Journal of Human Sciences and Extension

Volume 7
Number 2 *Special Issue: Credible and Actionable Evidence in Extension Practice*

Article 6

6-28-2019

Measurement and Credible Evidence in Extension Evaluations

Marc T. Braverman
Oregon State University, marc.braverman@oregonstate.edu

Recommended Citation

Braverman, M. T. (2019). Measurement and Credible Evidence in Extension Evaluations. *Journal of Human Sciences and Extension, 7*(2), 6. https://doi.org/10.54718/TLJT3272

Measurement and Credible Evidence in Extension Evaluations

Acknowledgments

I am grateful to Ben Silliman for his suggestion of an earlier version of the mode lthat appears on page 2. I also thank my Oregon State University colleagues Kathy Gunter and Shauna Tominey for their very helpful suggestions about sections of Table 1

Measurement and Credible Evidence 88

Measurement and Credible Evidence in Extension Evaluations

Marc T. Braverman
Oregon State University

This article examines the concept of credible evidence in Extension evaluations with specific attention to the measures and measurement strategies used to collect and create data. Credibility depends on multiple factors, including data quality and methodological rigor, characteristics of the stakeholder audience, stakeholder beliefs about the information source, and the evaluation context. Measurement planning involves a process of making thoughtful decisions about choosing study variables, measurement strategies, and specific measures that adequately reflect the content and goals of the program being evaluated. The use of specific measures may also entail implicit assumptions, e.g., that the respondent is being truthful and accurate, which must be accepted if resulting data are to be viewed as credible. The article discusses aspects of measurement quality, including reliability and validity, for both quantitative and qualitative forms of data. Program stakeholders should be encouraged to be attentive, reflective, and critical in their analysis of evaluation evidence, and their views on what makes data credible must be understood and considered. The use of common measures in evaluating multi-site programs can be valuable, but only if the measures are fully appropriate for local sites. The article concludes with a summary of implications and recommendations for Extension evaluation practice.

Keywords: common measures, credible evidence, evaluation planning, Extension evaluation, measurement strategy, validity

How both program processes and outcomes are measured will largely determine the degree to which a program is determined to be effective.

—Schwandt (2015, p. 82)

Introduction

The credibility of an evaluation refers to the likelihood that stakeholders will accept the evaluation results as convincing and will accept the conclusions and recommendations as reasonable and justified. Judgments about the soundness, credibility, and persuasiveness of evidence set the stage for follow-up action and utilization, and thus, one can speak of the *actionability* of evaluation findings (Mark, 2015). Credibility for a particular audience depends on numerous factors including its timeliness, the relevance of the primary questions, the use of a rigorous design to answer those questions, and the quality of both the evidence and the

Direct correspondence to Marc T. Braverman at Marc.Braverman@oregonstate.edu

Journal of Human Sciences and Extension Volume 7, Number 2, 2019

conclusions (Donaldson, 2015). Measurement is at the heart of the process and is the focus of this article. The choice of measurement strategies and instruments and the effectiveness of data collection produce the raw material on which the analyses and interpretations rest.

In this article, I discuss several dimensions of measurement in Extension evaluations, and the implications for how convincing—that is, how credible and actionable—the findings and recommendations will be for program stakeholders. I begin with some clarification about the concept of credibility and why it is broader than simply an assessment of data quality, and I present a model of the components of credibility with regard to evaluation evidence. I then describe some relevant principles of the measurement process in program evaluation and their relevance to credibility. Based on these analyses, I conclude with recommendations for increasing the credibility and actionability of Extension evaluations through thoughtful measurement decisions.

Isn't Credibility Just a Reflection of Data Quality?

