strata will be reduced based on analysis of optimal stratum boundaries. Specifically, in some states systems serving between 10,001 and 50,000 will be in one group rather than two.
10,001 – 25,000	Surface	
25,001 – 50,000	Ground	
25,001 – 50,000	Surface	
50,001 – 100,000	Ground	Random sample
50,001 – 100,000	Surface	
More than 100,000	Ground	Sampled with certainty
More than 100,000	Surface	

EPA’s sample design for small CWSs is also stratified based on the size of the population served and the source water of the system. In addition to source and population served, at least 25 percent of the counties selected will be in counties with high levels of arsenic. This is to ensure the collection of data regarding necessary infrastructure needs for systems serving 3,300 or fewer people affected by the Arsenic Rule. In 2003, infrastructure costs related to the arsenic regulation were determined using the Economic Analysis for the final rule. The workgroup decided this information was outdated and new estimates are necessary.

The proposed strata are as follows:

Water Source	Population Served
Surface Water Systems	25 – 1,000
	1.001 – 3,300
Ground Water Systems	25 – 1,000
	1.001 – 3,300

B.2.b.iv Sampling Method

[2] For the purposes of the 2007 DWINSA, purchased surface water systems are included with ground water systems. This design yields lower within-stratum variance.

As indicated above, all CWSs serving populations of more than 100,000 will be sampled with certainty.

For systems serving 3,301 – 100,000 persons, all CWSs will be allocated to 10 strata, based on population served and primary source. The sample size for each stratum in each state will be determined by the sampling strategy outlined above. The sampling method will be an equal probability random sample within each stratum. Anticipating a level of non-response, EPA will over-sample to achieve the desired number of completed data collection instruments. Since the expected response rate for systems serving 3,301 – 100,000 persons is 90 percent, EPA will draw a sample of 2,857.

All CWSs serving populations of 3,300 or fewer will be allocated to six strata, based on population served, primary source, and arsenic occurrence. The sample size for each stratum will be determined by the sampling strategy outlined above. The sampling method will be a two-stage probability proportional to size random sample within each stratum. Past response rates for these systems exceeded 90 percent. EPA will over sample to account for non-response, and will draw a sample of 600. [3]

B.2.b.v Multi-Stage Sampling

To achieve the required precision, reduce the burden to small systems, and to keep costs down, a two-stage cluster sample will be used for systems serving fewer than 3,300 people. The use of a two-stage sample design will result in slightly reduced precision for the stratum-level estimates.

First-Stage Sample

All small CWSs will be assigned to a county (or county equivalent in jurisdictions that do not have counties). Data on all small CWSs will be sorted by county so that EPA can determine the number of systems, by strata, in each county. If a particular county does not contain the required number of systems (a minimum of 6 systems), it is grouped with an adjacent county; the combined county group is referred to as a county-cluster. The first-stage sample will be approximately 120 counties, selected with probability proportional to size, where size is a composite measure of the number of small systems in each county. This method ensures that counties with more CWSs serving 3,300 or fewer people have a greater probability of being selected.[4]

States will be given a SDWIS list of small CWSs in the county (or counties) selected in the first-stage sample for their jurisdictions, and EPA will ask states to verify that the systems on the list are active CWSs with populations of 3,300 and fewer and assigned to the appropriate county. If the number of systems in a county is large (e.g., 100 or more), EPA will select a sub-sample of the systems in that county to reduce the burden on the state. This review by the states will produce a clean sample frame for the second-stage sample.

Second-Stage Sample

In the second stage, a stratified random sample of five systems is drawn from each of the counties or county-clusters selected in the first stage of sampling.

[3] For purposes of burden calculation, EPA assumes 100 percent response.

[4] This method is based on Folsom, R.E., F.J Potter., and S.R. Williams, “Notes on a Composite Size Measure for Self-Weighting Samples in Multiple Domains,” *American Statistical Association 1987 Proceedings of the Section on Survey Research Methods*, August, 1987, pp. 792-796.

