Best Practices in Measure Retention and Lifetime Studies: Standards for Reliable Measure Retention Methodology Derived from Extensive Review

Lisa Skumatz, Ph.D., Skumatz Economic Research Associates, Inc., Superior, CO, USA
John Gardner, Skumatz Economic Research Associates, Inc., Superior, CO, USA

ABSTRACT

In work for a large west coast agency, the authors[1] reviewed more than 94 measure retention/lifetime studies to assess conformance with prescribed protocols, methods, quality, and the justifiability of results and conclusions. The studies covered measures for a wide array of programs in the residential, low income, commercial, industrial, agricultural, and military sectors. For each study, the consultant team conducted an exhaustive review of: (1) program information, share of program savings covered by the analysis, measures included, and other topics related to justification and context for the studies; (2) sampling methodologies, sample quality, and justification; quality of field work, including data collection approach, treatment of sample, quality of program records, and field work practices; and (3) data validation and verification, treatment of sample attrition and sample, statistical approach, consideration of alternative models and treatments; and justifiability of the results.

In conducting the review, we found considerable variation in the quality and conduct of the studies. We reviewed the range of practices used, and identified what we view as "best practices" in the field of retention analysis. The results have implications for others conducting retention analyses, as well as those considering or revising protocols or standards related to these studies.

Introduction

Retention studies, also known as measure life studies, are a critical and highly useful component of energy efficiency research. Despite the various data collection and treatment methodologies employed, the fundamental purpose of all measure retention studies is to estimate the amount of time that a measure will be in place.

The overall approach taken by most measure retention studies in the energy efficiency (EE) field is to estimate the median effective useful life (EUL) of the measure in question. The EUL is usually defined as the median number of years[2] that a measure is likely to remain in-place and operable.[3] This amount of time is often calculated by estimating the amount of time until half of the units are no longer in-place and operable.

While this task may seem straightforward at first glance, there are often considerable complications involved with obtaining EUL estimates. Measures often last for a long time, making it impractical to simply wait until half of the units fail in order to determine the median survival time. Measure lives are also frequently interrupted prematurely by the owners or employees of the residence or business in which the measure was installed. Obtaining unbiased EUL estimates, therefore, can require delicate statistical analysis to control for exogenous factors that might affect measure lifetime

[1] Scott Dimetrosky and other staff from Quantec, LLC, a subcontractor on the assignment, conducted a number of the reviews. A review of realization studies was also conducted. Subcontractors Northwest Research Group and NAA / Emcor assessed the quality of data collection work for the studies.
[2] Or other time interval, as appropriate.
[3] "In-place and operable" is at least the most common definition of measure survival. Depending on the specific measure under inquiry, alternative formulations of the definition may be more appropriate.

and to predict measure lifetimes based on empirical data. Furthermore, applications for this work require projected results fairly early into the lifetime of much of the equipment installed as part of various programs, when a set of measures is young and only a relatively small portion of the installations may have failed. For example, protocols in California require periodic verification of EULs when measures have been installed for fewer than five years. While important, this poses a particular challenge, as EUL estimates are driven off failures and few measures projected to last 20 years or more would be expected to fail under that schedule. Developing unbiased estimates of EULs under circumstances of limited data early in measure lifetimes is particularly challenging.

This paper presents a set of best practices for measure lifetime and retention studies derived from the authors' experience evaluating more than 90 studies covering a diverse collection of energy efficiency measures. We were able to compare the different data collection, treatment, and analysis techniques used by various studies on the basis of their effectiveness in obtaining meaningful results, their ability to produce to produce reasonable EUL estimates, the degree to which they produced statistical models that fit the data and the defensibility of the conclusions drawn from them. The review of a large number of studies provided an opportunity to view the range of practices used for small and large, and simple and complex measures and programs over a period of nearly ten years. Examples of strong practices and problems encountered by other studies are presented in the text. The next sections include: a discussion of best practices in data collection; a parallel discussion of data treatment and analysis practices; a brief discussion of the most common mistakes that we encountered during our review; and a summary of an assessment of the results of an application of these "best practices" to lifetime studies conducted in California since 1994.

