Likert scales are too simplistic – better and more useful alternatives in four applications in energy efficiency

Dana D'Souza
Skumatz Economic Research Associates
P.O Box 1486
1577 Legend Lake Circle,
Silverthorne, CO 80498
USA
dsouza@serainc.com

Lisa Skumatz
Skumatz Economic Research Associates
P.O Box 1486
1577 Legend Lake Circle,
Silverthorne, CO 80498
USA
skumatz@serainc.com

Keywords

qualitative survey, selection bias, design process

Abstract

Likert Scales (5, 7, or 9-point) are in common use in energy efficiency (EE) evaluation, usually asking respondents to indicate a selection along a linear scale marked with labels on the extremes (very dissatisfied and very satisfied, or with labels at each numeric point. EE work commonly assigns equal differences in scores for each numeric increment, even though the originators warned against the practice. The evaluation, survey, and literature research conducted in this paper shows equally-spaced differences between the numeric or verbal extremes can be biased. This study provides an overview of Likert Scales, clarifies key analytical issues, and provides practical examples of more defensible approaches that should be considered in place of the simple analysis methods commonly applied with Likert scales. Specifically, we suggest adaptations of "labeled magnitude scaling (LMS)" or other Labeled Scales (LS) as more appropriate scoring values for the Likert increments. This paper suggests improved practices in four example areas of evaluation where Likert Scales are commonly used:

- Process Evaluations: Likert scales are common in process evaluations (e.g., applied to "satisfaction", and "ease of participation"). LMS or other labeled scale values provide more robust information and defensible estimates of average likelihoods and other factors.
- Net to Gross: The most sophisticated net-to-gross (NTG) methodologies use a combination of direct and corroborating questions to identify free ridership and spillover. Several questions involve Likert scales, and the current computations assume 25 %/50 %/75 %/100 % "likelihoods" in the NTG computations. Transitioning to LMS scaling would provide likelihood percentages that are defensible and would reduce bias.
- Quantifying Better/Worse Program Changes: In matrices used to analyze or score program options, Likert or linear scales are often used for those elements that cannot be monetized. A LMS scale better represents differences than the "one unit" differences arising from Likert scales.
- Barriers Analysis: Assessment of program barriers in process evaluations commonly rely on Likert scales. Changes in barriers from, say, 3.4 to 3.6 do not provide implementable information. Using an extension of LMS (non-energy benefits or non-energy impacts NEB/NEI) provides explicit recommendations on the dollar size of the barrier and the incentive/intervention value needed to erase the barrier.

While none of these changes may be large, the LMS approach allows for greater defensibility, may support additional computations and meaning from surveys and applications, and will lead to more robust and defensible Evaluation, Measurement, and Verification (EM&V) results.

Introduction

Likert Scales are straightforward and commonly used in Energy Efficiency (EE) evaluation surveys to measure frequency, importance, quality, likelihood, agreement, and are extended to net-to-gross and other factors. However, analysts frequently apply scores of equal value between the Likert points, and this

simplification of linear response values is not an appropriate analytical technique for Likert scales. Likert scales were developed by psychologist Rensis Likert in 1932. Initially a 5-point scale (numbered 1–5)[1], some studies use 7-point or sometimes 9-point scales[2]. These scales present a bipolar response scale with negative options presented first and allow quick responses by requiring respondents to read more carefully.[3]

Likert Scales, interpreted with equal-value differences, may fail to capture accurate values. Likert scales provide greater granularity and finer quantitative responses than yes/no, or agree/disagree. However, the "intervals between values cannot be presumed equal"[4]. Thus, for a scale where 1 = strongly agree, 2 = agree, 3 = neutral, 4 = disagree, and 5 = strongly disagree, a mark of 4 would be more negative than either 3, 2, or 1 (directionality). However, it cannot be inferred that a response of 4 is twice as negative as a response of 2. These scales do not say that the strength/intensity of an attitude is linear and can be measured using this ordinal scale response in a quantitative way, except to say that the responses can be ranked in order.

Applying Likert scales in a linear way requires further caution, because some research also suggests that people tend not to select the extreme categories in large rating scales, perhaps not wanting to appear extreme[5], so assigning unitary differences across the entire scale is particularly problematic. In addition, Likert-derived data may fail to meet other assumptions for parametric tests (e.g., a normal distribution).[6] Since data are ordinal, non-parametric statistics are typically considered the most appropriate option for analysis. If the data are interval, then parametric statistics can be used.[7]

Research indicates moving from strict Likert Scales to Visual Analog Scale (VAS) may better capture extremes. Some investigators have abandoned the Likert-type response in favor of a simple visual analog scale (VAS)[8]. The VAS typically has descriptive anchors only at the two extremes. The participant is free to mark the scale at any point desired resulting in a continuous interval measurement with scores constrained between 0 and 100 (3). The scale can be scored by manually measuring the participant's chosen mark from the left end. A modified measure using a VAS with verbal anchors only on the two extremes[9] has also been developed and used. Studies comparing Likert and VAS have found there appeared to be systematic end aversion bias in scorings on the Likert scales.[10] (They noted that a larger percentage of the respondents tended not to mark the two extreme ends of the Likert scales as compared with scorings on the corresponding VAS. Thus, on the Likert scales but not the VAS, respondents seemed to mark the three middle response categories more often, suggesting end aversion bias.

Overall, there is no conclusive evidence that either VAS or Likert based scales are superior to one another from a statistical point of view. Rather, the context of application and circumstances of use seems to be of greater importance. Some discussion continues regarding whether VAS data are "interval"[11], but it is suggested that, with VAS available, there seems little reason to use Likert-type, or other non-interval responses in most research applications.[12]

Research indicates there are better alternatives available that can avoid some of the weaknesses of Likert scales. Certainly, Likert scales can be used to collect data quickly, allowing responses from 1–5, 7, or 9. However, even though the literature warns against assigning equal value increments, users commonly treat a move from an average score of 4.2 to 4.0 the same as they would a move from 3.2 to 3.0. And in other applications (described below), evaluators routinely assign 1 as 25 %, 2 as 50 %, 3 as 75 %, and 4 as 100 %. This is not justified by the science, and comparisons presented later in this paper, show that the common treatment biases the resulting calculations and conclusions. VAS appears to provide some advantages over the Likert approach. It requests responses between 1 and 100 that allow unequal intervals; however, the larger and more flexible scale can be more troublesome for respondents (and potentially slow responses). However, VAS's non-linear/ non-interval responses map to non-linear scores metrics, providing greater flexibility and the ability to reflect more nuanced gradations in intensity.