B.2.c Precision Requirements

B.2.c.i Precision Targets

The sampling design for large and medium systems will be implemented at the state level. EPA's goal is to be 95 percent confident that the margin of error, when estimating the total capital needs facing these systems in each state, will be plus or minus 10 percent of the total need for these systems. For example, if the total need for these systems in a state is estimated to be $2 billion, EPA will be 95 percent confident that the actual total need is between $1.8 billion and $2.2 billion.

The size of the small system sample is driven by budget constraints, not precision targets. EPA estimates that the sample size of 600 will allow it to estimate the national capital need of these systems with a 95 percent confidence interval equal to plus or minus 15 percent of the national small systems need. This precision level will be less than the level for estimates developed for medium and large systems, but it will not materially reduce the overall precision for total cost estimates at the state level. Small CWS costs are a small portion of total system costs in each state. Thus, the lack of precision for these systems will not significantly reduce the overall precision of the state-level estimates.

B.2.c.ii Nonsampling Error

EPA has developed an assessment approach that will employ several quality assurance techniques to maximize response rates, response accuracy, and processing accuracy to minimize nonsampling error. A pre-test will supplement the experience of EPA and its contractor (The Cadmus Group, Inc.) in formulating a strategy to reduce non-sampling error.

Particular emphasis will be placed on maximizing response rates. Standard methods that have proved effective in other surveys of CWSs will be used, including the following:

- States will review the sample of systems to receive the mail data collection instrument and will ensure that the best person to receive the data collection instrument is determined in advance.

- EPA and the states will coordinate in the production of a cover letter for the 2007 DWINSA. EPA's opinion (shared by state drinking water administrators, trade associations, and PWSs) is that surveys on state letterhead will be better received than letters on EPA letterhead. Therefore, states can use state-level cover letters signed by a senior state official instead of the EPA letter.

- The data collection instrument design, content, and format have been reviewed by organizations representing CWSs. In addition, the data collection instrument design, content, and format were reviewed by states that participated in the 1995, 1999, and 2003 DWINSAs.

- The data collection instrument design, content, and format will be pre-tested to ensure that all questions are properly stated and can be answered by all systems in the mail survey.

- Items being asked are those that owners or operators of systems serving populations greater than 3,300 should know. EPA does not ask for items that require monitoring, research, or calculations on the part of the respondent.

- The data collection instrument design is limited to 12 pages. By limiting the information requested, EPA believe that the average respondent can complete the data collection instrument in approximately 4 hours.

- Toll-free phone numbers will be provided to help respondents with questions or problems. In addition, respondents will be encouraged to call state personnel who will be trained to answer questions.

- Pre-paid return envelopes will be provided to respondents to make returning the data collection instrument convenient.

Standard methods to reduce other sources of non-sampling error also will be used.

- EPA expects complete coverage of the target population using SDWIS, supplemented by state agency review of all systems.

- Data will be 100 percent independently keyed and verified.

- The data collection instrument is pre-coded to improve accuracy by eliminating unnecessary processing steps.

Supplementing these standard methods, EPA proposes several unique steps to eliminate non-sampling error, which have been developed in concert with organizations representing the states and CWSs. These organizations believe that the 2007 DWINSA is important and that a high level of participation by all CWSs is essential to its success. Because of the substantial commitment being made by states and CWSs to the 2007 DWINSA, EPA believes that response rates will be higher than most surveys of similar respondents. To ensure success, states and organizations representing CWSs are taking the following steps.

- *Participation of the states.* Because the DWINSA will be used to allocate DWSRF funds to states, each state has a strong interest in achieving a high response rate. EPA believes that state participation will be a key factor in guaranteeing high response rates and low item non-response. State personnel who work with CWSs every day are in a strong position to encourage systems to complete the 2007 DWINSA form. These states have committed to assisting EPA in achieving a high response rate by participating in follow-up activities. The states also will be available for technical assistance for any system that has questions about the 2007 DWINSA. All states have already agreed to participate in the 2007 DWINSA.

- *Participation of Organizations Representing CWSs.* EPA anticipates public support of organizations representing CWSs. The prior assessments were supported by groups such as the American Water Works Association (AWWA), the National Association of Water Companies (NAWC), and the Association of Metropolitan Water Agencies (AMWA).