The concept of credibility, widely discussed in the current evaluation literature (e.g., Donaldson, Christie, & Mark, 2015), is universally acknowledged to depend, at least in part, on rigorous methods of investigation. So a good place to start is to ask why we speak of the *credibility* of our results and conclusions rather than simply the *quality* and *rigor* of the data. Some writers who use the term "credible evidence" do take this approach, concentrating only on the rigor of the methods used to produce the evidence. These methodological factors include the selected measurement instruments, the sampling procedures, and the evaluation design (often favoring randomized controlled trials). However, my own view is that credibility, though it undoubtedly depends on data quality and methodological rigor, is a more complex and multi-faceted concept. Since it refers to the likelihood that evidence is to be believed and judged as accurate, there must be a human angle involved. Observers will often disagree on what constitutes the strongest methodologies for collecting data, or even on what a particular response means. Some stakeholders are more skeptical than others about program results, methods, or assumptions about evaluation data.

A Model of the Influences on the Credibility of Evaluation Findings

Stakeholder judgments about credible evidence in Extension evaluations may be influenced by four kinds of factors:

- ***Data quality.*** This term encompasses a wide-ranging set of considerations, including the reliability of measures, the formats of the data, the timing of data collection, and the validity of the conclusions stemming from the evidence.
- ***Characteristics of the stakeholder audience(s).*** Most Extension evaluations will have multiple primary audiences, which may include administrators, internal program staff, clients (including adult participants as well as parents of participants in the case

Measurement and Credible Evidence 90

of youth programs), elected officials, and other community members. Some of these stakeholders will be more informed, more skeptical, more interested, more invested in the success of the program, better able to understand evaluation methodology, and/or more actively engaged in the evaluation process than others. Their tendencies to accept results as credible will vary significantly as a result of these predispositions.

- ***Stakeholder beliefs about the information source.*** Communications will be more readily accepted by stakeholders if they accept the information source as objective, trustworthy, and knowledgeable. The information source might be an organization such as Extension, a public agency, a private business enterprise, or an individual contractor.
- ***The context for the evaluation.*** All program evaluations take place within a larger context, which includes the nature of the organization delivering the program, the time and resource constraints on the evaluation, the decisions that might be riding on the evaluation results, and so on. That context will influence both how the evaluation is conducted and how it is received and accepted by its stakeholders.

Only the first of these factors, data quality, is directly related to what is thought of as methodological rigor. The other factors, to varying degrees, are dependent on perceptions, potential biases, predispositions, and political priorities, thus adding to the complexity of the concept of *credibility* in the assessment of evidence.

Perspectives on the Measurement Process in Program Evaluation

Several important issues about the measurement process provide background context for making assessments about data quality and credibility.

The Links Between Constructs, Variables, and Measures

In an earlier paper (Braverman, 2013), I described a model for developing evaluation measures that involves a four-step process. "The measurement specification process generally begins with the broadly conceived target construct, which reflects, often in everyday language, the issue that the program is designed to address" (p. 102). Examples of these constructs could be "healthy eating," "parenting skills," "interest in science," "leadership skills," "knowledge about common garden pests," and so on. Once the target construct is decided on, ideally with participation from stakeholders, the evaluator must decide how the construct will be translated into a variable to be measured, which measurement approach will be used, and finally, what specific instrument will be included in the evaluation. The instrument will consist of the specific questions (or sometimes a single question) that will be used to measure the target construct. Some instruments do a far better job of representing the target construct than others. If an evaluation has sufficient resources and the construct is of central importance, the evaluation planners might decide to measure the construct using multiple variables and instruments. The sequence for some of the decisions to be made in this process is illustrated for several sample constructs in Table 1.

Journal of Human Sciences and Extension Volume 7, Number 2, 2019

Measurement and Credible Evidence 91

Measurement Rigor

Braverman and Arnold (2008) defined methodological rigor as "a characteristic of evaluation studies that refers to the strength of the design's underlying logic and the confidence with which conclusions can be drawn. An evaluation that incorporates attention to methodological rigor will be in a better position to afford evidence and conclusions that can stand up to critical analysis" (p. 72). With specific regard to aspects of measurement in evaluation, Braverman and Arnold described several components of rigor that relate to measurement strategies. These include the conceptualization of program outcomes, decisions about how those outcomes will be represented by the evaluation measures, and the data collection strategies to be used.

