Applications of Meta-Analysis to Methods of DSM Research

Daniel M. Violette, XENERGY, Inc.
Pamela A. Greene, RCG/Hagler, Bailly, Inc.
Shel Feldman, Wisconsin Center for Demand-Side Research
Philip Hanser, Electric Power Research Institute

Meta-analysis is a term used to describe a set of statistical techniques used to analyze, pool and leverage the results of research studies. While the term meta-analysis might suggest the use of sophisticated analytic methods, these techniques range from computing simple summary statistics to more complex analyses. In all instances, the objective is simply to efficiently extract the information from a group of related studies. Meta-analysis techniques use data and results from a number of studies to either answer the original research question with greater precision, or address new questions with the existing data. For example, applications of meta-analysis to DSM can estimate an average impact for a DSM measure across multiple studies, or produce a confidence interval for an estimated average effect size.

Two example applications are presented in the paper. These applications use representative data gained from literature reviews to illustrate potential gains from meta-analysis. The results used in these examples are not taken from actual sets of impact evaluations, and therefore, are illustrative only. The Wisconsin Center for Demand-Side Research is currently undertaking more rigorous meta-analysis studies using data from impact evaluations performed by Wisconsin utilities.

Entire books have been devoted to developing and presenting the equations and theory underlying the application of meta-analysis methods. In this short paper, the examples focus on what meta-analysis can accomplish and on the results, rather than on details of the calculation procedures; however these can be found in the referenced material.

Introduction

This paper discusses the applicability of a set of statistical techniques for pooling and leveraging the results of DSM evaluations. These techniques, termed meta-analysis methods, have the potential to:

- Increase the precision with which the impacts of DSM programs can be estimated;
- Accumulate and report information from multiple research studies using quantitative techniques;
- Provide information about factors that may influence program impacts across different applications; and
- Leverage and pool results from studies with small samples, such as end-use metering.

Other disciplines have used meta-analysis techniques. Applications are common in the fields of education, psychology, medicine, and even the physical sciences. For example, in physics, the Particle Data Group has been using meta-analysis techniques since 1957 to synthesize and aggregate findings on particle properties (Rosenfeld 1975). This information is used by physicists world-wide, with estimated values revised annually as additional data become available. Each year, the estimates become more precise. By compiling and analyzing data, and making both the data and the analyses available to interested parties, particle physicists have saved millions of dollars and years of research. It is possible that these same benefits may be available to DSM researchers. This paper introduces the concepts of meta-analysis and explores potential applications of meta-analysis to DSM research.

Meta-Analysis: An Overview

Meta-analysis is a term used to describe a set of statistical techniques used to analyze, pool and leverage the results of research studies. While the term meta-analysis might suggest the use of sophisticated analytic methods, these techniques range from computing simple summary statistics to more complex analyses. In all instances, the objective is simply to efficiently extract the information from a group of related studies.

Traditional literature reviews typically involve qualitative examinations of the results of a number of studies, and a comparison of their trends and conclusions. Rather than provide a subjective review of studies, meta-analysis methods impose systematic standards and increased rigor on research efforts aimed at synthesizing research results. These criteria are not dissimilar from the standards imposed on the original research. One stated purpose of meta-analysis methods is the resolution of the dichotomy between scientifically rigorous original research and then the largely unsystematic, mostly subjective reviews of this research. Even calculating the average impact estimate from a group of studies examining the same DSM measure raises both analytical and statistical questions that most often are ignored in traditional literature reviews. Meta-analysis provides a means of holding the "second users" of the data as accountable for their methods and conclusions as were the first (Cooper and Hedges 1992).

Using meta-analysis, the reviewer can draw conclusions about relationships that appear in the literature and can characterize the strength of these relationships. The reviewer can also describe the extent to which study characteristics and DSM program characteristics may affect the phenomenon (e.g., program impacts) being examined. For example, given sufficient data, meta-analysis can characterize the extent to which programs for lighting and programs for refrigerators have similar effects on energy consumption per dollar spent.

