Supporting Statement Part B
Weatherization Assistance Program ARRA-Period Evaluation
OMB Control Number: XXXX-XXXX

B. <u>Collections of Information Employing Statistical Methods</u>

B.1. Describe (including a numerical estimate) the potential respondent universe and any sampling or other respondent selection methods to be used.

The WAP ARRA-Period Evaluation is designed to be largely similar to the ongoing WAP Retrospective Evaluation, which was in turn designed to be similar to an evaluation conducted in 1993, the most recent fully complete WAP evaluation.[1] All three evaluations require or required statistical sampling of weatherization agencies, subsampling of agency staff and weatherized and comparison home occupants, and weatherized and comparison home utility billing data. The design of the ARRA-Period Evaluation should also enable efficient comparisons with the Retrospective Evaluation so that differences between the two evaluations can be estimated, and commonalities estimated with data from both studies to increase statistical precision.

For the Retrospective Evaluation, a sample of 400 agencies was selected from a target population of 904 agencies. The sample size of 400 was based on the sample size in the 1993 study, the response rate (90%) for which was considered adequate, and the estimates from which were considered sufficiently but not excessively precise.[2] We therefore plan to sample 400 agencies again for the ARRA-period study.[3]

Updated lists of WAP agencies are maintained in the DOE's WinSAGA[4] database. Fiscal year 2008 WinSAGA data was used to sample agencies for the retrospective study. If the target population of weatherization agencies were static, then the same sample of agencies could be used for both the retrospective and ARRA studies, and differences between the ARRA and retrospective study periods could be estimated more efficiently than would be possible, for example, with two independent cross-sectional samples of agencies. However, although the target populations for the two evaluations contain many of the same agencies, new agencies have come into existence since 2008, and some agencies that existed in 2008 are now defunct.

[1] Brown, Marilyn A., Berry, Linda G., Balzer, Richard A., and Faby, Ellen, "National Impacts of the Weatherization Assistance Program in Single-Family Dwellings," ORNL/CON-326, Oak Ridge National Laboratory, Oak Ridge, Tennessee, May, 1993 (http://weatherization.ornl.gov/pdf/ORNL_CON-326.pdf).

[2] See Section B.3.2 below.

[3] Response rates for key information collection instruments implemented during the retrospective evaluation are: S1 – State Program Information Survey: 100%; S2 – Agency Program Information Survey: 92%; S3 – Subset of Agencies Detailed Program Information Survey: 90%; DF4 – Electricity & Natural Gas Bills Information from Agencies: 95%; and DF2/3 – Housing and Building Information Data Forms: 85%.

[4] "WinSAGA" refers to DOE's Systems Approach to Grants Administration for Windows. WinSAGA data kindly provided by Christine Askew, Office of the Weatherization and Intergovernmental Program, Energy Efficiency and Renewable Energy, U.S. DOE.

Table B.1.a shows numbers of agencies with FY10 funding allocation for new and original agencies according to whether they were sampled in the retrospective study. Table B.1.b shows the corresponding percentages of FY09-10 planned allocations for the same classification of agencies.

Table B.1.a. Numbers (and Percentages) of New and Original WAP Agencies with FY10 Dollar Allocations

Agency Status	Not Sampled in Retrospective Evaluation	Sampled in Retrospective Evaluation	Total
New	130 (13.3%)	0 (0%)	**130 (13.3%)**
Original	465 (47.6%)	381 (39.0%)	**846 (86.7%)**
Total	**595 (60.9%)**	**381 (39.0%)**	**976 (100%)**

Table B.1.b. Percentages of Total Planned Allocation for FY09-10 for New and Original WAP Agencies with FY10 Dollar Allocations

Agency Status	Not Sampled in Retrospective Evaluation	Sampled in Retrospective Evaluation	Total
New	20.1%	0.0%	**20.1%**
Original	35.5%	44.4%	**79.9%**
Total	**55.6%**	**44.4%**	**100.0%**

To ensure that estimates computed for the ARRA period are unbiased, new agencies must be sampled for the ARRA study, though there are a number of reasons to more heavily sample agencies that were sampled and responded in the retrospective study:

- Because of changes in WAP operating parameters that were made to accommodate increased ARRA-period budgets (e.g., maximum income and unit spending limits were both increased), it is imperative to compare WAP ARRA-period with earlier WAP performance. Comparisons of ARRA and retrospective period performance will be much more efficient with agencies sampled in both evaluations because agency effects subtract out and thus don't contribute to the statistical error in the comparisons.

- Contacts with already-sampled agencies have already recently been established in the Retrospective Evaluation.

- Because the data requests in the two evaluations will be very similar, agencies that responded in the Retrospective Evaluation will be more likely than agencies in general (either new or original) to respond in the ARRA-Period Evaluation.

These features make agencies already sampled in the retrospective study more attractive as samples for the ARRA-period study than either new agencies or original agencies that were not previously sampled or were sampled but did not respond.