Data Collection Practices

As with any empirical study, effective data collection is absolutely crucial to the effective execution of a measure retention study.[4] Retention studies present unique problems that make accurate data collection particularly exigent. Such studies are by necessity longitudinal, and since some measures may last longer than ten years on average, surveys and inspections might need to be spaced out over intervals of five or even ten years. This requires careful planning and coordination on the behalf of the researchers making the data collection effort.

Population Source and Instrument Type

The first step to conducting a successful data collection effort for a retention study is to obtain a high quality/well-documented population list, if possible. Naturally, the feasibility of this step may vary depending on the situation. Measures that are installed through an energy efficiency program, particularly a publicly funded program, are likely to require recipients to give detailed contact information; however, this may not be the case when tracking measures are purchased independently or through purely commercial channels. It is important to avoid using a population source list that contains a built-in bias such as warranty or registration cards sent in to a manufacturer. If multiple measures are under investigation, or if measures are distributed as part of a comprehensive program that installs different measures in different locations based on perceived recipient needs and usage, a measure-based population list may be preferable to a location-based population list, because sampling locations might lead to bias issues (e.g. clustering). Retention studies are often dependent on program records as the key source for the population. To support retention analysis, several key items are critical to a

[4] Note that this paper assumes the measures and the measures required for study are already selected. In California, the protocols require EUL work on measures sufficient to account for more than 50% of savings. So measures in HVAC, lighting, and other key end uses are relatively common. However, as a consequence, there are multiple studies of refrigerator lifetimes, but none of a variety of other less-frequent measures. It may be appropriate to modify procedures to make certain un-studied measures receive attention, perhaps before more new studies of previously much-estimated measures.

successful retention analysis: participants, measures/practices, and dates. Lifetime analyses examine failures based on elapsed time since installation, and without good records on elapsed time since particular measures or practices were installed, the analysis is critically hampered.[5]

Once the appropriate population source list has been obtained, a data collection strategy must be formulated. In our experience, data collection takes one of three forms: (1) a survey (usually telephone, but occasionally mail), (2) on-site inspections or audits, or (3) a combination of inspections and surveys. While no single data collection method is clearly superior, retention studies for certain types of measures are better served by one method than another. For large measures for which only one is usually installed in a location such as residential refrigerators or central air-conditioning systems, a telephone interview may be completely appropriate. These types of large measures are unique and easily identified by residents or businesses. Recipients of these large-scale measures are unlikely to have forgotten about the measure, and can usually say with accuracy whether the measure is still in place and operable and, if not, when it ceased to be so.

Other measures are not as memorable or unique. For example, a store manager receiving thirty compact fluorescent light-bulbs or an office in which hundreds of T8s are installed is not very likely to be able to give an accurate account of the operating status of each individual bulb. For such a situation, an on-site interview, in which technicians physically visit the installation site to determine or verify the status of the measure, is far more appropriate (and more accurate).

Given the importance of dates for failures in the analysis, several suggestions have been made to gather accurate data and facilitate the use of less expensive telephone surveys versus more expensive data collection approaches. For CFLs and a number of other measure types, it has often been suggested that "tags" be affixed to the measures requesting a telephone call to report equipment removal. In theory, this should help provide a more accurate data series than a poorly recalled date gathered "after the fact". However, as far as the authors could determine, this had not been implemented in any of the studies reviewed.

Sampling Strategy

Once the population source and the data collection method have been determined, an appropriate strategy for sampling must be chosen. As in any empirical research, the ultimate goal of sampling for a retention study is to collect a sample that is representative of the relevant population. The measure population is usually known and a number of steps are needed to obtain a representative sample.