VAS may be an improvement on Likert, but this paper suggests adaptations of another approach – labeled scaling (LS) – may provide an even better alternative for Evaluation, Measurement, and Verification (EM&V) applications that currently use Likert Scales. This approach is described below, along with a series of examples of suitable applications in which this approach can provide more robust and *actionable* recommendations than Likert or VAS.

Identifying Labeled Scaling (LS) options as a More Robust Scaling Alternatives for EM&V Applications

Researching Non-Energy Benefits/Non-Energy Impacts (NEBs/NEIs) revealed weaknesses in valuation questions. Starting in 2002, the authors (Skumatz 2002, Skumatz and Gardner, 2005), were seeking a practical and robust option for asking NEI valuation questions. NEI research focuses on measuring and monetizing effects beyond energy savings that are delivered by energy efficiency programs (low income, residential, and commercial programs). For a group of NEIs (like the dollar value of changes in comfort from EE measures), participant surveys are virtually the only source of the needed information. The tradi-

1. Scales truncated to an even number of categories (typically four) to eliminate the "neutral" option in a "forced choice" survey scale. Rensis Likert's original paper clearly identifies there might be an underlying continuous variable whose value characterizes the respondents' opinions or attitudes and this underlying variable is interval level, at best (4).

2. The seven-point scale has been shown to reach the upper limits of the scale's reliability (5).

3. Kunz 2015; DeCastellarnau 2017.

4. Jamieson, S, 2004.

5. Jamieson, S. 2004; Kunz 2015.

6. Jamieson, S. 2017.

7. Allen and Seaman, 2007.

8. Hayes and Patterson (1921) is usually cited as the origin of the VAS method. Its acceptability as a generic pain measure was demonstrated in the early 1970s.

9. One objection to the use of VAS responses is the challenges of doing this ("making a mark") on computerized questionnaires. For computerized surveys or other instruments Reips and Funke 2008 recommend their website, www.vasgenerator.net, which generates VAS usable on the computer. They also offer information on the precision of these scales along with others (8, 22). This should alleviate some of the issues of large-scale computerized measurements.

10. Hasson and Arnetz 2005; Bishop and Herron, 2015.

11.Reips and Funke 2008, Hayes, Allen, Bennet 2013.

12. Bishop and Herron 2015.

tional literature suggested carefully-constructed willing-to-pay or willingness-to-accept (WTP/WTA) questions as the preferred option for obtaining monetized values for certain experiences, public goods, and other hard-to-measure, non-market items. However, phone interviews with program participants made it clear that WTP/WTA approaches had major disadvantages in actual practice. Even with clear and accessible wording, respondents were confused about what they were being asked, tried to ask for clarifications, and had difficulties providing and dealing with dollar values. This respondent difficulty and uncertainty led to extra time on each response, which negatively affected survey completes, time, and as a result, project budgets[13]. This finding led the authors to explore other question alternatives, including options related to Likert or VAS approaches, but ultimately, to a more robust approach that turned out to have applications beyond the initial NEB/NEI work.

Exploring other valuation options indicated that questions asking *relative* valuations were easier for participants to answer than WTP/WTA, and this can be accomplished with a version of Likert scales (extremely valuable, no value, etc.). Directionality is important in this application (four is greater or more valuable than three), but the inability to assign units or multipliers that measured the differences in intensity in a defensible way was a fatal flaw in this monetization approach and led to research of approaches that are several steps beyond Likert or VAS. A brainstorm on a ferry ride led to a possible solution. While participants might not be able to answer the dollar value of a less drafty window, or the comfort from the program, they might be able to answer a question about value of, for example, "comfort" in relative terms. Is "comfort" more or less valuable to them than another item "X" – and if we asked relative to an "X" that we did have a dollar value for, we might be able to calculate a monetary value for "comfort". The authors tried asking about percentage multipliers to express the relative value (Pearson and Skumatz 1998, Skumatz 2002), but that required responses in number terms – the very thing identified as difficult for respondents.

With that nugget of an idea, we researched the academic literature and found a field called "labeled magnitude scaling" (LMS) and later, labeled hedonic scaling (LHS), Labeled affective magnitude (LAM) and other related literature.[14] This field match our basic, simple concept. And the research indicated that people assign consistent ratios to specific "labeled" relative terms – phrases similar to "like very much" vs. "like moderately". Since these ratios are known from large populations (from numerous academic publications), then if the same language is adapted to valuing the NEIs, the researched "multipliers" might be able to be applied to the value of "X" to estimate the respondent's dollar valuation for changes in comfort from an EE measure.

There are several key points about this LMS/LHS/LAM work (we will call it "Labeled Scaling" LS in this paper) that are important:

1. These academic value ratios are well-estimated with confidence intervals and other statistics, and are estimated and tested in multiple studies.
2. Our research on NEIs, in which we often ask both LMS and percentage values from within the same sample, has found that these academic values are extremely consistent with the within-sample LS multipliers.
3. The values are NOT linear; instead, the distances vary with different labeled modifiers.
4. The work for NEIs requires a few more adaptations, but this is not the focus of this paper.[15]

Table 1 shows the multiplier values associated with up to five positive and negative points and one neutral on several scales developed and in use (11-point scale). Notice they are non-linear, and also, that the negative and positive values differ slightly for similar labels. Figure 1 demonstrates the "non-linear" nature of these labeled factors. Figure 1 shows how the values compare to linear Likert multipliers (the dashed line).

Scaling methodology work (in taste and other areas) was undertaken to provide methods or scales for measuring degrees of difference or approximating magnitudes (the difficulty associated with Likert); to be able to apply the techniques to measuring sensation (intensity) differences and hedonic experiences (experiences that cannot be shared directly with others), and ultimately; and to be able to permissibly substitute identifiers with numbers. The underlying approach depends on responses to stimuli as greater than or less than a standard. This is the fundamentally attractive metric that relates to NEIs, which attempt to measure and value sensations that relate to comfort and other difficult-to-measure factors. [16]

A Stretch? Applying Labeled Scaling Outside its Usual Wheelhouse

With our work focus in energy efficiency and sustainability, the authors weren't interested in relative "tastiness" of foods, severity of pain, or similar rankings in the food tasting, psychology, and other fields; however, our review of this literature implied this concept might have properties that would be useful in NEI work. Perhaps the same types of language could be applied to relative "value", not just "liking" a taste. We adapted and applied labeled magnitudes, guided by this language, to NEI questionnaires for several programs. We conducted several tests of the concept, asking participants to identify whether the comfort they experience from the program was slightly more valuable than X[17], moderately more valuable, and other "labeled magnitudes" along the scales. Within the same surveys and respondent groups, we also asked by what percent more or less valuable they assessed the value of "com-

13. The authors note that the WTP/WTA responses were slowest among the three alternatives used in the surveys conducted for the 2002 study, and led to the most attempts by respondents to ask clarifying questions. Skumatz 2002, Skumatz and Gardner, 2005.