Of the 400 agencies sampled in the retrospective study, 15 are now defunct, and 3, though still listed in the WinSAGA database, are listed for FY10 as having zero planned allocations or units. Of the remaining 381 sampled agencies, about 90 percent or 343 are expected to respond in the retrospective study.[5] From Table B.1.a, of the 976 agencies with FY10 funding allocations, 130 (13.3%) are new agencies, and from Table B.1.b, the new agencies account for 20.1% of the total funding allocation planned for FY09-10. For an ARRA-period study sample of 400 agencies, this suggests that if we sample all of the approximately 343 agencies that will have been sampled and will have responded in the retrospective study, then we can still sample 57 new agencies, which is in the range of 53-80 (13.3% to 20.1% of 400) and is thus an appropriate number of new agencies to sample. Therefore, for the ARRA-period agency sample we will sample 343 agencies sampled previously in the retrospective study and 57 new agencies. The procedure for sampling new agencies is discussed in Section B.2 below.

Sampled agencies (both original and new) will be asked to provide lists of current and, for comparisons, future (i.e., wait-listed) clients, corresponding electricity, natural gas, or bulk-fuel utilities, and physical characteristics of the client homes. For a sample of clients, requests to the utilities will be made for the client billing data. In the 1993 evaluation, a 50% response rate by utilities was encountered in requests for billing data. We are assuming about the same utility response rate for the ongoing Retrospective Evaluation, and preliminary results are consistent with that assumption. Therefore we assume this 50% response rate again for the ARRA-period study. Utility non-response is important because it reduces the final acquired number of billing records. Utility non-response tends to be nonbiasing, however, because it does not reflect on the performance of the weatherization agencies themselves.

In the 1993 and retrospective studies, clients were/are sampled from the client lists provided by the sampled agencies at the rate of one in three for electricity and natural gas and one in four for bulk fuels. The billing data for the sampled clients was/is then requested from the utilities. Note that the effort for agencies or utilities to pull a set of records does not increase in proportion to the number of records in the set. For this reason and for the sake of simplicity, the same 1-in-3 or 1-in-4 sampling rates were used for all agencies, regardless of how many clients they had. At their option agencies and utilities may and sometimes do also provide data for all weatherization and comparison clients, not just sampled ones.

For the ARRA-period survey, client lists will be longer (i.e., more clients) than the lists for either the 1993 or retrospective evaluations, and so the net number of client billing series obtained will be larger for the retrospective study. Nevertheless, the same 1-in-3

[5] For the retrospective evaluation as of April 29, 2011, 347 agencies had submitted validated responses, responses for 8 were pending validation, requests for 30 were still in process, 11 agencies were deemed non-responders, and 16 were not contacted. This is consistent with the 90% response rate expected on the basis of the 1993 evaluation.

and 1-in-4 sampling rates will be used again, because reducing the rates would not substantially reduce the effort in providing either the client lists or billing data.

In addition to clients and utilities for billing data analysis, agency staff and home occupants will also be subsampled after the agencies themselves are sampled. Additional sampling will also be needed for various special evaluation studies, for example, of under-performers and for the Weatherization Innovation Pilot Project (WIPP). Additional sampling is discussed in Section B.3.2 below.

Table B.2, which shows proposed sample sizes for agencies, occupants, utilities, and agency staff and for the special studies, summarizes the sampling for the ARRA-Period Evaluation.

Table B.2. Weatherization Assistance Program Evaluation Proposed Sample Sizes for Sampled Study Components

Evaluation Component	Proposed Sample Size	Basis Discussion (in this document)
Agency Survey	400 agencies: client lists and building characteristics All agencies: short form	B.1 (this section)
Billing Data Analysis	Utility bills for 1 in 3 or 1 in 4 clients of sampled agencies	B.1 (this section)
Weatherization Staff Survey	271 staff surveys (same as in retrospective study)	B.2.3.1
WIPP Study	68 home inspections and 243 staff surveys and 267 occupants	B.2.3.2
Sustainable Energy Resources for Consumers (SERC) Study	103 home inspections + 384 occupants	B.2.3.3
Under-Performer Study	42 “over-performer” homes + 73 “under-performer” homes + occupant survey for these	B.2.3.4
Air Conditioning Monitoring	132 + 132 = 264	B.2.3.5
Persistence Study	228 homes (blower-door measurements)	B.2.3.6
Indoor Air Quality Remediation Cost Study	Continuation of retrospective study	B.2.3.7
Deferral Study	203 deferral homes	B.2.3.8
Territories study	Treatment and comparison homes for billing analysis + 258 occupants	B.2.3.9

B.2. <u>Describe the procedures for the collection of information including:</u>

B.2.1. Statistical methodology for stratification and sample selection

To be eligible for inclusion in the ARRA-period study, agencies must have had FY10 funding allocations. The target population of 976 such agencies was stratified into two groups: 846 "original" agencies, that is, agencies in the target population for the Retrospective Evaluation, and the remainder of 130 "new" agencies. The 846 original agencies were sampled in the retrospective study using probability proportional to size (PPS) sampling stratified by state, with "size" taken as FY08 funding allocation, and allocation of sample to states in proportion to size. When the sample for the ARRA-period survey is actually selected, approximately 343 original agencies of the 400 originally-sampled agencies will have responded. Those 343 will comprise the sample of original agencies for the ARRA-period study. New agencies will also be sampled, also with PPS sampling. The PPS size measure for the new agencies will be the FY10-11 planned funding allocation.[6] Because the distribution of new agencies is irregular over states, however, new agencies will not be stratified by state. The total number of agencies sampled will be 400, and thus the number of new agencies sampled will be 400 – 343 = 57.