If the total number of installations of the measure in question is small, a census may be conducted. Gathering a census obviates the need for many sample adjustments (e.g. weighting, stratification, standard-error adjustments, etc.) that might be necessary if a probability sample is conducted. However, it is not always possible to conduct a census of the measures that have been installed. Moreover, even where a census is technically possible, if many measures have been installed, a census may not be cost-effective. In such cases, development of a well-designed sample is appropriate.

The most thorough retention studies attempt to stratify their samples based on obvious characteristics of the population. The appropriate strata to use depend on the measures being installed and the recipients of those measures. Weatherization measures, for example, vary in their effectiveness and longevity depending on the climate in which they operate. A sample of climate measures should therefore be stratified based on the climate zone distribution of the population. Especially in the commercial and industrial sectors, site energy demand may play a large role in the operation of energy efficiency measures. Studies that define peak kWh demand bins and sample based on energy demand strata, therefore, are more likely to obtain reliable median EUL estimates. Other common strata used in the studies we reviewed included sector (industrial, commercial, agricultural, residential) and business

[5] We found at least one example in which critical information was missing from the program records, and as a result, there was little that could be done by the retention study to provide a reliable, high quality analysis.

type. Operating hours may also be a critical sampling strata, which may or may not be reflected in business type.

Another decision that researchers must make when designing a retention study is whether to conduct a measure-based sample or a site-based sample. There are valid reasons to choose either strategy. Measure-based strategies are more likely to result in a random sample of the measures installed, therefore avoiding potential biases caused by exogenous site-related factors. A measure-based sample may not always be possible, though. In some cases, program tracking data will only contain information about participants, or may be set up in a way that makes it very complex to sample on a measure basis. If this is the case, a measure-based sample would require a pre-sample audit of every participant location in order to determine the population. Such an audit may not be possible, particularly in residential programs, and is likely to be expensive.

If a measure-based sample is not possible, a two-stage sampling method might be the best solution. First, a random sample of locations is drawn. From this sample, another random sample of the measures installed at the chosen populations' location is taken. Such a strategy may help avoid possible biases associated with site-based sample. However, the distribution of measure installations throughout the population of sites may be far from uniform. Because measure failure can occur for non-technical reasons (e.g. the measure was removed because the proprietor of the business did not like it), and because some measures may be interdependent (e.g. a bulb and a lamp), sampling at the site level – even if accomplished through a two-step procedure – may result in biases such as clustering. While it is possible to treat these effects in the data analysis process, it is easier to avoid them from the start with a well-designed sampling strategy.

However, site-based sampling has several advantages, particularly if on-site approaches are used. The statistical work for a retention study is driven by failures, and given that no fewer than one measure is installed at any one site, sampling a full "site" can increase the number of measures (and failures) identified for each observation/data point. To further increase cost-effectiveness, the number of measures surveyed could be maximized by over-sampling from sites with the highest number of measures installed. This may be tempting, but would not be an appropriate approach unless it can be determined that the failures of measures would be unrelated to the number of measures installed. However, if this case can be credibly made, the costs for data collection can be reduced considerably.

A sufficient number of measures must be sampled; if regression analysis is to be used to obtain EUL estimates, there must first be enough valid observations for convergence. When retention studies are being conducted for regulatory and other institutions, they often need to meet pre-established standards of statistical accuracy. Wherever possible, necessary sample-size calculations should be performed and every effort should be taken to meet or exceed this number of surveys.

When phone interviews are used for data collection, we have found that a few simple steps greatly increased the response rate and sample size. The most effective strategy was to employ a scheduler to call in advance, explain the purpose of the study to participants, and schedule a mutually agreeable time to conduct an interview. Studies that did not take this approach, but still had a high response rate, used at least three or four callbacks (and often five) before making a replacement, and in addition, the calls varied in time of day and day of week. When replacements must be made, they should be taken from the same strata as the original (if stratification is being used).