14. Moskowitz HR. 1982; Green, et al. 1993; Schutz and Cardello, 2001; Lim, Wood, Green 2009.

15. See, for example, multiple other publications, including Skumatz et. al., 2009, Skumatz 2014, Skumatz et.al., 2019, and the works referenced therein.

16. For a longer description of development, differences and advantages/disadvantages, see Juyun Lim 2011.

17. Pioneering this in 2002, we used "the energy savings from the program", later "energy savings from the measure" for the X value for most of our EE work. We have used other X language as we adapted this work to other types of projects. (Skumatz 2002, Gardner and Skumatz 2005, Skumatz et. al., 2009, Skumatz et. al. 2019 and others.)

Table 1. Semantic Phrases and Scale Values for Multiple Positive/Negative Descriptors.

	Labeled Hedonic Scale (LHS)	Labeled Affective Magnitude (LAM)	Oral Pleasantness and Unpleasantness Scale (OPUS)		Generalized Labeled Magnitude Scaling (g-LMS)
(Like) Greatest Imaginable	100.00	100.00	100.00	Strongest Imaginable(Like) =>	100.00
Extremely	65.72	89.58	82.29	Very Strong=>	52.08
Very much	44.43	56.25	63.54	Strong=>	35.42
Moderately	17.82	37.50	42.71	Moderate=>	18.75
Slightly	6.25	10.42	22.92	Weak=>	6.25
Neutral	0.00	0.00	0.00	Neutral=>	0.00
(Dislike) Slightly	-5.92	-9.38	-20.83	Weak=>	-6.25
Moderately	-17.59	-29.17	-38.54	Moderate=>	-14.58
Very much	-41.58	-54.17	-64.58	Strong=>	-31.25
Extremely	-62.89	-83.33	-83.33	Very Strong=>	-48.96
(Dislike) Greatest Imaginable	-100.00	-100.00	-100.00	Strongest Imaginable (dislike)=>	-100.00
Language / Use =>	Like/Dislike	Like/Dislike	Pleasant/Un.		Strength

Source: Skumatz calculations (2021) from multiple sources.

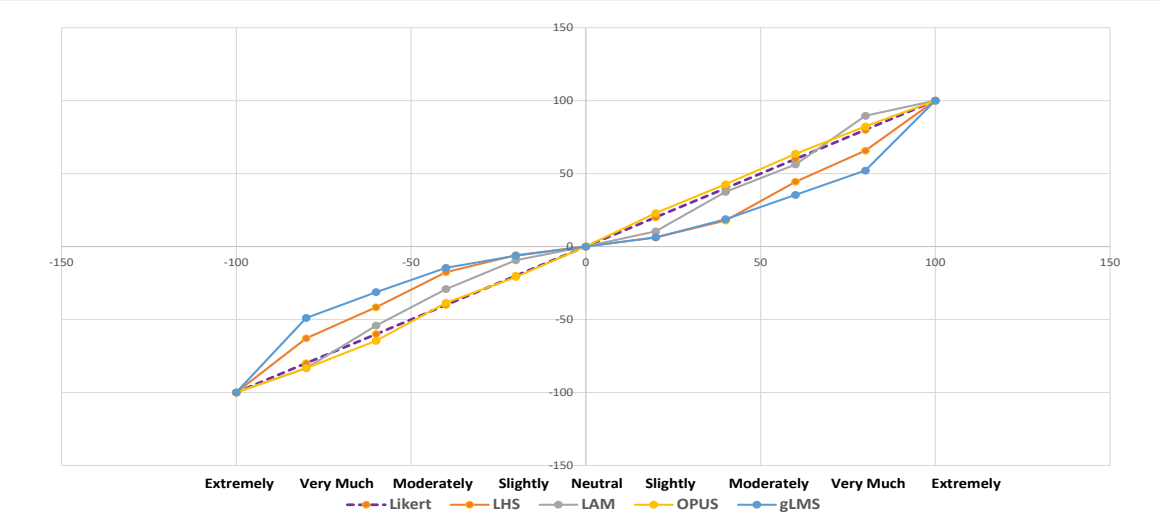

Figure 1. Semantic Phrases and Scale Values for Multiple Positive/Negative Descriptors.

fort" and other NEIs – a harder question, but necessary to be able to compare multipliers with academic values. The results for these NEI comparisons showed that the average multiplier assignments to labeled verbiage was remarkably close to these non-linear values. One example follows in Figure 2. The first two bars in each group are representative academic values, and the others reflect estimated in-sample average multipliers from different groups of respondents associated with a five-point Labeled Scale.

Note that we have selected some or all of the various scaling language options in different projects. For phone surveys, we selected fewer options; for web surveys we have been able to include most or all of the labels (up to extremely, not "most imaginable"), with similar capabilities to closely match the academic values. Whether or not these LS options are strictly applicable to NEIs, they do seem to work, they offer a simple link to get monetized values for NEIs (and multipliers and ratios for other applications), and they offer options that participants can answer – a very big advantage.

We have now applied this approach to more than 50 programs – initially using the g-LMS multipliers, and later tending to apply language and multipliers for the LHS multipliers; however, based on more recent literature,[18] we now tend to use

18. See Juyun Lim, 2011. The 9-point Labeled Hedonic Scale is widely used, and is considered a simple an effective measuring device, but there are concerns about its performance at extreme values. The gLMS is a similar scale, with similar concerns about end points. Because some researchers were concerned about the suitability of the gLMSs adjectives and their suitability for pleasantness/unpleasantness applications, the Oral Pleasantness and Unpleasantness Scale (OPUS) was developed. A final hedonic category-ratio scale that was designed to address specific limitations of each of the existing scales is the Labeled Affective Magnitude (LAM) scale. Research asserts the LHS appears to have strong properties (better discrimination and better resistance to ceiling effects). However, reliability and sensitivity studies suggest that the LAM has equal reliability and sensitivity, greater discrimination at the extremes, and is judged by consumers to be as easy to use as hedonic scales and easier than magnitude estimation.