Subsampling (of occupants, agency crew, etc.) will be by simple random sampling (SRS) within the various strata and in some cases further stratified (e.g., by crew function, weatherized-vs-comparison subjects). Occupant sampling for the ARRA-Period and Retrospective Evaluations will differ in that in the Retrospective Evaluation, WAP occupants were sampled twice, about a year apart, whereas in the ARRA-Period study occupants will be sampled only once.

B.*2.2. Estimation procedure*

Most of the statistical analysis of the ARRA-period survey data will be to compute summary totals and means (e.g., of energy and cost savings). All such analyses will account for the stratified sampling design and sampling weights (e.g., with the SAS[7] Surveymeans or Surveyfreq procedures). Sampling weights (which are computed when the PPS samples are generated) will be adjusted for non-response and for the original agencies that are now defunct or have zero FY10-11 funding allocation.

Billing data and other fuel consumption analyses require regression analysis to adjust for differences in weather across seasons and locations. These adjustments will be made using multiple approaches (as an internal check), for example the Princeton Scorekeeping Method[8] as well as a straightforward linear model

[6] Slight departures from PPS sampling may be necessary to accommodate very heterogeneous agency sizes. For this ARRA-period evaluation, agency sampling will also be constrained so that 5 of the 57 new agencies are SERC grantees. See SERC discussion in Section B.2.3.3 below.

[7] 2004 SAS/Stat 9.1 User's Guide, Cary, NC: SAS Institute, Inc.

[8] Fels, M., K. Kissock, M. Marean, and C. Reynolds, "PRISM Advanced Version 1.0 User's
Guide," Princeton University, Center for Energy and Environmental Studies, Princeton, NJ,

approach with billing days and heating and cooling degree days as independent variables.

B.2.3. Degree of accuracy and sample sizes needed for the purpose described in the justification

The most important endpoint of all in the WAP evaluations is probably electricity and natural gas savings estimates. Estimates and standard errors of the mean control[9]-adjusted natural gas and electricity savings per weatherized unit per year, computed for the 1993 evaluation, are listed in the following table:

Table B.3. Accuracy of Estimates Computed in 1993 WAP Evaluation[10]

Primary Heating Fuel	Average Savings per Weatherized Unit (1)	Standard Error of Average (2)	Relative Error: (2)/(1) as Percent
Natural gas	17.8 million Btu (MMBtu)/year	1.8 MMBtu/year	10%
Electricity	1,830 kilowatt-hours (kWh)/year	358 kWh/year	20%

Because of increased ARRA-period funding levels, there will be more clients per agency in the ARRA-period study than in the 1993 study. Because the standard errors in Table B.3 involve fairly complicated degree-day-adjusted regression analyses, however, it would be difficult to quantify how standard errors computed for the ARRA-period survey will differ from the standard errors in the table. Nevertheless, it is reasonable to assume that if the same number (400) agencies are sampled again, then the ARRA-period survey will have the same or somewhat better precision than the 1993 study. Therefore 400 agencies will be sampled again for the ARRA-period survey.

As discussed below (see Section 2.3.3) a number of additional agencies will be sampled as part of the SERC study. All WIPP grantees will also be samples (Section B.2.3.2). Additional agencies will also be samples for a Territories study (see Section B.2.3.9). Otherwise, with the exceptions of the complete sampling of states and the complete sampling of agencies themselves (short form only), all other components of the evaluation will require subsampling through the primary agency sample. The remainder of Section B.2.3 describes the sample size requirements for the various, subsampled, component studies.

1995.

[9] "Control" and "comparison" will sometime be used interchangeably in this document.

[10] Brown et al, op. cit.

B.2.3.1. Weatherization Staff Survey

For the Weatherization Staff Survey a subsample of agency staff members will be identified from lists compiled from agency staff contact information. To ensure adequate representation, the sample will be stratified by staff functional classification (crew, supervisor, auditor/inspector) with equal-size strata. A web-based survey of the sampled staff will be conducted to characterize current staff understanding and awareness of weatherization methods and technologies. The primary endpoint of interest is the combined proportion of correct responses to technical questions.

As an approximation, we determine a sample size necessary under a model with simple random sampling from the lists of agency staff members concatenated for all responding (as yet undetermined) agencies. In practice, however, a sample of that size will be apportioned across the responding agencies. Thus the net sample size will not be affected by agency non-response.

The proportions of correct responses in each sampling strata (crew, supervisor, auditor/inspector) will be estimated to within approximately five percentage points with at least 90% confidence. The standard error of the combined proportion of correct responses can be no greater than the standard error of an individual (correct/incorrect) response, which cannot exceed $.5/n^{1/2}$ (maximum standard error of binomial proportion). Therefore the estimate will be accurate to within approximately five percentage points with at least 90% confidence if $Z_{0.95} \times .5/n^{1/2} = .05$, where $Z_{0.95}=1.645$ is the 95th percentile of the standard normal distribution. This implies a sample size of n = 271 per stratum and a total sample size for crews, supervisors, and auditors/inspectors of 271×3 = 813.