Meta-analysis is a set of techniques that use data and results from a number of studies to either answer the original research question with greater precision; or address new questions with the existing data. Meta-analysis is essentially a statistical analysis of findings from a number of empirical studies. Applications of meta-analysis to DSM can:

- Estimate an average impact for a DSM measure across multiple studies;
- Produce a confidence interval for this estimated average impact;
- Combine study results to generate an overall probability for the existence of an effect;
- Compare the observed variability in "impacts" or "effect sizes" to the variability that would be expected if sampling error alone were operating; and,
- Explain variability in effect sizes through study characteristics, research designs, or program characteristics.

Meta-analysis techniques account for the fact that some of the assumptions underlying traditional inferential statistics may be violated when the data used in the analyses are results from two or more related studies. Conventional statistics rely on the assumption that observations are drawn from a population with the same underlying distribution, and hence, the same variance. This assumption, termed the homogeneity of variances, may be violated when the data are the results of different studies, with each study having different sample sizes and different variances for the estimates produced. Meta-analysis techniques do not rely on the assumption of homogeneity of variances.

Conditions under which meta-analysis is best applied include: (1) a set of studies address the same question; (2) the individual study comparisons test the same conceptual hypothesis; and (3) the studies are independent analyses. The meta-analyst typically will use judgment to determine the independence of studies; for example, if two studies are conducted by the same researcher using the same methods, but with different treatment groups, are those studies independent? One criterion that can assist with this determination is whether the initial assumptions made by the primary researchers are valid. If they are not, and/or if they are found to bias the study results in some way, then the study should be either corrected (when possible) or excluded from the meta-analysis.

The **effect size** has traditionally been the primary variable used in meta-analysis. In the context of DSM evaluation, the effect size is the magnitude of the DSM program impact. Most systematic relationships in meta-analysis involve estimating a population effect size. The effect size for the population is most often described as the "standardized" difference between the means of the treatment group (μ_t) and the control group (μ_c) normalized by the standard deviation (σ), often estimated, for the population, i.e.:

Standardized effect size $= (\mu_t - \mu_c)/\sigma$

Standardizing effect sizes puts variables that are measuring essentially the same thing on the "same scale," making them comparable across studies.

In the case of DSM literature, the variables of interest are often already on the same scale (e.g., kWh or kW reduction in demand). Thus, they may not need to be standardized unless there is some other reason to believe that the effect sizes are not directly comparable.

Meta-Analysis Techniques

Meta-analysis methods typically are categorized into techniques that address three different sets of questions: (1) How consistent are the estimates or findings across studies; (2) Where results of studies do vary, are there factors that explain the variation; and (3) What is the best way to use the information from a set of related studies to produce a combined or average effect estimate? The techniques used to address these questions are termed diffuse tests, focused tests, and synthesis procedures in the meta-analysis literature (Rosenthal 1984). Each is discussed below.

Consistency of Results - Diffuse Tests

Suppose four studies were conducted to determine the energy impact of residential weatherization programs. Three of these studies indicated that statistically significant savings resulted from the program, while one indicated that the savings were not significant. A reviewer conducting a traditional literature review might conclude that residential weatherization programs probably have a significant effect on energy use, but that some study findings were contradictory.

A diffuse test can be used to compare the size of the energy savings and determine whether observed effect sizes differ significantly among themselves and/or whether they consistently reject the studies' null hypotheses. This meta-analysis test could produce several conclusions:

First, it is conceivable that the energy savings found in all of the studies do not differ significantly among themselves, even though three of the studies found that energy savings to be significantly different from zero, and one did not. Suppose the three studies with statistically significant estimates produced savings estimates ranging from 300 to 600 kWh per year. The fourth study estimated energy savings to be 200 kWh per year, which was found to be not significantly different from zero. The diffuse test could reveal, however, that the energy savings found in the three significant studies do not vary significantly from that found in the nonsignificant study. In other words, the findings in the fourth study are not inconsistent with those in studies one, two and three. In this case, the estimated effect sizes across all four studies are said to be homogeneous.