  This support by the organizations representing the respondents for the 2007 DWINSA can be helpful in many ways to minimize non-sampling errors. For example,

  - These associations are likely to agree to prepare a letter for each system in their membership, stressing the importance of the 2007 DWINSA of drinking water infrastructure needs. This letter, along with the letter from the states, should make systems more likely to respond.

  - In the past DWINSAs, the largest association representing CWSs serving populations greater than 3,300—AWWA— provide support of its national organization behind the DWINSA. To improve the response rate, the AWWA enlisted of the support of its state affiliates (called "Sections") in telephone follow-up to encourage response. AWWA assisted in past DWINSAs to help achieve the overall response rate of 94 percent. EPA hopes to secure similar AWWA support for the 2007 DWINSA.

- *Communications Strategy*. EPA has developed a comprehensive communications strategy that will inform likely respondents of the need for their participation. This strategy includes articles in magazines, newsletters, and bulletins of all major organizations that represent (or communicate with) CWSs. This includes publications of all of the organizations mentioned above, plus the state and local affiliates of these organizations. The strategy is designed to develop widespread peer-group support for participation in the 2007 DWINSA.

B.2.d Data Collection Instrument Design

Questions about system characteristics (name, population served, number of connections, and other customary business information) will be pre-printed on all data collection instruments. The respondent needs only to enter accurate information if any pre-printed information is not correct.

The 2007 DWINSA is based on matrices that request a list of capital projects that the system plans for the period 2007 through 2026. For each project listed, the system is asked to provide: type of need; documentation of need and cost (if necessary); if the project is for new infrastructure or rehabilitation of existing infrastructure; if the project is needed now to protect public health or if it is needed over the next 20 years to continue to provide safe drinking water; the federal regulation or state requirement if the project is to meet a current regulation or state requirement; design capacity of source, storage, and treatment projects; cost of the project; and date of the cost estimate. For most of these variables, respondents will choose the appropriate "documentation," "type of need," or "regulation or requirement," from the Lists of Codes. All matrices have been designed to be concise, to avoid jargon, and to avoid ambiguous words or instructions. Terms and formats have been standardized to the extent possible. There is no intentional bias in the ordering of the items.

This page intentionally left blank.

B.3 PRE-TESTS AND PILOT TEST

B.3.a Pre-tests

EPA conducted two pre-tests of the data collection instrument for the 2007 DWINSA. The 2007 DWINSA pre-tests were conducted by EPA's contractor, The Cadmus Group, Inc. The pre-tests gathered feedback on the effectiveness of the data collection instrument, highlighted imprecise, ambiguous, or redundant questions, and indicated where further inquiry is needed. A pre-test was held in both Maine (four participants) and Montana (three participants). These states were chosen because they are both "opt-out" states, and because most of their systems will not need to participate in the 2007 DWINSA. Also, the contractor conducting the pre-tests has offices in both these states and by conducting the pre-test in these states they were able to reduce costs. The names of the seven systems were provided to EPA by the state 2007 DWINSA contacts. Based on the comments received EPA made modifications to the data collection instrument.

B.3.b Pilot Test

To eliminate unnecessary burden on states and CWSs, it has been decided that no pilot test for the 2007 DWINSA will be conducted. A pilot test was conducted for the 1995 DWINSA and consisted of 60 CWSs from New York and Texas. The procedures for mailing the data collection instruments and collecting the data are the same as those used for the 1995, 1999, and 2003 DWINSAs. EPA believes these procedures are well tested and have proven to be successful; therefore, it is not necessary repeat this testing step.

This page intentionally left blank.

B.4 COLLECTION METHODS AND FOLLOW-UP

B.4.a Collection Method

The proposed collection method for medium and large systems is a mail survey. The study data collection instrument and lists of codes will be mailed to all systems in the sample. State drinking water agencies will begin follow-up if the mail data collection instrument has not been returned in 30 days. For a complete description of the follow-up procedures proposed to increase the response rate, see section B.2.c.ii.