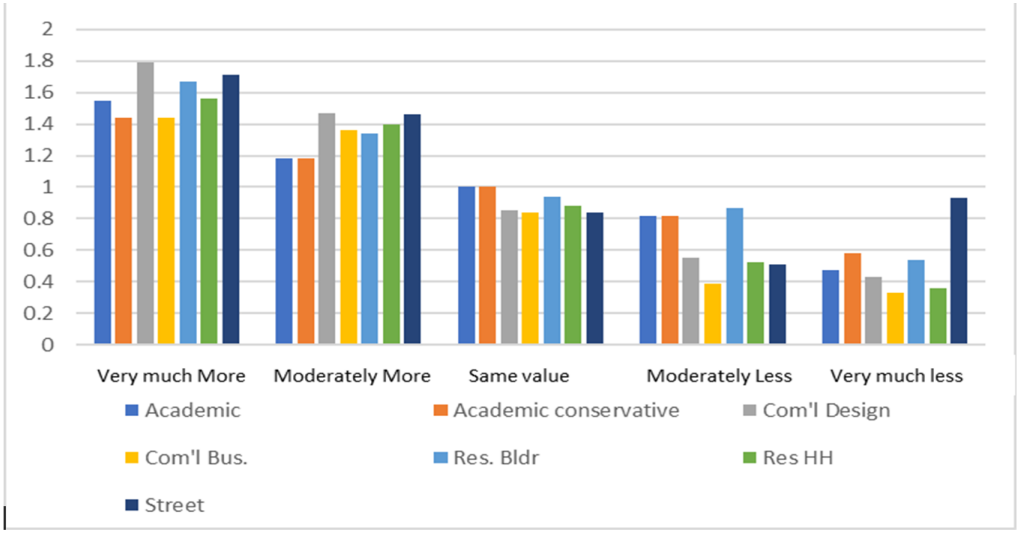

Figure 2. Comparison of Academic LS multiplier Values with In-Sample Values for a SERA NEB/NEI study of LEDs for Multiple Categories of Interviewees. Source: Skumatz et.al., 2020.

the LAM scale. Our early choice of the g-LMS multipliers was initially made for two reasons: one, because the fit seemed most important based on the academic underpinnings, and two, because the multipliers were more conservative than others, and we were interested in (consciously biased toward) making sure the NEI multipliers didn't overstate the valuation of NEIs (especially in the field's early years).

As we have applied versions of LS to our EM&V work, we find the following compelling advantages worth considering:

- Labeled magnitudes – words – are easy for respondents to answer, and presumably at least as easy as a number scale.
- The ratios associated with the labeled magnitudes are directly translatable to numbers, allowing less biased estimates of intensity, proportions, etc.
- These ratios are justified based on academic literature, and are more defensible than ad hoc assignments of 25 %, 50 %, etc.

In the paragraphs below, we have applied these various Labeled Scaling (LS) scale values to a number of topics in energy efficiency evaluation to illustrate some of the performance advantages and some of the complexities.

Effects of Moving from Likert to Labeled Scaling in Key EM&V Applications

PROCESS EVALUATIONS THAT BETTER REFLECT INTENSITIES

Likert scales are routinely used to estimate awareness, satisfaction, and other values in EE process evaluations, commonly using a 5-point scale (very, somewhat, and neutral). Performance scores across populations are presented and compared as simple averages of values from 1 to 5. Using LS values provide more robust interpretation of the ratios of intensity associated with the factors. For example, in the 5-point scale selected below in Table 2 related to satisfaction with the program's application process, "extremely" is more than twice as large a multiplier as "moderately". Similarly, if a 7- or 9-point scale was selected, the "distances" or intensity ratios between any pair of labeled scales is non-linear. Using a labeled magnitude approach is fairly straightforward and nearly as easy to include or code as a Likert scale, the questions are clearer to the reader, and the interpretation is quite clear and the multipliers are defensible and presumably less biased than *ad hoc* values.[19] Average participant scores are calculated by multiplying the percent responses for each response category times the weight or score for each response category. Under the Likert scale, 3.0 is commonly "neutral"; under the Labeled scales, 0 is "neutral".

NET TO GROSS WITH MORE DEFENSIBLE MULTIPLIERS

Net-to-Gross (NTG) calculations[20] often includes elements relying on Likert scales.[21] In the most sophisticated NTG methodologies, exemplified by the array of questions and corroborating elements for estimating free ridership and spillover in Massachusetts, several inputs rely on Likert scales. The inputs and steps assigned for the Massachusetts NTG methodology are included in Table 3. In the MA methodology, 25 % increments are assigned to the Likert scores in steps in the Free Ridership portions of the NTG calculation; these ad hoc increments introduce inaccuracies or unsupported assumptions into the NTG results. The authors reviewed the methodology, recognized the bias introduced by use of these Likert scales and recommended use of an improved methodology for NTG computations being used going forward for Connecticut NTG work. The Labeled scaling approach has the potential to provide clearer questions and more defensible percentages and NTG results.

19. And recall that the foundational literature for Likert Scales states up-front that differences between rankings are only more vs. less, and implies that ratios between the values cannot be assigned (and are not linear).

20. The NTG ratio represents the share of gross program savings that are attributable to the program. 1-NTG is the share of savings that would have happened without the program (but incentives were paid). The ratio is the combination of Free Ridership (participants and incentive recipients that would have bought the EE measure even without the program), and spillover (savings added without incentive payments, e.g., induced by gained knowledge, word of mouth, etc.).

21. Skumatz, et. al., 2009, NMR 2020.

Table 2. Applying Labeled Scaling to Process Evaluation Questions.

A. 5-Point Likert Label for "Satisfied with Program Application Process"	B. Common Likert % values	C. Suggested Labeled Scaling Labels for "Satisfied with Program Application Process" – 5 point*	D. g-LMS values	E. LHS values	F. LAM values	G. OPUS values
Very Dissatisfied – 1	0 %	Extremely dissatisfied	-0.52	-0.66	-0.90	-0.82
2	25 %	Moderately dissatisfied	-0.15	-0.19	-0.29	-0.39
3	50 %	Neither satisfied nor dissatisfied	0.00	0.00	0.00	0.00
4	75 %	Moderately satisfied	0.19	0.18	0.10	0.23
Very Satisfied – 5	100 %	Extremely satisfied	0.52	0.66	0.90	0.82

(*) Note that Labeled options support 5, 7, 9-point scales, with special wording.