Non-response (if any) among staff members will be partially accommodated by making additional selections to fill in for the non-responders so that non-response, though potentially biasing, does not affect ultimate sample sizes. For agencies that have responded at all, however, staff member non-response is expected to be minimal because staff contact information is being provided by the agencies themselves. Thus non-response bias will be negligible for agency staff.

B.2.3.2. Weatherization Innovation Pilot Project (WIPP) Study

The WIPP evaluation will entail an occupant survey, home inspections, utility bill analyses, and weatherization staff surveys. As the WIPP focuses on innovation, each of its various project components tends to be unique. Therefore the WIPP surveys will be stratified by project component. Table B.4 summarizes the sixteen project components. Numbers of units weatherized, staff members (jobs), and total Federal request are listed for each project component. The numbers in the table are known, though the 18,528 units and 2,306 staff members referred to in the table cannot yet be listed in a population frame. The unit and staff listings will be obtained from the WIPP component grantees.

Billing data for each weatherized unit, which is being collected by the WIPP grantees as part of the project, will also be requested and analyzed by the evaluation team.

Comparison data will be collected from an equal or greater number of units from the comparison group for the main ARRA-period evaluation. Units will be matched as closely as possible by location, heating and cooling degree days, size, construction type, and other characteristics.

Table B.4. Weatherization Innovation Pilot Project Component Statistics

WIPP Components	Number of Units Weatherized	Number of Jobs Created or Saved	Federal Request ($)	Percent of Total Federal Request
Green and Healthy Homes Initiative	220	96	2,400,000	8.00
In Home Monitoring	2,500	120	2,400,000	8.00
Performance-based Revolving Loan Pilot...	450	38	850,000	2.83
Energy Pioneer Solutions Weatherization...	250	25	2,400,000	8.00
SAHF Energy Performance Contracting...	2,500	123	810,000	2.70
Connecticut Green and Healthy Homes...	2,285	593	3,000,000	10.00
Streamlined Weatherization Improvements...	800	25	2,000,000	6.67
Leveraging Smart Grid Technology to...	550	16	720,000	2.40
YouthBuild USA Weatherization...	998	74	1,400,000	4.67
Habitat for Humanity Weatherization...	1,770	168	3,000,000	10.00
Community Environmental Center...	1,200	63	3,000,000	10.00
Building Deep Efficiency...	425	28	600,000	2.00
Project with the City matching federal...	300	10	1,015,746	3.39
Tackling the Problem of Weatherizing...	1,700	85	1,898,938	6.33
Replicable, Innovative, Sustainable...	2,240	169	3,000,000	10.00
People Working Cooperatively...	340	673	1,500,000	5.00
All Projects	**18,528**	**2,306**	**29,994,684**	**100.00**

Minimum necessary sample sizes for the WIPP surveys are calculated below as the minimum sample size needed under SRS with finite population correction. In practice this sample size will be allocated across the sixteen project components (strata) in proportion to their Federal requests (and rounded up to the next whole integer). As stratification is expected to increase precision, the necessary sample size calculated on the basis of SRS is slightly larger than the minimum needed under stratified sampling.

For each survey, the SRS sample size is computed for estimating a yes/no binary response proportion: For home inspections, does the home pass inspection? For occupants, is the occupant (i.e., client) satisfied with the service? For the staff survey, is the respondent satisfied with his/her job? For the occupant survey, which will be conducted by telephone interview, and for the staff survey, which will be a web survey, we use a five percentage point margin of error and 90% statistical confidence. For home inspections, which require site visits, we use a 10 percentage point margin of error and 90% statistical confidence. The SRS-based sample sizes for these criteria are 68 for home inspections, 267 for occupants, and 243 for staff survey.[11] Table B.5 shows the

[11] The SRS-based sample size is larger for occupants because of the finite population correction and because the target population is larger for occupants.

allocations for these surveys across the sixteen WIPP component strata. The totals are slightly larger than the SRS-based sample sizes because stratum-specific totals are rounded up to the next whole integer. Non-response is expected to be minimal for these surveys, so that non-response bias will be minimal, and the sample sizes can be achieved by sampling additional subjects to replace any that fail to respond.