There are a few sampling concerns that are especially relevant to retention studies. As mentioned previously, retention studies are by necessity longitudinal. Consequently, data attrition must be carefully addressed. Over the course of a ten-year retention study, significant attrition can occur that is beyond the control of the research team. Businesses can close, for example. In one study we reviewed, a house containing some of the measures sampled in the initial study was destroyed by a fire. While this type of attrition is by no means preventable, it may lead to large biases – especially if a site containing a disproportionate number of measures suddenly becomes inaccessible. If such attrition occurs, any research output should make explicit mention of it and try to account for the biases that it caused.

Another issue is movers and non-movers. Depending on the context of the study, an operable measure that has been moved to another location may be counted as a failure.[6] In such cases, careful precautions should be taken to correctly code and record the status of measures. In the studies that we reviewed, the flagship example of this issue was residential energy efficient refrigerators. If a recipient moved – taking their refrigerator with them to an area outside the domain of the study – the measure needed to be counted as a failure. However, if they left their refrigerator for their house's next inhabitants, the new owners needed to be contacted in order to determine if the refrigerator was still operable. New residents, however, may not be as acutely aware of the status of the refrigerator, and may not be willing to participate in the survey at all, since they were not part of the program through which it was installed. If the original owners of the refrigerator had moved somewhere else within the study domain, taking the appliance with them, they needed to be tracked down and contacted – a difficult task that resulted in lower response rates.

Given that protocols in California require periodic conduct of retention studies, it may be appropriate to develop a well-designed panel approach. A panel provides several advantages for retention studies.[7] The first round of visits can help identify program-installed measures – a task that becomes more difficult as time elapses. They may "tag" measures, or may develop labels or maps of the location of program-installed measures on-site. Earlier observations at the same site can also assist when measures are later identified as missing, but the occupant cannot recall the date of failure. Information from previous surveys can be reviewed, and the date of the previous visit provides a bound for the failure date of the measure; we know the measure did not fail before the date of the previous visit. This is considerably more information than is known if a new sample is visited each time – in that case, it would have been unknown whether the measure failed as early as a day after initial installation. This significantly improves the data for the retention work. Contact is simpler, reducing survey costs. If it is believed that frequent contact does not affect failure (unlikely) or removal (a less straightforward case), more frequent contact may be implemented to gather even more accurate dates and retention information. Of course, appropriate replacement procedures must be devised for the panel approach.

Surveys, Inspections, Fieldwork and Validation

Careful data validation, regardless of whether phone or site surveys are used, is of fundamental importance. Especially if the measure population is small, a handful of bad data points might result in inaccurate EUL estimates. Worse, if validation and fieldwork are inconsistent between years, EULs may prove inestimable due to data incompatibility.

The critical information to be gathered includes, by measure, whether the equipment is still in place and operable, and the date of failure for any equipment that is not. For measures no longer in place and operable, information should be recorded describing the reasons for removal, or a description of the situation or rationale of the removal. Given the relatively small number of removals, collecting these additional data elements is not onerous. Additional data may be collected, depending on the modeling expectations. If operating hours or other factors are important exogenous factors expected to affect lifetimes (e.g. operating hours for light bulbs, air conditioning or HVAC equipment; climate for HVAC, etc.) then these data must be collected as well.

All data collection instruments should be pre-tested. If conditions affecting measure "operability" arise that cannot be recorded by an interviewer or auditor, data analysts may be unaware of these conditions and wouldn't be able to correct for them. Since retention studies frequently cover multiple measures, testing should be done to determine the best instrument for each one, since different measures may be sensitive to different conditions, and may require different information before they can be analyzed.

[6] This might arise, for example, if the purpose of a retention study is to estimate the effective useful life as it relates to creating energy savings within a county, or within a utility's territory.

[7] Panel approaches are valuable for any retention analyses, not just for California.

As mentioned, instruments should include information about conditions – and especially changing conditions -- that may have impacted the failure or survival of the measure. These conditions may include changes in climate, changes in ownership of purpose of site and changes in economic conditions that might have affected the operation of the measure.