The proposed collection method for small systems is to visit to each small system in the sample. An EPA contractor, accompanied by state personnel that choose to participate, will interview the owner or operator and fill in the data collection instrument for all costs except treatment costs. (Costs of treatment will be modeled, using methods similar to those used by the OGWDW for regulatory impact analyses for new regulations.)

B.4.b Survey Response and Follow-up

The target response rate (defined as the ratio of responses to eligible respondents) for the 2007 DWINSA is 90 percent. EPA realizes that this is an ambitious target, but EPA believes that there are special circumstances that warrant such a target. Also, overall response rates of 94, 97, and 96 percent were achieved in the 1995, 1999, and 2003 surveys, respectively. In the first three surveys, EPA conducted the following proposed activities to achieve that high response rate.

- *Support from the Respondent Population*. This is a national survey of infrastructure needs for drinking water systems. The medium and large systems, as well as all national organizations representing these systems, understand the importance of the DWINSA results. All national organizations have endorsed the DWINSA and have communicated the importance of a high response rate to their members. As discussed in Section B.2.c, organizations have provided access to their newsletters and magazines to publicize and endorse participation in the DWINSA. EPA will ask national organizations representing smaller CWSs (e.g., the National Rural Water Association (NRWA) and AWWA) to help communicate the importance of a high response rate to their members.

- *Follow-up by States and Respondent Peer Groups.* Since a majority of participating states have indicated their willingness to participate in follow-up activities, these procedures will be implemented by state personnel, most of whom are personally familiar with the respondents. Procedures that states will use include reminder letters and telephone follow-up. In states that elect not to participate in follow-up, the EPA contractor will conduct these activities. If the follow-up fails after three attempts (one reminder letter plus two telephone follow-ups), EPA will shift to a second approach: peer-group follow-up by members of a trade association, such as AWWA. Procedures to be used by the association include a reminder letter followed by telephone calls. Such involvement is likely to improve the 2007 DWINSA's response rate.

- *Recruitment by States and Respondent Peer Groups of Small Systems.* In participating states, scheduling of site visits will be conducted by state personnel, most of whom are personally familiar with the respondents. If state personnel cannot schedule a visit with a system in the sample, EPA will turn to respondent peer groups.

This page intentionally left blank.

B.5 ANALYZING AND REPORTING SURVEY RESULTS

B.5.a Data Preparation

State personnel will check all cost data and documentation to ensure that it is consistent with state and national standards. States will then send the completed and reviewed data collection instruments to EPA for a second round of review by EPA contractor staff.

Once data have been checked, the contractor will key and verify the data. Senior data entry staff will be used for the verification process to improve quality control. Editing will include automated logic and range checks and checks for missing data. Missing cost data will be modeled, using other information provided by the respondents on the data collection instrument. When modeling is insufficient, missing data will be imputed using the standard methods such as cell means and regression. The sample of CWSs will be weighted so that stratum estimates can be summed to prepare state-level estimates.

B.5.b Analysis

EPA will prepare a report that tabulates the results of the 2007 DWINSA and explains the precision of the state-level estimates of total capital needs. Examples of statistics that will be produced include:

- Eligible capital needs by state and by types of need.

- Total capital needs by state and by types of need.

- Total capital needs by domains within the total population, e.g., systems serving populations greater than 100,000.

- Mean and median statistics on total capital needs (by type of need) for systems of various sizes. (These data will be of particular interest to participating respondents who will receive a short summary of these statistics.)

- Standard errors calculated for key statistics.

The analysis will be similar to that done for previous DWINSAs.

B.5.c Reporting Results

The 2007 DWINSA results will be made available to EPA and the public through:

- A printed report that is submitted to Congress on drinking water infrastructure needs. This report will be distributed to all participants in the 2007 DWINSA and all interested offices at EPA.

- Micro-computer access to state data (each state can access only its own data).

- Micro-computer access to the entire database (EPA only).

A report containing all technical information (data collection instrument, sampling plan, response rates, and variances) will be prepared and distributed. Record layouts, codes, and complete file documentation will be developed for data users (both micro-computer and mainframe users).