Table B.5. Allocation of WIPP Samples Across the WIPP Component Strata

WIPP Project	**Home Inspections (10/90)**	**Occupant Survey (5/90)**	**Staff Survey (5/90)**
Green and Healthy Homes Initiative	6	22	20
In Home Monitoring	6	22	20
Performance-based Revolving Loan Pilot...	2	8	7
Energy Pioneer Solutions Weatherization...	6	22	20
SAHF Energy Performance Contracting...	2	8	7
Connecticut Green and Healthy Homes...	7	27	25
Streamlined Weatherization Improvements...	5	18	17
Leveraging Smart Grid Technology to...	2	7	6
YouthBuild USA Weatherization...	4	13	12
Habitat for Humanity Weatherization...	7	27	25
Community Environmental Center...	7	27	25
Building Deep Efficiency...	2	6	5
Project with the City matching federal...	3	10	9
Tackling the Problem of Weatherizing...	5	17	16
Replicable, Innovative, Sustainable...	7	27	25
People Working Cooperatively...	4	14	13
All Projects	**75**	**275**	**252**

B.2.3.3. Sustainable Energy Resources for Consumers (SERC) Study

The SERC program is similar to the WIPP in that both programs explore new and unique treatment technologies not standard to the WAP. The evaluations of both programs will entail home inspections, an occupant survey, and energy usage (e.g., billing-data) analysis. Thus the statistical design objectives for the two evaluations are quite similar. However there are 92 SERC grantees (as opposed to 16 WIPP grantees), and SERC grantees are actually WAP agencies. The number of units planned to be weatherized in FY10-11 by the 92 SERC agencies is 56,570, many more than for the WIPP.

Of the 92 SERC grantees/agencies, 36 are original agencies previously sampled in the Retrospective Evaluation, 51 are original agencies not previously sampled, and 5 are new agencies. As discussed in Section B.1, the ARRA-period agency sample will consist of the approximately 343 original agencies that were sampled and responded in the retrospective study and 57 new agencies to be sampled from the population of 130 new agencies (see Table B.1.a). The 36 previously sampled SERC agencies will be resampled in the ARRA-period study, and the 5 new agencies will be sampled as part of the 57 new agencies to be sampled. Because none of the 51 original but not previously sampled SERC agencies can be used in pairwise comparisons of the retrospective and ARRA study periods, they will be treated as a supplemental sample of agencies.

However, once the original but not previously sampled SERC agencies are contacted as part of the SERC evaluation, the additional effort for agency staff to answer the questions for the regular evaluation will be minimal, and so that information will be requested from them as well.

We assume unstratified SRS as an approximation in determining a sample size. In practice homes and occupants will be sampled by SRS stratified within agency. As in the approach in Section B.2.3.2 for the WIPP, standard criteria, 10-95 (percentage-point margin-of-error and statistical confidence) for home inspections and 5-95 for occupant sampling, along with standard calculations suggest sample sizes of 97 homes and 385 occupants, before apportionment across agencies.

Because the total number of SERC agencies (92) is relatively large, much larger than the number of WIPP grantees, we allocate the sample across strata, simply rounding to the nearest whole number rather than rounding up to the next whole number. As SERC agencies are regular WAP agencies, billing data will be handled for SERC agencies as for other agencies in the evaluation.

B.2.3.4. Under-Performer Study

"Under-performers" (or "over-performers") in the context of weatherization refers to weatherized homes whose savings turn out to be less (or more) than predicted on the basis of energy audits and after making weather and any other adjustments. By its very nature then, a study of under-performers is exploratory. Target populations of under- and over-performing homes will be identified through examinations of expected savings in conjunction with agency staff judgment. The population of homes so identified will be sampled by SRS stratified by agency, accounting for the PPS sampling of agencies. However, unstratified SRS is considered here as an approximation in determining a sample size.

Sampled homes will be inspected and many characteristics about them will be recorded. What is of particular interest are characteristics that can explain the discrepancy between the original prediction and the actual adjusted savings. That is, we seek additional variables that might serve as additional predictors in modeling energy savings.

So, consider the estimate of the slope coefficient B in a simple linear regression model

$Y = M + XB + \text{Error},$

where Y denotes adjusted energy savings, X is a generic potential new predictor variable, M is the model intercept, and B is the slope (coefficient) of X. The variance of the least squares estimate b of B can be shown to be

$V_Y/(n^* V_X)$

where n is the sample size, V_Y is the variance of Y (i.e., of the Error term), and V_X is the average of the squared deviations of the sampled X's from their sample mean.

Consider the test of the hypothesis B = 0 against the alternative hypothesis B = B* > 0. Let $Z_{1-\alpha}$ and $Z_{1-\beta}$ denote, respectively, the 1-α and 1-β quantiles of the standard normal distribution. Then the test that rejects if $b > Z (V_Y/(n^* V_X))^{1/2}$ has size approximately (for reasonably large n) α, and (it can be shown that) if $n = (V_Y/ V_X)((Z_{1-\alpha}+ Z_{1-\beta}) / B^*)^2$, then that test rejects with probability approximately 1-β. Without loss of generality, we can also express the alternative hypothesis in terms of $B' = B^* / (Vy/Vx)^{1/2}$.

If α and β are both 0.10, and B' = 0.30, we get $Z_{1-\alpha} = Z_{1-\beta} = 1.282$, and $n = 4(1.282/0.30)^2 = 73$ (rounded up to the next whole number). If α and β are both 0.10, and B' = 0.40, then n = 42. We use the former, larger value (73) as a sample size for under-performers and the smaller value (42) as a sample size for over-performers.