Greater time-of-failure accuracy results in greater EUL estimate accuracy. Interval censoring techniques can be used to aid EUL estimation if measures are simply recorded as being alive or dead at the time of the inspection or survey. Additional information is extremely helpful. In phone interviews, respondents were often able to report the year and possibly the season in which a measure failed; the greater the accuracy – month, for example – the better the estimates that can be derived. Since surveys are conducted at intervals often greater than 5 years, this level of accuracy is substantially greater. Some studies tagged measures with a phone number that could be called if the measure was removed. These strategies, of course, work best in situations where only a few measures were installed at each location. Program participants are unlikely to recall when they removed each of the twenty (or hundreds of) light bulbs that were installed in their store or facility, thus making this an ideal application of the "tagging" approach.

Proper follow-up survey or inspection interval choices can also increase EUL estimate accuracy. Perhaps the best design for a retention study might include at least three data collection phases: an initial survey, a follow up survey a short time thereafter, and a final survey conducted after a greater interim. The purpose of the second survey is to detect any equipment that failed anomalously early, perhaps due to equipment malfunction or owner distaste for the equipment. The final survey is more spaced out in order to pick up the overall trend of measure failure or removal. Obviously, the measure under review will dictate the appropriate time intervals. Measures like light bulbs may not be expected to last much more than three years, whereas a refrigerator may remain in place and operable for twenty years. EUL estimates from the manufacturer of the measure, from protocols (or the DEER or other databases) or from prior EUL studies, should be consulted to aid in determining survey intervals. Plans and procedures may need to vary by type of measures (lifetime categories); if the interval is too short, there may be insufficient (or no) failures, and a statistical model may not be estimable. Likewise, if the interval is too long, too many (or all) of the measures may have failed.

Where fieldwork is conducted, we find that the best studies all met certain criteria that improved the reliability of the data. Auditors should be trained in coding and data recording practices. The specific practices vary according to the measures involved and the firm conducting the data collection, but should always include (1) clear definitions/descriptions of the measures, (2) the definition of a failure and (3) how to reconcile discrepancies (e.g. if a measure is still in place when a phone interview suggested that it had been removed). Engineers may need to accompany the fieldwork team to site visits for supervision and to aid in the reconciliation of apparent discrepancies for especially complex[8] or important measures. Measures should be physically verified as in place and operable or not.[9] To the extent that it is possible, having the same staff conduct audits between years may resolve uncertainty regarding measures whose status is questionable. Finally, tagging the measures with a unique identification label during the initial inspection (generally performed during or just after installation) can greatly reduce uncertainty regarding whether original measures are still operable.[10] Tagged fixtures can also be mapped within the site, making subsequent surveys faster, easier, and more accurate.

Lastly, the best studies we reviewed employed several data management techniques that increased our confidence in the research. Double blind data entry should be used. This is a final check against data input problems that might substantially affect the final outcome of the retention research.

[8] An approach like this may be needed for whole building measures that are not easily identified, for example.

[9] While most on-site audits are conducted via physical inspection already, some of the studies that we reviewed relied on staff accounts of measures. While a maintenance manager or other staff member might be able to recall the status of the measures, employee turnover or a very large number of measures decreases the reliability of staff accounts.

[10] Tagging is especially useful in the case of measures that are easily replaceable, such as light bulbs. If the original light bulb failed and was replaced, it may be incorrectly coded as a survivor in the absence of an identifying tag.

Measures, or participants as the case may be, should be assigned identifying codes. Data merging should occur across these codes. Follow-up phone calls should be made to survey respondents if a questionable answer arises upon review of the data.

Data Treatment and Analysis

Outliers

A sensitive and often overlooked issue in retention studies is the treatment of outliers. As retention studies tend to use statistical models to estimate survival curves for the measures for which EULs are to be estimated, they rely on data containing a number of failures and a number of survivals. If there are only a few of either, the confidence level surrounding the EUL estimates is enlarged. If one of those few failures or survivals is, in fact, an outlier, the entire EUL estimate can be off; outliers can be very influential in the results, and not in a positive way.