***Table 1. Measurement Planning: Identifying Potential Options in the Progression from* Construct *to* Evaluation Measures**

General Construct	Specific Variables That Might Be Used to Represent the Construct (Selected)	Related Variables That Could Potentially Also Be Used as Relevant Outcomes (Selected)	Potential Measurement Strategies

Parenting skills	• Identification of parenting style
	• Parenting self-efficacy
	• Parent-child communication
	• Parent-child interactions:
	  • Expression of warmth
	  • Empathy
	  • Responsiveness
	• Discipline practices
	• Monitoring
		• Parental stress
		• Parenting satisfaction
		• Parent-child relationship quality
		• Positive child behaviors
			• Survey self-report questionnaire (scales or specific items):
			  • Self-ratings of knowledge gain
			  • Behavioral self-report
			• Observation of parent-child interaction:
			  • Live observations
			  • Videotaped interactions
			• Interview

Physical activity	• Daily, weekly, or monthly total minutes of *Moderate to Vigorous Physical Activity* (MVPA)
	• Number of days per week with at least 1 hour MVPA
	• Average or total number of steps per day
	• Physiological tracking (heart rate)
		• Body mass index
		• Sedentary behavior (e.g., sitting time) per day
		• Overall physical fitness
			• Survey self-report questionnaire (scales or specific items)
			• Activity logs or diaries
			• Interview
			• Activity monitors:
			  • Pedometers
			  • Accelerometers
			• Direct observation

Healthy eating	• Overall eating patterns
	• Food consumed in past week (or day or month)
	• Meal observation
	• Food available in home
	• Eating intentions
		• Family eating practices
		• Knowledge of:
		  • Nutrition
		  • USDA's MyPlate
			• Survey self-report questionnaire (scales or specific items):
			  • Food frequencies
			  • Dietary recall (e.g., over 24 hours)
			• Tracking of food purchases (e.g., from debit card)
			• Pantry inventory inspection

Journal of Human Sciences and Extension Volume 7, Number 2, 2019

Measurement and Credible Evidence 92

Kinds of Data: Quantitative and Qualitative

The measurement process will differ in significant ways depending on whether the data to be collected and analyzed are quantitative (in numerical form) or qualitative (in text form). The form of the data may change between data collection and analysis. For example, short-answer responses may be coded into categories or numerical quantities for certain kinds of data analysis. Qualitative data will be the product of free-form, open-ended responses on surveys, extended answers to interview questions, daily log entries, text-based descriptions from observers, and so on.

There are multiple and varied approaches to analyzing both categories of data, but in general, the analyses of quantitative and qualitative data tend to be distinctly different processes. Indeed, they often reflect different goals for what is to be described and learned. Quantitative data analyses generally take the form of summarizing the dataset in terms of descriptive or inferential statistics, e.g., through calculating mean scores to present a picture of the sample "on the average," or of exploring quantitative relationships between variables. By contrast, qualitative data analyses examine interviews, raw video and audio evidence, narratives, and observational notes to generate "rich" and "thick" descriptions of programs. These approaches can often provide unexpected insights that cannot be captured by checklists, surveys, or tests.

Mixed-methods evaluation designs, which make use of both quantitative and qualitative forms of data (e.g., Mertens, 2018), benefit from the distinct strengths and advantages of each format. However, because the analysis methods for the two forms of data differ greatly (e.g., Bazely, 2017), the criteria and approaches for making judgments about data quality are very different as well. Nevertheless, crucial data quality considerations, such as the relevance of the measurement approaches for addressing the major evaluation questions, the recognition of implicit assumptions, the awareness of ambiguities in the data, the recognition of potential biases, the appropriateness of interpretations and conclusions, etc., are of primary importance for both types of data (Creswell & Creswell, 2018).