B.2.3.5. Air Conditioning Monitoring

Because of inconclusive studies and wide variability in estimated weatherization savings, it has been hypothesized that weatherization in warm-climate states does not achieve any air conditioning energy savings at all. In one air conditioner (AC) study, for a sample of 22 weatherized homes, the mean AC energy savings was –31 kWh, with a standard error of 167.2 kWh.[12] The mean AC savings for a sample of 19 comparison group homes was 106.7 with a standard error of 112.1. These results are consistent with the hypothesis of no AC savings. This study was conducted in Oklahoma, but results were similar in a study of AC savings in North Carolina.[13]

The object of the proposed AC study is to test the null hypothesis that mean AC savings in warm-climate states are zero against the alternative that the mean savings are positive, with a probability of at least .90 of detecting a savings of ten percent of the pre-weatherization AC consumption. The Oklahoma study's mean pre-weatherization AC consumption estimate combined for both the weatherized and comparison groups is 1,652.4 kWh, ten percent of which is 165.2 kWh.

[12] Ternes, Mark P., and Levins, William P. (1992), "The Oklahoma Field Test: Air-Conditioning Electricity Savings from Standard Energy Conservation Measures, Radiant Barriers, and High-Efficiency Window Air Conditioners," Oak Ridge National Laboratory, ORNL/CON-317, August 1992 (http://weatherization.ornl.gov/pdf/ORNL_CON_317.pdf).

[13] Sharp, T. (1994), "The North Carolina Field Test: Field Performance of the Preliminary Version of an Advanced Weatherization Audit for the Department of Energy's Weatherization Assistance Program.," ORNL/CON-362, June 1994 (http://weatherization.ornl.gov/pdf/ORNL_CON-362.pdf).

The proposed study will be conducted through weatherization agencies, the primary sampling units. As an approximation in reckoning sample sizes, we ignore the primary agency sampling, but it will be accounted for in the analysis. Assuming a level 0.1 one-sided hypothesis test, and using the Oklahoma study for preliminary estimates of the standard error and pre-weatherization AC savings, the sample size necessary for detecting a weatherization effect of 165.2 kWh or more can be estimated as follows.

From the Oklahoma study's weatherized and comparison group sample sizes (22 and 19) and standard errors (167.2 kWh and 112.1 kWh), it can be shown (using an F-test) that the sample variances are not significantly different. Therefore, a single (pooled) standard deviation will be assumed for the proposed study, and the same sample size will be assumed for both weatherized and comparison groups. The pooled standard deviation estimate is $[(22\times21\times(167.2)^2 + 19\times18\times(112.1)^2)/(21 + 18)]^{1/2} = 664.4$. The standard error for weatherized-comparison-group difference of mean AC savings can therefore be estimated as $664.4/N^{1/2}$, where N weatherized and N comparison group units are to be sampled (2N units in all).

The usual one-sided normal-theory test at the 0.1 level rejects the null hypothesis when the difference of means divided by the standard error (SE) of the difference exceeds the .90 normal quantile $Z_{.90} = 1.28$. For a true mean difference of 165.2 (i.e., ten percent of pre-weatherization NAC), P(Reject) =P[difference /SE > $Z_{.90}$] = P[(difference – 165.2)/SE + 165.2/SE > $Z_{.90}$] $\approx$ P[Z > $Z_{.90} - 165.2/(664.4/N^{1/2})$] = 1 – P[Z $\leq$ $Z_{.90}$ – $(165.2/664.4)N)^{1/2}$], where Z denotes a standard normal random variable. This implies N = 106. That, is 106 weatherized and 106 comparison group homes will be needed for the air-conditioned study. The sampling will be implemented by random sampling from air conditioned homes identified in the dwelling information data provided by a PPS-subsample of agencies from warm-climate states.

Prior experience[14] with in-home AC metering studies has shown that AC metering instruments may fail or be damaged up to twenty five percent of the time. This kind of non-response can be considered random and nonbiasing. To ensure an adequate sample size, however, increasing the sample size by 25% seems advisable. Thus 132 (106×1.25) homes will be sampled in each of the treatment and comparison groups (264 homes total).

B.2.3.6. Persistence Study

A study of weatherization persistence will be based on a comparison of a treatment group of homes weatherized circa 1995 with a comparison group of homes selected from recent WAP-applicant homes to match the treatment group on age and other characteristics. Sample sizes will be based on the primary comparison measurement, blower-door test Cfm50 (cubic feet per minute at 50 pascals of pressure) values. As no circa-1995 blower-door tests were conducted for comparison homes, a comparison group will be selected from homes for which blower-door tests are being conducted currently (i.e., circa 2011-2012).

[14] Ternes and Levins, op. cit.

The treatment-comparison group matching for the persistence study could be conducted in two ways: (i) The treatment and comparison homes could be matched on a case-by-case basis, with the data then analyzed, for example, with a paired t-test, or (ii) the treatment and comparison homes could be matched more loosely, with age and other distributions of the two groups kept the same, but without case-by-case matching or even the same sample sizes in the two groups, with the data then analyzed, for example, with a two-sample t-test. If case-by-case matching could be done precisely and with many characteristics, then approach (i) would likely lead to more precise comparisons. However, the looser approach (ii) is likely to be more feasible. Furthermore we have pilot study data to support approach (ii) only. Therefore we assume approach (ii) for reckoning sample sizes. If approach (i) is ultimately used instead, the sample size suggested here may be a little bigger than necessary.