Tests should be performed to detect outliers. Fortunately, at least in the case of most energy efficiency measures, a common sense reality test can detect obvious outliers (e.g. an air conditioner that lives for 90 years). In some cases, formal testing for influential data points may be necessary. The analysis should be conducted with and without suspected outliers to test for differences in outcomes. If the results are appreciably different and there is a high-degree of suspicion that a particular observation is an outlier, it should be removed or addressed in an appropriate manner.

Model Selection

Among the studies that we reviewed, many had too few failures to estimate EULS, or among those with sufficient failures, many estimated only one functional form. However, among those estimating multiple models, the most common distributions used to estimate survival curves for energy efficiency measures were the exponential, log-logistic, log-normal, Weibull and gamma distributions. These models all differ in the flexibility offered by forecasts derived from them. These are standard models for fitting survival and hazard curves, and are all estimable with standards statistical and regression packages such as SAS. The model selection process can be subjective, but based on our review work, we have developed three categories that can inform selection.

Theoretical Expectations. The first category is congruence between theoretical expectations and forecasted results. If theory dictates that the hazard rate should increase over time, then the model should facilitate that shape. Of course, this selection criterion requires *a priori* expectations about the hazard rate – if none exist, then other selection criteria should be used.

Implications. The second category is implications of results. If a particular model suggests an unreasonably long or short EUL, it should be removed. Likewise, if a particular model suggests an EUL that is far from the estimates suggested by the other models, and the assumptions behind the other models seem justified, the model in question might be determined to poorly represent the failure shape or hazard function. This may reflect the different data requirements or data intensiveness of the various functional forms. This selection criteria category may not provide guidance in every case, but it can frequently be used to eliminate weaker models.

Formal testing. Finally, formal testing can provide valuable – and less subjective – insight into the technical effectiveness of the different models. Techniques, such as residual analysis, the max of log-likelihood, or the likelihood ratio test can be employed to determine which model or models best fit the data.

At least one of these selection techniques should be used for every retention study, and where reasonable, all three. Regardless of the selection technique used, multiple models should be estimated and prepared for each measure studied. It is not necessary to pick one model as the overall best model. Results from multiple models can be reported in summary, though if all models produce noticeably different results, some explanation or recommendation should be given if the research is to be of practical use.

Specifications

The models used in the studies that we reviewed usually contained few covariates, if any, in the final specifications. This is appropriate if the inclusion of additional regresses does not prove to have any significant effect on the estimates and fit statistics of the model. However, tests should be conducted to determine whether exogenous factors affected the data.

The kinds of exogenous factors that should be controlled depend on the context of the retention study – was the study implemented through a public program, are the measures involved likely to be sensitive to economic fluctuations, were weatherization measures installed, etc. The most common exogenous factors included as regressors in our review were (1) changes to the facility, structure, or business, (2) changes in occupant or use of the space in which the measures were originally installed, and (3) seasonal effects such as weather/climate and hours of operation.

Another choice that might affect EUL estimates is how dependent failures are treated. Although technically this is a coding decision and not a model specification characteristic, it is best tested at the estimation phase. It is not uncommon for multiple units installed at the same location (or at a cluster within a location) to fail simultaneously (or simultaneously according to the nearest time-of-failure data). Such cases may require different treatment, especially if it is likely that there is reason to suspect a causal relationship between the failure of one unit and the failure of others. Not only must different models be estimated (one that combines all failures, one for dependent failures and one for independent failures) but different definitions of dependency must be tested. Some studies may define failures as dependent only if someone from the installation site can confirm the dependency. Others use numerical definitions, such as 40% or more of a group of measures failing.