Second, meta-analysis might show that, among a set of studies all finding statistically significant results, the estimates from the studies may still be heterogeneous, i.e., the results are found to differ significantly across the studies. For example, suppose the average energy savings estimated by four studies of weatherization programs ranged from 300 to 1000 kWh. All four studies rejected the null hypothesis of no savings. A diffuse test can reveal, however, that the savings estimates found in the studies vary significantly when compared to what would be expected from sampling error alone. The meta-analyst would then use focused tests to try to determine if the effect sizes differ in a systematic way.

Third, the two types of findings outlined above lead to a third potential conclusion, which is that the estimated effect sizes from a set of studies all showing limited statistical significance, when taken together may actually produce consistent, relatively precise estimates. Meta-analysis methods for combining study results may increase the statistical significance and the confidence in a given relationship. This will be illustrated in the example applications, but to provide a simple illustration, assume that three independently conducted studies all estimate program savings to be 300 kWh. Also assume that each study produced limited statistical significance, i.e., there is only a 50 percent probability that the impacts are significantly different from zero. Now, a meta-analysis approach would ask the following question: "What is the likelihood that these three studies all could have been conducted and produced these results, if the savings were in fact zero?" If the studies can be assumed to be independent, then the answer is the probability of a significant finding for each study multiplied by the number of studies. In this case, this is .5 x .5 x .5 or .125. Given the results of all three studies, there is only a 12.5 percent probability of the null hypothesis being true, and an 87.5 percent probability that program impacts do, in fact, differ significantly from zero.

Examining Variation in Study Results - Focused Tests

Meta-analysis can explore whether study results show systematic variation across the entire set of studies or within subgroups of studies. Focused tests are used to identify factors associated with variation in effect sizes. Independent of whether effect sizes are found to be homogeneous, there still may be some systematic variation in study results. Thus, meta-analysis can be used to test specific hypotheses concerning relationships between effect sizes and factors such as research design or program characteristics.

As an example, suppose that a set of impact evaluations of residential weatherization programs used different methods for estimating energy savings. Of the four studies, two used a cross-sectional treatment/control group research design, while the other two used a time-series research design where energy use of participants before and after participation in the program was compared. A focused test can be performed to examine whether the study designs are associated with variation in effect size estimates; then a diffuse test for homogeneity within each research design sub-group can be applied. The analyst may discover that the effect sizes are heterogenous between groups, but homogeneous within these subgroups.

Techniques for Combining Studies

These methods involve estimating an average effect size across a set of studies, and a confidence interval for this estimated combined effect size. The average affect size indicates the overall "trend" or "result" of the entire set of studies. Combined effect sizes can be calculated for both homogeneous and heterogeneous sets of studies; however, combining heterogenous effect sizes may mask factors causing variation in effect sizes, and need to be viewed with care.

Steps in a Meta-Analysis

Four steps are common to most meta-analyses.

Step 1: Issue Identification

This involves identifying the question to be addressed. Often, the effect of a change to the *status quo*, such as the implementation of a DSM program is the issue to be examined. Meta-analyses can:

- Combine results of multiple studies to better estimate the impact of a specific DSM measure or program.

- Determine if different program characteristics result in different outcomes, e.g., program impacts or free ridership.

- Determine the effect of study design on outcomes, e.g., are self-reported estimates of free-riders different from those obtained through discrete choice analysis?

Step 2 - Study Identification and Collection

This step involves retrieving studies and establishing criteria for determining relevant studies. The retrieval of reports on DSM programs may not be straightforward. First, all reports should be retrieved, not just those with significant results. Studies not producing significant results, or studies showing results in the "wrong" direction, may not be reported. This is called the "file-drawer problem." To produce unbiased estimates, these studies should be included in the meta-analysis. Their exclusion can bias the results toward significance. Second, more studies provide more data and results, allowing the analyst to be more confident about the conclusions. It is important to make an effort to obtain all applicable studies.