The current, residual effect of the 1995 weatherizations will thus be measured as the difference between the averages of the current blower-door measurements for the 1995-weatherized (treatment) homes and the pre-weatherization comparison homes. Persistence can then be estimated by comparing that difference to the average of either the 1995 post-pre blower-door differences for the treatment homes or the current post-pre blower-door difference for the comparison homes. The average of the current differences for the comparison homes today is a better reference, however, as blower-door tests conducted in 1995 are considered to be less accurate than those tests conducted today.

We can use the Retrospective Evaluation's Indoor Air Quality (IAQ) Study for pilot data for calculating sample sizes. This data is summarized in the following table:

Table B.6. Indoor Air Quality Study Pre and Post-Weatherization Cfm50's

Variable	**N**	**Mean**	**Standard Deviation**
Pre-Wx Cfm50	2,324	3,466.7	1,829.28
Post-Wx Cfm50	2,251	2,362.5	1,003.41
Difference	2,220	1,093.9	1,177.64

From the table we see that IAQ Study blower-door Cfm50 flow rates were reduced by 1,093.9 Cfm on average after weatherizations. Fifteen years after weatherization we would expect the effect of the weatherization to be reduced from its initial effect. Thus we would like to detect a mean difference Δ between the treatment and comparison group blower-door Cfm50 measurements on the order of, say, Δ=250, 500, or 750 Cfm.

A sampling frame will be constructed from lists compiled from agency records. As an approximation we consider SRS from this frame. In practice sampling will be stratified and appropriately weighted to account for agency sampling.

To determine a sample size, we consider a statistical test of the null hypothesis Ho: Δ=0 versus the alternative hypothesis H_1: $\Delta > 0$, with the requirements that (1) if Δ=0, then the probability α that the test rejects (i.e., finds a difference) is low, say α=.10 or .05, and (2) if Δ=250, 500, or 750 Cfm, then the probability β that the test rejects is high, say .90 or .95. We use a two-sample Z-test as an approximation to the two-sample t-test.

Let N_W and N_C denote the numbers of treatment and comparison group homes (possibly but not necessarily the same), and let $\rho= N_W / N_C$. Let V_W and V_C denote the variance of the treatment and comparison group Cfm50's respectively. Then by straightforward calculation the smallest N_W that satisfies conditions (1) and (2) is

$$N_W = (V_W + \rho\ V_C)((Z_\alpha - Z_\beta)/\Delta)^2.$$

where Z_α and Z_β are the α and β quantiles of the normal distribution. Substituting the squares of the Pre-Wx Cfm50 and Post-Wx Cfm50 standard deviations in Table B.6 for V_C and V_W respectively then leads to the approximate necessary and sufficient sample sizes. For Δ=500 Cfm, α=0.10, β=0.90, and ρ=1 (same size treatment and comparison groups), we get $N_W = N_C = 114$ for a total of 228 homes.

B.2.3.7. Indoor Air Quality Remediation Cost Study

This study is a continuation of the ongoing Indoor Air Quality Component of the Retrospective Evaluation for homes sampled in that in that study that are found to require remediation.

B.2.3.8. Deferral Study

Ten states and ten WAP agencies will be sampled for this deferral study, and the deferral incidence and process (e.g., quality assurance) will be examined for a random sample of weatherized units from each sampled agency. Site visits to deferral homes will be made to verify deferral classifications. Agency selection will be purposive for agencies for which deferrals are understood to be troublesome. Inferences will therefore be restricted to the ten sampled agencies. However the ten sampled agencies will serve collectively as anecdotal evidence about the extent to which deferrals can be a problem, and the overall deferral rate for the ten sampled agencies will thus be a parameter of primary interest in the analysis.

We would like to estimate the overall deferral rate for the ten agencies to within five percentage points with 90% confidence. The following sample size calculation is based on unstratified simple random sampling, the stratification by ten agencies assumed to improve the survey precision slightly. In practice sampling will be stratified by agency with proportional allocation according to agency size.

Agency staff from some agencies have reckoned deferral rates to be possibly as high as 20%. We will assume that actual deferral rate is no higher than 25%. This implies that the variance of the overall deferral rate estimate is maximum when the deferral rate is 25% for each sampled agency, and the worst-case (i.e., maximum) variance of the overall deferral rate estimate is .25(1-.25) / N, where N is the total number of units sampled.

Hence, letting Z = 1.645 (95th percentile of the standard normal distribution), the worst-case required sample size for the 5-percentage-point-90%-confidence specification is

$n = Z^2(0.25(1-0.25)/(.05)^2) = 203$ units

Therefore we propose an overall sample size of 203 units.