This change forces recalculations of several parameters. The "timing", "efficiency level", and "quantity" questions for the Free ridership questions rely on scales of "not at all likely (scored at 0), slightly likely (scored at 0.25), somewhat likely (scored at 0.5), very likely (scored at 1), and 0.5 for don't know or refused.[22] While the system appears to have adopted some of the language of Labelled Scaling (LS), it has not adopted the scoring. According to Figure 1, if a 4-point scale is preferred (and we use the LAM scale), then the language and scores should be revised to Slightly likely (.10), moderately likely (0.38), Very (much) likely (0.56), and Extremely likely (.90). Not at all likely could remain with a 0. The scores may differ somewhat based on the relative appropriateness of LAM, gLMS, etc., but LAM has appeared to reduce bias in some of the extremes, as noted elsewhere in this paper. The assignments of do not know and refused can be set using a weighted average if those respondents are "random" or non-systematic, or another value; however, a pre-determined value of 0.5 (as currently included in the MA calculations) seems inappropriate/ad hoc. IF, to provide continuity in NTG values over time, the minimum and maximum must be set to 0 and 1, then the intermediate scores can be re-set to LS ratios using 1/0.9 times the in-between values.

Two elements of the spillover questions also rely on Likert-type scales and equi-distant valuations. These are "influence" and "actions in the absence of program participation"[23]. The former phrases the increments as Not at all important (score 0), slightly important (score 3), somewhat important (score 6), and very important (score 10). Under a LS system, which recognizes the increments intended or expressed by respondents are not equidistant, better labels and scores for these elements would be revised to not at all important (0), slightly important (1.04*10/8.96=1.16), moderately important (3.75*10/8.96=4.18), very important (5.6*10/8.96=6.25), and extremely important (10) – including an adjustment to make sure the scale still ranges from 0 to 10 to not introduce discontinuity in the time series of spillover values and calculations by introducing an entirely new range. Finally, the language for "actions in the absence of program participation" is phrased as not at all likely (score 10), slightly likely (score 6), somewhat likely (score 3) and very likely (score 0). The improved verbiage and scores would be revised to (here, again, assuming the same range is needed, but going from 10 to 0): not at all likely (10-0=10), slightly likely (10-(0.94*10/8.3)=8.87), moderately likely (10-(2.9*10/8.3)=6.5), very likely (10-(5.4*10/8.3)=3.5), and extremely likely (0). Note that initial calculations indicate that substituting LAM values for the intensity values (but not timing values) changes the NTG results by about 5.5 %.[24]

The existing Connecticut (CT) NTG elements rely on slightly different language, and contains some similar elements, but also includes elements of the NTG computations that are assigned without strongly-defensible links to the scores (for instance, the influence and timing scores).[25] Using revised scoring that directly link to justifiable scoring of likelihoods, is probably worth considering in other states. If fewer response options are desired, options can presumably be skipped. Although not all NTG, free ridership, or spillover questions are phrased or used exactly as MA, the calculations above provide an example of how to translate Likert or equi-distant scales to less biased versions.

QUANTIFYING BETTER/WORSE PROGRAM CRITERIA MORE MEANINGFULLY

In matrices used to analyze or select among program options, criteria often include both quantitative data and qualitative factors. Some of these qualitative factors may be "pass/fail" or "satisfactory/unsatisfactory". However, many of these qualitative factors can be relative – better, worse, much worse, than another option or the status quo. Likert or linear scales are often used for those elements that cannot be monetized; for example, options are scored better or worse on a 5, 7, or 9-point Likert scale, or may be simply assigned a rank from 1st to last. As noted above, both of these scoring systems are unable to reflect the non-linear nature of how much better or worse an option may be. Extreme-

22. For MA responses to question about timing: 'Without the <incentives (etc.)> how likely is it you would have installed any type of equipment at the same time'; efficiency: 'Without the <incentives etc.> how likely is it that you would have installed the exact same high efficiency level of equipment?', and quantity: 'Without the <incentives, etc.> how likely is it that you would have installed the exact same quantity of/amount of high efficiency measures/?

23. For MA responses to question about influence: 'On a scale of 0–10 where 0 is not at all influential' and 10 is 'very influential', how influential was <program attribute> in your decision to install the high efficiency measure'. Program attributes include program rebate, marketing materials, etc. the question for actions in absence of program: 'How likely is it you would still have purchased/installed energy efficient <equipment> that did not receive a rebate from <program sponsor>?

24. Lisa Wilson-Wright and Melissa Meeks, NMR, 2020.

25. The CT logic for Influence is 1–5 for "no influence" to "had a great influence" and a 3 or higher is associated with "influenced by the program". The CT logic for efficiency levels is asking if participants would have installed more efficient model, same efficiency, less efficient, would not have installed/purchased, or don't know/refused. The calculation treats "if more or same" scored as 0 to 1.0; if less, 0–0.75, and if none, score of 0. Both options would likely be improved by a move in the direction used in MA.

Table 3. Steps in MA NTG Calculations (Free Ridership and Spillover).

Subject	Sub-Element	Key Differences	Uses Likert 25%/50% etc.
Free-ridership	Prior intentions	Uses prior intentions only to resolve respondent inconsistencies; not part of primary FR algorithm	No
	Timing	Respondents who say they are slightly or somewhat likely to have installed the equipment at the same time in the absence of the program are asked about doing it within six months, between six months and a year, or more than a year later. They are not given the option of doing it sooner. Those who say they would have done it more than a year later skip the efficiency and quantity questions (max FR of 0.25)	Yes
	Efficiency levels	Examines the likelihood of purchasing the same efficiency level and does not allow for an option of purchasing a more efficient model	Yes
	Quantity	Examines the likelihood of purchasing the same quantity in the absence of the program	Yes
	Influence	Uses a scale of 0 to 10 to rate program influence	No
Spillover	Screening	Examines only purchases or changes where the respondent believes program participation influenced their decision	No
	Identifying measures	Uses general categories to prompt respondents on possible spillover measures	No
	Ascertaining measure efficiency	Uses detailed, open-ended questions to examine spillover measure efficiency	No
	Influence	Offers four options (not at all important, slightly important, somewhat important, very important) to rate program influence	Yes
	Actions in the absence of program participation	Examines likelihood of taking actions (not at all likely, slightly likely, somewhat likely, very likely) in the absence of program participation	Yes
	Consistency check	Examines if the measure would have qualified for the program and limits spillover to only those measures.	No

(*) – CT uses many similar steps, but based on this research, is modifying MA's methods to change from Likert to Labeled scaling for three steps.