Bias Corrections

Bias corrections are not always necessary, especially if appropriate precautions are taken during the sampling design and data collection phase. Nevertheless, biases sometimes arise despite the best efforts of the research team in charge of gathering data. The most common bias remedy is the application of weights to the data before model estimation. The appropriate weighting strategy is dictated by the research situation. The most common weights observed were business type, sector, energy demand, measure type (where multiple measures were aggregated and studied together), and the reciprocal of the probability of inclusion, or combinations thereof. Weighting can be useful to remedy attrition from one survey to the next, as well as any other number of sample distortions that might occur. When weighting schemes are used, however, studies should report comparative results between weighted and unweighted data.

Another important bias correction issue that surfaces frequently in retention studies is clustering. As discussed earlier, clustering can occur if measures are sampled based on their location. Site audits may disproportionately reflect certain measures, and measure lifetime may be correlated within sites, or within site segments (groups of related measures installed at a site). If clustering is suspected to be an issue, standard errors should be estimated using adjustors such as the design effect factor.

Common Mistakes

Our experience in reviewing many retention studies was that, despite their analytical approach, the majority of the studies that attempted to follow the guidelines and suggestions presented above were successfully able to provide reasonable and useful EUL estimates. However, we also noticed several common mistakes in the studies.

Small Sample Size. The most common problem with the studies that we evaluated was an insufficient sample size. In some cases, a small sample size was the result of an inadequate data collection effort, and therefore easily avoidable. However, other studies worked from poor population lists (usually obtained from program tracking data). In such cases, sample size complications were far beyond the control of the research team. Still, any possible effort should be taken to ensure a sufficient

sample size. Inadequate samples can lead to several insurmountable analytical problems, from large confidence intervals to models that do not converge.

Failure to Test Other Models. Another common mistake was the failure to test several models using different distributions when estimating survival functions. Often, this occurred when a research team tested different functions for one measure, then applied that function to the rest of the measures covered by the study. Because different measures act differently, the same model assumptions will not always be justifiable from one type of equipment to the next. This caveat is especially important when parametric models are being used. Although failure rates may accelerate with time for both refrigerators and air conditioners, they may not accelerate in the same way.

Ambiguous Failure Dates. There is often a tendency for inspections and surveys to fall short in their attempts to obtain approximate failure dates. Even if the exact date of failure is unknown, any additional information regarding when the measure in question failed can be used to narrow the censoring interval. Follow-up questions, such as year of failure, season or month of failure may produce more accurate responses than simply asking whether the measure is still in place at the time of the interview. Accurate failure date responses are easier to obtain when the measure being studied is more noticeable. Even the maintenance supervisor for a large and busy building is likely to know approximately when an energy management control system stopped working. The best remedy for large failure date intervals when measures are small and numerous (such as light bulbs) is more frequent surveys – though this course of action can be expensive. However, if the measure is common or responsible for a large share of savings (and potentially a large share of earnings claims) the extra investment may be well justified.

Poor Documentation. Although this criticism applies more to the *report* than the *study*, having reviewed many retention studies, we feel that it is important to note that, the biggest problem that we encountered was documentation that was inadequate to determine exactly what procedures had been followed, hypotheses tested, modeling applied, coding adjustments made, weighting schemes used, etc. Some reports had included formulae that were not relevant to the models estimated. If anyone from outside the research team will need to read the report produced from an EUL study, thorough documentation can greatly facilitate both (1) understanding of how the study was conducted and (2) the conclusions drawn from the study. Regardless of whether a potential reader is reviewing the study for accuracy, to assess shareholder earnings claims, or simply trying to gain insights from its conclusions, the documentation step is frequently overlooked but extremely important.

Failure to examine results in context. Very few studies looked "outside themselves". There are now many retention studies that have been conducted across the nation for a large number of measures (including previous studies for the same program in many cases). Discussion of results is improved if results are reviewed and compared to other studies to identify similarities, patterns, and differences, and provide a context for the findings.

Table 1 summarizes the assessment of a review of the scores for assessment studies undertaken as part of the California Public Utilities Commission (CPUC) Protocols. The table shows the percent of studies reviewed that received scores on a point system running from 1 (low) to 5 (high). The studies tended to do relatively poorly on methodology, and well on sampling / survey-related tasks. The bulk of the studies received scores between 2.5 and 4.0 (C- to B range).