Step 3 - Data Extraction, Effect Size Calculation, and Significance Level Determination

The data required to calculate an effect size or significance levels must be extracted from the report. The types of data that are needed for these calculations include sample sizes, sample means, standard deviations, and any significance tests and probability levels (p-levels) that are provided.

Information on the design of the study or data that can help explain variations in effect sizes or significance levels are also useful in meta-analyses. These data can include:

- Report author or sponsor;
- Report date;
- Sampling procedure;
- Description of sampled units (participating customers, dealers);
- Demographic information for the sample;
- Characteristics of the participants;
- Study design;
- Type of program being evaluated; and
- Description of program details (type or level of incentive, eligible end uses, etc.).

These data may provide information that will help explain any heterogeneity in study results.

Step 4 - Analyzing Effect Sizes and Significance Levels

Two classes of meta-analysis techniques have been developed. One class of methods is aimed at analyzing effect sizes, while the other is structured to analyze significance levels and probabilities. These analyses include the application of diffuse and focused statistical tests, and meta-analysis methods for combining study results.

Example Meta-Analysis Applications

Two example applications are presented here. These applications are more fully documented in EPRI (1992). These applications use representative data gained from literature reviews to illustrate the potential gains that might stem from meta-analysis. The results used in these examples are not taken from actual sets of impact evaluations, and therefore, are illustrative. The Wisconsin

Center for Demand-Side Research is currently undertaking more rigorous meta-analysis studies using data from impact evaluations performed by Wisconsin utilities.

Entire books have been devoted to developing and presenting the equations and theory underlying the application of meta-analysis methods. In this short paper, the example applications focus on the results of meta-analysis, rather than on the calculation procedures.

Example Application #1 - Residential Weatherization Results

Data for this example are shown in Table 1. This table portrays data from nine studies. All studies but one (study 9) used a control group. Also, all studies but one reported the standard deviation of the estimated program effect. Without information on the standard deviation, few meta-analysis applications are possible.[1] In Table 1, the effect size is simply the difference between the change in energy use for program participants and the change in energy use for non-participants. The effect size variances shown in Table 1 use the standard equation for calculating the variances for the difference between the two means.[2] Confidence intervals for the estimated effects for each individual study can be calculated for seven of the ñine studies. These are presented in Table 2.

The combined effect size for "k" studies is calculated from the following equation:

$$d_{comb.} = \frac{\sum_{i=1}^{k} \frac{d_i}{v_i}}{\sum_{i=1}^{k} \frac{1}{v_i}}$$

where di and vi are the effect size and the variance for study i. Using this formula, the combined effect size shown in Table 2 is calculated to be 337 kWh.

A confidence interval for this combined effect estimated can be calculated using the variances of the combined effect size estimates. This variance is:

$$v_{comb.} = \frac{1}{\sum_{i=1}^{k} \frac{1}{v_i}}$$

A confidence interval is calculated of $d_{comb.}$ as:

$$v_{comb.} \pm Z_\alpha \sqrt{v_{comb.}}$$

where Z_α is the 100 (1-α) percent two-tailed critical value for the standard normal distribution. Using these equations, the combined effect size estimated and the resulting confidence interval using meta-analysis methods are 337 kWh ± 25 kWh, for a precision level of ± 7 percent. As can be seen from Table 2, the confidence interval for the meta-analysis combined effect estimate shows a substantial increase in precision when compared to the

Table 1. Effect Size and Variance Calculations Example #1 - Residential Weatherization Studies