Source: NMR and Tetra Tech 2020. Likert notes added by this paper's authors.

ly better (say, on hassle factor or complexity to install a measure), is not simply "one unit" better than Very much better – the LS table indicates a move from "moderately" to "very much" is a 1.5 to 2.4 times increase in value, and from "moderately" to "extremely" is an increase of 1.3–3.7 times. A LMS scale better represents differences in intensity of difficulty or ease than the "one unit" differences arising from Likert scales.

In our very simple example (Table 4), we show differences that might arise when using a Labeled Scale vs. a simple Likert Scale. The example is for a city identifying what strategies – from among a limited set of energy/energy efficiency and solid waste strategies – might be most suitable when considering what options to use to move toward greenhouse gas (GHG) reduction goals.[26] The criteria weights are included near the bottom of the table. Criteria A, B, and C are all quantitative. Column A is the ratio of the cost-per-metric ton of carbon equivalent, normalized to have the cost for Commercial lighting be "1". Column B is also numeric, and calculated in a similar way. Column C is expressed in years. Finally, Column D (or Column E) is used as an example of a non-quantitative factor that is important for a City the options. Note our case assumes a community without a City Municipal Energy Utility. In that case, then, it is very difficult for a City to get a commercial lighting in place at their servicing utility. Presumably they would need to go through influencing the state-level PUC to influence the utility to develop such a program (assuming these programs do not exist). However, a city has the ability to pass an ordinance for a special trash rate program, or to put out an RFP to contract for new curbside recycling service, and so on. It is harder for cities to gain authority in the commercial trash sector, but not very difficult, and that is reflected in the scores below. With a Likert scale, we find the ratio from easiest to hardest is only 5:1. However, using the labelled scale, extremely difficult is actually something that is about 9 times worse than "slightly difficult". And the scores in-between are also not linear. Looking for the lowest score among the weighted scores on the far right, we find a difference in the lowest (best) scoring option (largely constructed by our large weight on "difficulty), but the point of the example is to note that a simple Likert or a High/Medium/Low assignment for some qualitative factors is easily replaced by a more robust, meaningful, and defensible labelled magnitude system and it can better reflect unequal, and non-linear (almost exponential) differences between "low" and "high" associated with qualitative factors. This refinement is very easy to substitute and we use it all the time.

BARRIERS ANALYSIS – MOVING TO IMPLEMENTABLE RESULTS

Process evaluations often involve analysis of barriers. Likert scales are the most common method used to assess the importance of barriers. Using a scale similar to Column A in Table 4, respondents mark a 1–5 (or 7 or 9) in response to the question, and the weighted average response is reported as the "score" for the particular barrier. Recognizing that a numbered is most useful when compared to another number, the list of barriers may be ranked by their score, indicating that, perhaps, the paperwork barrier is a 3.2 and other barriers score from, say,

26. Note this example comes from a Skumatz ACEEE paper from 2010. Therefore, some of the ratios of dollars-per-MTCE are now outdated; solar/wind are likely less expensive (Skumatz, 2010).

Table 4. Example of Refined Options for "Better/Worse" in an Options-Comparison Context with some Qualitative Indicators.

City Selecting Strategy for GHG Reduction - Energy vs. Recycling-Related								
1 is fastest / best / cheapest	A. Ratio $/MTCE	B. Relative Jobs/$ spent	C. Relative Speed to implement	D. Difficulty for City to Implement (Likert, 1=easy; 5=difficult)	E. Difficulty for City to Implement (LS words)	F. Labeled Scaling score/10	G. Weighted Final Score - Columns A, B, C, D - Using Likert	H. Weighted Final Score - Columns A, B, C, F - Labeled Scale
Commercial Lighting Program	1	1	3	5	Extremely difficult	9	3.4	5.4
LI Weatherization	3	2	3	5	Extremely difficult	9	3.9	5.9
Wind	7	3	10	5	Extremely difficult	9	6.2	8.2
PV	17	11	10	5	Extremely difficult	9	9.0	11.0
Pay As you Throw Trash Rates	0.5	5	0.5	1	Slightly difficult	1	1.2	1.2
Curbside Recycling	0.7	2	1	1	Moderately difficult	3.8	1.0	2.4
Curbside Yard Waste	0.7	0.5	1	1	Moderately difficult	3.8	0.9	2.3
Commercial Recycling	2	1	3	3	Very difficult	5.6	2.6	3.9
Commercial Food Scraps	9	0.5	3	3	Very difficult	5.6	4.0	5.3
Criteria Weights	**0.2**	**0.1**	**0.2**	**0.5**		**0.5**		
(Based on study from 2010; some values no longer accurate								
Rationale for Extremely difficult for City to implement energy programs - assumes city does not have municipal energy utilty and can only influence through PUC								

Source: Skumatz 2010.

1–4.3. Unfortunately, using a Likert scale does not allow the analyst to suggest that the barrier with a score of 1 is twice as bad or dire as the barrier receiving a score of 2. Comparisons over time can also be made. The paperwork barrier may have been 3.0 last year, and is 3.2 this year. This 0.2 change in score can indicate the situation is worse (assuming the confidence intervals show a difference), and its change could be compared to changes in score for other barriers to identify those that worsened most dramatically in a year. However, the differences are not meaningful in ratio form, and provide no information about how much more or less important some barriers are than others to the participant or non-participant respondents, and do not directly identify remedies that would resolve the issue or get potential participants to "neutral". The first two difficulties could be resolved if Labeled Scales were used instead.