Table 1. Distribution of Scores of Retention Studies Reviewed for CPUC (Skumatz, et.al. 2004)

Score	Protocols	Survey, Sampling	Data Coll'n	Modeling	Total, unwtd	Total, Wtd[11]
1.0	2%	0%	0%	0%	0%	0%
1.5	0%	0%	0%	0%	0%	0%
2.0	4%	0%	6%	26%	2%	2%
2.5	0%	6%	0%	15%	6%	24%

[11] The "weighted" scores weigh methodology scores more heavily. Average is not on quite the same scale because "letter" grades were assigned and then translated into numbers for the overall grades

3.0	48%	48%	56%	19%	55%	19%
3.5	30%	4%	2%	4%	24%	6%
4.0	7%	35%	28%	26%	10%	30%
4.5	6%	6%	6%	7%	6%	19%
5.0	0%	2%	2%	4%	0%	2%
Average	3.21	3.46	3.39	3.13	3.30	"3.50"

Summary

The authors recently completed a review of more than ninety measure retention and lifetime studies. Through the review process, we examined the studies for conformance to basic requirements, and also developed a set of best practices for retention studies. Although there is never a "right" way to conduct a study, the guidelines and suggestions presented in this paper establish a baseline strategy upon which measure retention and lifetime studies can build. Much of our experience evaluating retention research has shown that particular design and analysis choices depend on the context and purpose of the study and the measures under investigation; however, the "best practices" suggested in this study can serve as a guideline for key issues and approaches for retention studies. Table 2 presents a generalized summary of the practices discussed throughout this paper.

Table 2. Summary of Best Practices

Best Practices
• Obtain a strong and unbiased population source list from which to conduct a draw a sample.
• If the number of measure installations is small, conduct a census. Otherwise, use a probability sample. Stratify the sample based on important population characteristics, such as climate zone and energy demand. Consider establishing a panel survey.
• If possible, use a measure-based sample, rather than a site-based sample.
• If phone interviews are conducted, use call management. Schedule phone calls in advance, use at least 3-5 callbacks, and leave sufficient time between callbacks.
• Pretest survey instruments for each measure under investigation.
• Ask about conditions that might affect the operations of the measures.
• Try to get the most accurate information about measure-failure dates and explore causes / reasons.
• Conduct follow-up interviews at time intervals appropriate to the measures under investigation.
• Use trained and supervised auditors.
• If on-site inspections are used: Physically verify the status of each measure; Affix identifying tags to measures and create a map of the measures sampled.
• Use standard data-management practices, such as double-blind data entry and follow up calls regarding questionable responses.
• Test for outliers (either visually or with a formal procedure) and remove obvious outliers.
• Compare different models and model specifications with respect to their congruence with theory, implications for results, and results from formal tests.
• Include influential variables as regressors to control for exogenous factors.
• If failure dependency is suspected to be an issue, estimate a combined model as well as models for dependent and independent failures.
• If the sample does not accurately reflect the measure population, weight the data using the most appropriate means, and report both weighted and unweighted results.
• If the sampling strategy resulted in clustering, use common standard error adjustments to compensate
• Compare results to previous studies and discuss differences, considerations.[12]
• Clearly document the study and methods, alternatives considered, rationale, and discuss in context of results from other similar studies.

[12] See, for example, other reports by the authors for quantitative lifetime results, including Skumatz, et.al, 2005.

REFERENCES

Skumatz, Lisa A., John Gardner, and David Bell, "Revision / Updating of EULs Based on Retention and Persistence Studies Results / Draft", prepared for Southern California Edison for the Four IOUs, March 31, 2005.

Skumatz, Lisa A., Ph.D., Rose A. Woods, and Scott Dimetrosky, "Review of Retention and Persistence Studies for the California Public Utilities Commission (CPUC), Final Report", October 20, 2004.