	Treatment Group			Control Group				
Study	kWh Change (mean)	Standard Deviation	Sample Size	kWh Change (mean)	Standard Deviation	Sample Size	Estimated Impact	Variance
1	-300	150	4	50	100	8	350	6875
2	-275	250	25	60	100	20	335	3000
3	-170	160	70	200	75	68	370	448
4	-225	200	10	65	70	10	290	4490
5	-250	300	100	35	90	82	285	999
6	-350		85	-60		90		
7	-400	170	75	-50	150	62	350	748
8	-310	180	40	0	70	45	310	918
9	-325	130	60					

Table 2. *Calculated Confidence Intervals for Estimated Effects*
Example #1 - Residential Weatherization Studies

	Estimated Savings (kWh) and	Precision
Study	95% Confidence Intervals	Level
1	350 ± 163 kWh	± 46%
2	335 ± 107 kWh	± 32%
3	370 ± 42 kWh	± 11%
4	290 131 kWh	± 45%
5	285 ± 62 kWh	± 22%
6	Not Available	
7	350 ± 54 kWh	± 15%
8	310 ± 59 kWh	± 19%
9	Not Available	
Meta Analysis Combined Effect	337 ± 25 kWh	± 7%

individual studies. The average precision for the individual studies is ± 27 percent, where the precision level for the meta-analysis combined estimate is ± 7 percent.

A "diffuse test" that examines the homogeneity of the effect sizes across a group of studies uses the following test statistic:

$$H_T = \sum_{i=1}^{k} \frac{1}{v_i}(d_i - d_{comb.})^2$$

This is simply the weighted sum of squares of the effect size estimates d_1, ..., d_k about the combined effect size estimate [See Hedges (1982), pages 490-499; and Rosenthal and Rubin (1982) pages 500-504]. The statistic H_T has approximately a chi-square distribution with k-1 degrees of freedom. If all studies do not have a common effect size, then H_T will tend to be larger than expected under the condition of homogeneity. Thus, the test rejects the homogeneity of the effect sized at significance level α if H_T exceeds the $100(1-\alpha)$ percent critical value of the chi-square distribution with k-1 degrees of freedom [See Hedges and Becker (1986) page 34].

Example Application #2 - Commercial Programs

The data for this example are presented in Table 3. In this case, standard deviations are not reported for two studies (#3 and #9). Also, these data come from three types of commercial programs - audit, rebate, and interruptible rate programs.

The diffuse test for homogeneity produces a test statistic of 467, exceeding the critical value and resulting in a rejection of the hypothesis of homogeneity across study results. Focused tests that examine homogeneity within study subgroupings can be used to determine if the estimated effects are homogeneous within the different program types. This calculation showed that the estimated energy savings within each of the program groups is more homogeneous than for the total set of studies. Significant differences between the savings estimates for each program group were found as well.

The combined effect estimates for the audit and rebate program subgroups are presented in Table 4. Table 4 also compares the combined estimate and variance t to those of the individual studies. The variances for the meta-analysis combined effect estimates for both programs are smaller than the variance of any individual study for each program. Again, the application of meta-analysis methods that combine information from multiple studies increases the confidence in the estimated program savings.

Conclusion

Meta-analysis methods can be used by DSM researchers to enhance the information provided by impact evaluations. They can provide increased confidence in the estimates of DSM program accomplishments and provide

Table 3. Effect Size and Variance Calculations Example #2 - Commercial Programs

	Treatment Group			Control Group					
Study	kWh Change (mean)	Standard Deviation	Sample Size	kWh Change (mean)	Standard Deviation	Sample Size	Estimated Impact	Variance	Program Type
1	-600	400	30	50	100	45	-650	5556	Rebate
2	-430	250	75	60	100	75	-490	967	Audit
3	-1000		50	35		47	NA	NA	
4	-500		100	200		78	-700	648	Rebate
5	-100	250	30	50	175	10	-150	5146	Interrupt
6	-375		5	-20		5	-355	19620	Audit
7	-300	170	65	85	100	75	-385	578	Audit
8	100	180	45	0	70	40	100	842	Interrupt
9	-250	195	15				NA	NA	