To go a step further, Likert (and simple Labeled Scale) responses do not provide much implementable information about what to do next, or how much value respondents associate with the barrier. This question can be answered, however, with an enhancement on the Labeled Scale system. If a well-designed NEI set of questions[27] were incorporated into the process evaluation survey and used to assess barriers[28], the resulting NEI value would represent a direct estimate of the monetary value associated with the barrier. Assume the estimate is $75 per business. This value could be interpreted to mean that, on average, this barrier (e.g., maintenance of high-tech HVAC equipment) is a $75 barrier to participation.[29] The clear and direct program recommendation would be that if the utility It could be interpreted to indicate that, on average, the barrier could be erased (or customers brought to "neutral") with an increase in incentive of $75, or a similarly-valued "free" warranty visit, or other such incentive relevant to the program and measure/barrier. A drill-down on the NEI value would also provide information on the distribution, not just the average, value of the NEI. Program staff could potentially find that 50 % of the customers could be brought to neutral with an incentive of perhaps $20, and other findings and recommendations could be derived that would be far more specific than the information derived from a Likert (or VAS) or sampled Labeled Scale analysis. These values can be – and have been – calculated using a version of the question framework presented in Table 5[30, 31].

Conclusions and Recommendations

Likert-based survey questions are common in EM&V work, including process evaluations, barriers analysis, and net-to-gross applications. Simple rankings or Likert scales are also occasionally applied to the non-quantitative elements of program, measure, or project analysis on a wide variety of topics. Proper analysis of Likert scales allows the researcher to identify that a score of 4 is higher than a score of 3, but does not support estimating how much higher with any confidence, and analysis techniques are limited. VAS approaches offer an improved option, but this paper suggests that one of a set of Labeled Scales (LAN is our most preferred as stated earlier[32]) provides advantages in a number of EM&V analyses.

27. Which employ questions that use Labeled scales as a key component, as described above.

28. A negative NEB/NEI can be interpreted as a barrier; and a not uncommon negative NEI from SERA NEI studies are concerns that maintenance for very high-tech energy efficient HVAC equipment in (small) commercial buildings might be beyond the capabilities of in-house custodial staff or local HVAC contractors.

29. or to selection of this measure or one measure vs. another, or however the NEI questions are posed.

30. Skumatz, 2015a and other citations.

31. "What, if any, positive, negative, or no effects did you receive from <measure / program> above and beyond what you would have received from installing a standard efficiency measure": Note the proper baseline should be used for replacement on failure vs. early replacement, etc. baselines.

32. LAM was shown to have equal reliability and sensitivity to the hedonic scale, provided somewhat greater discrimination among highly liked foods, and resulted in data that were similar to magnitude estimation in terms of the obtained ratios among rated stimuli. The LAM scale was also judged by consumers to be as easy to use as the 9-pt hedonic scale and significantly less difficult than magnitude estimation.

Table 5. Framework for Barrier Question using NEB/NEI Approach – Simplified Version.

Content

A. INTRO ONCE AT TOP
- Short Intro on NEBs concept
- Identify the program/measure/measure group they are in the sample for to link to correct NEB sub-list.
- What, if any, positive, negative, or no effects did you receive from <measure / program> above and beyond what you would have received from installing a standard efficiency measure? (open end)

B. FOR EACH MEASURE, FOR EACH NEB
- Some people who received the same measures as you in the program say they experienced a change in <NEB1>. Thinking about <NEB1>, would you say you received a positive or negative change in <NEB1> from the measure, or no impact compared to what you would have received from a standard efficiency model? (no impact, go to next NEB). (+/-/0).
- The following questions will ask about effects above and beyond installing a standard efficiency measure. Positive value: - Was the NEB1 effect you received from the measure more or less valuable than energy savings you received from the measure. Drop down or multiple choice: for level of value - <LABELED SCALE>

IF POSITIVE VALUE:
- Extremely more valuable than the energy savings (LAM multiplier=1.90).
- Very much more valuable (1.56)
- Moderately more valuable (1.37)
- Slightly more valuable (1.10)
- About the same value – value and savings (1)
- Slightly less valuable (.9)
- Moderately less valuable (.56)
- Very much less valuable (.37)
- Extremely much less valuable (.1)

IF NEGATIVE VALUE:
- Extremely more costly than the energy savings (-1.83). The negative effects are much larger than the energy savings.
- Very much more costly (-1.54)
- Moderately more costly (-1.29)
- Slightly more costly (-1.09)
- About the same costliness – the negative effects or costs are about balanced by the energy savings) (-1)
- Slightly less costly (-0.83)
- Moderately less costly (-0.54)
- Very much less costly (-0.29)
- Extremely much less costly (-0.09). The negative effects or costs are only a small share of the value of the energy savings/energy savings more than balances out the negative effects)

- Then repeat for other NEBs for the measure/measure group they are sampled for.
- (Also asking about any outside the list they mentioned, and place under "other")
- For every 3rd NEB, ask percent more/less valuable or costly for in-sample multiplier to compare to academic.

C. ONCE, at the end:
- You mentioned positive and negative effects/values from a number of NEBs above and beyond what you would have received from standard efficiency equipment. Thinking about all of them … would you say the total value of the effect was positive or negative or no value/effect? (+/-/0)
- Use positive/negative valuations from above.
- Ask percent more/less valuable or costly.
- Use to normalize to assure sum of the individual NEBs remains equal to total NEB value.

The refinements suggested in this paper will not lead to dramatic differences in calculated results, but the suggested approach is more robust and the calculations more defensible and justifiable than the more common Likert methods used for these process and evaluation analyses. We suggest that readers consider modifying their current EM&V surveys to incorporate Labeled Scaling verbiage and approaches in lieu of Likert options where feasible. This paper illustrates the LS approach in a number of examples, but we suspect analysts will be able to expand these examples to many more evaluation applications offering improvements over traditional linear scale methods.

References

Allen, I. Elaine, Seaman, Christopher A., 2007, "Likert Scales and Data Analyses", *Quality Progress*. July, 2007. http://rube.asq.org/quality-progress/2007/07/statistics/likert-scales-and-data-analyses.html

Bishop, Phillip A., Herron, Robert L, 2015. "Use and Misuse of the Likert Item Responses and Other Ordinal Measures", *International Journal of Exercise Science*. 2015; 8 (3): 297–302. Published online 2015 Jul 1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4833473/

DeCastellarnau, A., 2018. "A classification of response scale characteristics that affect data quality: a literature review". Qual Quant 52, 1523–1559. 2018. https://doi.org/10.1007/s11135-017-0533-4; https://link.springer.com/content/pdf/10.1007/s11135-017-0533-4.pdf

Green, B.G., Shaffer, G.S., Gilmore, M.M., 1993. "Derivation and Evaluation of a Semantic Scale of Oral Sensation Magnitude with Apparent Ratio Properties", *Chemical Senses*, 18 (6), 683–702 December, 1993. https://doi.org/10.1093/chemse/18.6.683

Hasson, Dan, Arnetz, Bengt B., 2005. "Validation and Findings Comparing VAS vs. Likert Scales for Psychosocial Measurements", *International Electronic Journal of Health Educations*, 2005; 8: 178–192. https://files.eric.ed.gov/fulltext/EJ794094.pdf

Hayes, M. H. S. & D. G. Patterson, "Experimental development of the graphic rating method". 1921. *Psychological Bulletin*, 18, 98–9.