B.2.3.9. Territories Survey

There is only one WAP agency in each of Puerto Rico and Guam. These agencies will be sampled, and utility billing data will be subsampled for them at the same 1-in-3 or 1-in-4 rates used for the States and DC in the main ARRA-period and previous WAP evaluations.[15] An occupants survey will also be conducted in Puerto Rico. Approximately 8,700 clients are to be weatherized in Puerto Rico. Following the approach above for occupants in the WIPP component, 258 occupants will be sampled, which is the minimum necessary so that the estimate of the proportion of satisfied occupants is within a 0.05 margin of error of the true proportion with 90% confidence.

B.2.4. Unusual problems requiring specialized sampling procedures

None.

B.2.5. Use of periodic (less frequent than annual) data collection cycles to reduce burden

The Weatherization Assistance Program evaluation is conducted occasionally, not annually. The previous evaluation is the Retrospective Evaluation, which is currently underway. The previous evaluation before that was in 1993, using data from the 1989-1990 program year. Although the Retrospective Evaluation is being conducted currently, the ARRA-period evaluation is necessary because of gross changes in WAP operational parameters (minimum income, maximum per-unit spending) made to accommodate greater ARRA-period funding.

B.3. <u>Describe methods to maximize response rates and to deal with issues of non-response.</u>

B.3.1. Maximizing response rates

With the exception of billing data sampling (utilities), nonresponse in the ARRA-period study is expected to minimal (see Footnote 3 for response rates observed during the retrospective evaluation). Because the great majority of agencies to be sampled have been sampled before and have responded in the Retrospective Evaluation, we are expecting an agency response rate higher than in previous evaluations. Otherwise, procedures for maximizing response rates will largely be as in Retrospective Evaluation. Subsampled

[15] Note that there is exactly one weatherization agency in each of Puerto Rico and Guam.

clients or staff that do not respond can generally be replaced with other clients or staff, so that non-response, though potentially biasing, does not affect ultimate sample sizes. Furthermore, because both staff and clients are beneficiaries of the program, and because staff and client contact information will be provided by the agencies, subsampled clients and staff have generally high response rates and non-response bias is minimal.

It should also be noted that several decisions, as explained in Item 12 of Supporting Statement Part A, will significantly reduce respondent burden. It is expected that the reductions in burden will reduce nonresponse.

Contacts with states, local agencies, and utilities have been or will be established to help promote the data collection process. Assistance will also be sought from professional organizations and, in the case of utilities, from regulatory commissions. Data requests will be designed to minimize the demand on respondents. Requests to the same party for multiple data installments will be coordinated to minimize the workload. Electronic data delivery will be encouraged, but data will be accepted in any standard format. Multiple follow-up requests will be made to agencies that have not responded. Web staff surveys will be made available for easy response. Computer-assisted telephone interviewing and callback scheduling will be used for occupant surveys.

B.3.2. Methods for dealing with non-response

The same procedures for accounting for non-response in the Retrospective Evaluation will also be employed in the ARRA-Period Evaluation. Most of the data analyses for ARRA-Period Evaluation will be to estimate per-stratum and overall totals, for example of energy savings, cost savings, numbers of clients (i.e., units weatherized) or satisfied clients, and proportions based on these totals, such as the proportion of satisfied clients among all clients. Ratio estimates of the totals will be computed by multiplying known population unit or funding totals by sample rate-per-unit or rate-per-dollar estimates. With this approach, observations that are missing and cannot be used to estimate rates are nevertheless accounted for when totals for all observations (including missing ones) are multiplied by corresponding rate estimates computed for the sample observed. Total estimates and related proportions are thus implicitly adjusted for non-response, as are their standard errors.

PPS sampling weights will also be adjusted to account for nonresponse. To the extent that non-responders and responders are alike, these adjustments completely correct for non-response. But of course non-responders and responders are not necessarily alike. If they aren’t alike, and if response rates are too low, non-response bias may be an important consequence. In this evaluation most of the sampled subjects (e.g., agency staff or clients) will already have interacted with the WAP program before they are sampled. For this reason and on the basis of the 1993 and retrospective evaluations, response rate are expected to be high for them.

Response rates for utilities are an exception. However, utility non-response is generally due to customer confidentiality and other issues unrelated to performance of WAP agencies or savings of weatherized homes and is thus generally nonbiasing.

Partial checks that non-response in nonbiasing will be made by (1) comparing responses of early and late responders and by (2) comparing responders and non-responders in terms of characteristics known without responses for subjects in both groups. These checks will be made as part of the data analysis.

B.4. Describe any tests of procedures or methods to be undertaken.

This evaluation will be very similar to the ongoing Retrospective Evaluation. Although some new populations are being surveyed (e.g., WIPP grantees), the questions being asked are essentially the same and have been extensively tested in the Retrospective Evaluation.

B.5. Provide the name and telephone number of individuals consulted on statistical aspects of the design and the name of the agency unit, contractor(s), grantee(s) or other person(s) who will actually collect and/or analyze the information for the agency.

Richard L. Schmoyer, Ph.D. (Statistics, 1980), of Oak Ridge National Laboratory (ORNL) developed the statistical components of the evaluation plan. He can be reached at 865-576-5327; ric@ornl.gov. ORNL will also provide oversight to the evaluation contractor (as yet to be determined) that administers the survey and performs data analyses.