Table 4. Estimates and Variances - Combined vs. Individual Example #2 - Commercial Programs

I. Audit Studies	Estimated Savings	Variance
2	490 kWh	967
6	355 kWh	19,620
7	300 kWh	578
Meta Analysis Combined Effect	423 kWh	355
II. Rebate Studies	Estimated Savings	Variance
1	600 kWh	5,556
4	700 kWh	648
Meta-Analysis Combined Effect	695 kWh	580

a better understanding of why different program designs produce different outcomes. These procedures provide a consistent approach for accumulating information from multiple evaluation studies examining similar DSM programs. Meta-analysis methods can produce a consolidated estimate of program and/or measure impacts using information from several DSM evaluations. This combined estimate takes advantage of information gained from conducting multiple studies on a single program type and usually is more precise than the estimates from any single study, i.e., the confidence interval is generally narrower. The influence of different evaluation methods on estimates of impacts can be analyzed to see if they contribute to systematic differences in study results. Finally, these procedures can be used to examine how different program attributes and designs influence the resulting impacts from the program. In conclusion, meta-analysis methods have made significant contributions to other scientific disciplines and it can be expected that similar contributions may be possible by applying these methods to DSM evaluation research.

Acknowledgments

The project on which this paper is based benefitted from the critical advice of many professionals in the field of DSM and meta-analysis. Of special value were the reviews of Dr. Harris Cooper, University of Missouri, Dr. Larry Hedges,. University of Chicago, and Dr. Robert Rosenthal, Harvard University.

Endnotes

1. If a t-statistic is reported, then the standard deviations can be "backed out" from the statistic.

2. This equation is simply the square of the standard deviation of the treatment group (sd_t) divided by the sample size (n_t), plus the standard deviation of the control group squared divided by it's sample size, or

$$\frac{(sd_t)^2}{n_t} + \frac{(sd_c)^2}{n_t}$$

For study 1, the effect size variance is

$$\frac{150^2}{4} + \frac{100^2}{8} = 6{,}875$$

References

Cooper, H. M., and L. Hedges, forthcoming, 1992. *Handbook of Research Synthesis*, Manuscript. Sage Publications, Newbury Park, CA.

Electric Power Research Institute, 1991. *Impact Evaluation of Demand-Side Management Programs: A Guide to Current Practice*. EPRI CU-7179, Volume 1.

Electric Power Research Institute, 1991. *Impact Evaluation of Demand-Side Management Programs: Case Studies and Applications*. EPRI CU-7179, Volume 2.

Electric Power Research Institute, 1992. *Review of DSM Program Evaluations: Database of Evaluation Findings and Approaches for Synthesizing Research Results*. EPRI RF-3269, Volume 1.

Hedges, L. V., and I. Olkin, 1985. *Statistical Methods for Meta-Analysis*, Academic Press, Orlando, FL.

Hedges, L. V., and B. J. Becker, 1986. "Statistical Methods in the Meta-analysis of Research on Gender Differences." *The Psychology of Gender: Advances through Meta-analysis*, J. S. Hyde and M. C. Linn, ed, The Johns Hopkins University Press, Baltimore, MD.

Hedges, L. V., 1982. "Estimation of Effect Size from a Series of Independent Experiments." *Psychological Bulletin*, Vol. 92, No. 2, pp. 490-499.

Rosenfeld, A. H., 1975. "The Particle Data Group: Growth and Operations--Eighteen Years of Particle Physics." *Annual Review of Nuclear Science*, Volume 25, pp. 555-598.

Rosenthal, R., 1984. *Meta-Analytic Procedures for Social Research*. Sage Publications, Newbury Park, CA.

Rosenthal, R., and D. Rubin, 1991. "Further Issues in Effect Size Estimation for One-Sample Multiple-Choice-Type Data." *Psychological Bulletin*, Vol. 109, No. 2, pp. 351-352.