Hayes, John, Allen, Alissa L., and Bennett, Samantha M., 2013. "Direct comparison of the generalized Visual Analog Scale (gVAS) and general Labeled Magnitude Scale (gLMS)", National Institute of Health, Food Qual Prefer. 2013 April 1; 28 (1): 36–44. doi:10.1016/j.foodqual.2012.07.012.

Jamieson, S. 2017. "Likert scale." *Encyclopedia Britannica*, September 27, 2017. https://www.britannica.com/topic/Likert-Scale.

Jamieson, S., "Likert scales: how to (ab)use them?", 2004 *Medical Education*, 38 (12), pp. 1217–1218. (doi: 10.1111/j.1365-2929.2004.02012.x) http://eprints.gla.ac.uk/59552/1/59552.pdf

Kunz, T., 2015. "Rating scales in Web surveys. A test of new drag-and-drop rating procedures" . Technische Universität, Darmstadt [Ph.D. Thesis]. 2015.

Ledbetter, Marc R., Lisa A. Skumatz, Ph.D., et.al., 2019. "Energy Saving Opportunity from Advanced LED Lighting Research", prepared for Pacific Northwest National Laboratory, October. https://www.pnnl.gov/main/publications/external/technical_reports/PNNL-29342.pdf

Lim, Juyun, 2011. "Hedonic scaling: A review of methods and theory", *Food Quality and Preference*, 22 (2011) 733–747.

Lim, Juyun, Wood, Alison, Green, Barry G., 2009. "Derivation and Evaluation of a Labeled Hedonic Scale". , *Chemical Senses* 34 (9): 739–51. DOI:10.1093/chemse/bjp054; November. https://www.researchgate.net/publication/38014936_Derivation_and_Evaluation_of_a_Labeled_Hedonic_Scale

McLeod, Dr. Saul, 2019. "Likert Scale Definition, Examples, and Analysis", *Simply Psychology*, Updated 2019. https://www.simplypsychology.org/likert-scale.html

Moskowitz, HR.,1982. "Utilitarian Benefits of Magnitude Estimation Scaling for Testing Product Acceptability". Philadelphia (PA): American society for testing and materials. 1982.

NMR Group and Tetra Tech, 2020. "Consistent Methodology for Self-Reported Net-to-Gross Measurement", Submitted to Massachusetts Program Administrators and the Energy Efficiency Advisory Council. https://ma-eeac.org/wp-content/uploads/MA19X03-B-RSRNTG_Residential-SR-NTG-Report_FINAL_2020.5.28.pdf

Nunnally, Jum C., Bernstein, Ira H., "Psychometric Theory", Tata – McGraw Hill, 1978. Third Edition.

Reips, U., Funke, F., "Interval-level measurement with visual analogue scales in Internet-based research: VAS Generator", Behavior Research Methods, 2008. *Psychonomic Society*, Inc., 2008, 40 (3), 699–704, doi: 10.3758/BRM.40.3.699. http://www.frederikfunke.de/papers/pdf/reips&funke_interval_level_vas.pdf

Schutz HG., Cardello AV., "A Labeled Affective Magnitude (LAM) Scale for Assessing Food Liking / Disliking". *Journal of Sensory Studies*. 16:117–159. 2001. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-459X.2001.tb00293.x

Skumatz, Lisa A., Ph.D., and Sami Khawaja, Ph.D., "AESP webinar on NEBs in Low Income Programs", September, 2010.

Skumatz, Lisa A., Ph.D., 1997. "Recognizing All Program Benefits: Estimating the Non-Energy Benefits of PG&E's Venture Partners Pilot Program (VPP)", Proceedings of the 1997 Energy Evaluation Conference, Chicago, IL.

Skumatz, Lisa A., Ph.D., 2002. "Comparing Participant Valuation Results using Three Advanced Survey Measurement Techniques: New Non-Energy Benefits Computations of Participant Value", Proceedings of the 2002 ACEEE Summer Study on Building Conference, Asilomar, CA.

Skumatz, Lisa A., and John Gardner, 2005. "Methods and Results for Measuring Non-Energy Benefits in the Commercial and Industrial Sectors", Proceedings from the IEPEC Conference, August.

Skumatz, Lisa A., Ph.D., M. Sami Khawaja, and Jane Colby, 2009. "Lessons Learned and Next Steps in Energy Efficiency Measurement and Attribution: Energy Savings, Net to Gross, Non-Energy Benefits, and Persistence of Energy Efficiency Behavior", prepared for CIEE Behavior and Energy Program, CIEE, California Public Utilities Commission, Berkeley, CA, November.

Skumatz , Lisa A., 2010. "A kWh is not just a kWh: Comparing Energy Efficiency Programs in Terms of GHG, Job Impacts, and Policy Achievements (NEBs and Beyond)", Proceedings of the American Council for Energy Efficiency Summer Study on Buildings (ACEEE), Asilomar, CA, August.

Skumatz 2014. Non-Energy Benefits / Non-Energy Impacts (NEBs/NEIs) and their Role and Values in Cost-Effectiveness Tests: State of Maryland, prepared for Natural Resources Defense Council (NRDC), New York, March.

Skumatz, Lisa A., 2015, "Efficiency Programs' Non-Energy Benefits: How States are Finally Making Progress in Reducing Bias in Cost-Effectiveness Tests, *The Electricity Journal*, September.

Skumatz, Lisa A., 2015a, "Estimating Participant Non-Energy Benefits for Households and Businesses: Labelled Scaling Approach", Toolkit, prepared for use by CT EEB NEI contractors, August.

Skumatz, Lisa A., 2020, "NEB Values for Next Generation LEDs: Residential, Commercial, and Street Lighting" Proceedings of the ACEEE Summer Study on Buildings, Asilomar, CA, August.

Wilson-Wright, Lisa, and Melissa Meeks, 2020, NMR. Personal communication with Skumatz, May.