Implicit Assumptions in the Use of Measures

The use of any measurement strategy or instrument entails some assumptions if we are to accept the resulting data as accurate and valid. Consider, for example, survey self-report, which is frequently used in Extension evaluations to measure attitudes, opinions, values, behavioral histories, behavioral intentions, assessments of programmatic success, and other types of outcomes. Several implicit assumptions are involved, and observers who disagree about the reasonableness of these assumptions will also disagree about the credibility of the responses. These assumptions include the following:

1. The respondent is trying to be truthful in reporting. In most cases, evaluators and evaluation audiences assume that the survey respondent is being truthful and honest. However, there are times when this assumption might be questionable, e.g., because respondents may be motivated

Journal of Human Sciences and Extension Volume 7, Number 2, 2019

to provide a socially desirable response (SDR; Dillman, Smyth, & Christian, 2014; Tourangeau, Rips, & Rasinski, 2000). Questions that ask about sensitive topics, such as household income, financial habits, drug use, sexual activity, or the respondent's compliance with regulatory requirements, are especially vulnerable to this form of bias. In these instances, the honesty of the survey respondents—or at least a subset of those respondents—might be reasonably judged to be open to question.

Social desirability scales are sometimes used to estimate the degree of bias due to SDR (Perinelli & Gremigni, 2016), but these add length to the survey and have limited effectiveness. In addition, a technique known as *randomized response* (Höglinger, Jann, & Diekmann, 2016) has been suggested as a formal methodological strategy to deal with sensitive survey questions, but it is complex and cumbersome. Most often, especially in Extension settings, common-sense strategies are employed to minimize social desirability bias, such as making the questionnaire anonymous, with the underlying assumption being that respondents will perceive that there is no reason to be dishonest if they cannot be identified. In practice, making questionnaires anonymous for sensitive question content, while undoubtedly helpful, is not fully satisfactory, because respondents' inclination to be honest and forthcoming is not entirely guided by logic (Tourangeau et al., 2000).

In addition to motives of self-protection, respondents may be motivated to answer in a way that they perceive as desired by the evaluator or program staff. For example, participants in Extension health education classes may be aware that it will benefit the program if they report positive personal impacts, such as increases in healthy eating and regular exercise. Thus, sometimes respondents may be trying to protect themselves by providing what they perceive as socially desirable responses, and sometimes they may be trying to protect the program in which they have participated.

2. The respondent is able to be reasonably accurate in answering the questions and is willing to make the effort to do so. Some questions that appear on surveys require cognitive effort to respond accurately, e.g., involving thoughtful judgment or memory recall. Examples include autobiographical behavior questions that ask about behavioral history or behavioral frequency, such as: "Which of the following foods did you eat yesterday?", "How many minutes of moderate-to-vigorous physical exercise do you usually engage in each week, on average?", and "When was the last time you had a physical examination from your doctor?". The last of those questions, which requires the respondent to accurately recall the amount of time that has passed since a previous event, is especially subject to errors of either overestimating or underestimating time periods, a phenomenon known as telescoping (Braverman, 1996).

Assuming that respondents are able to provide the necessary information, one must also trust that they are willing to engage in the concentrated effort needed to answer accurately. In many cases, respondents have been found to expend just enough effort to provide what they consider a "good

enough" answer, which might not reflect the degree of accuracy that the evaluator desires or expects; this phenomenon has been called *satisficing* (Krosnick, Narayan, & Smith, 1996; Tourangeau et al., 2000). Thus, the evaluator must be prepared to consider the question: Even if respondents are able to answer these questions, can we justifiably assume that they are motivated to do so?

3. The respondent is an appropriate source for the information desired. In addition to cognitive inaccuracies, respondents might also simply not have access to the requested information. For example, parents may be asked to estimate the number of minutes of screen time in which their children engage each week, their children's average amount of exercise, or the hours they spend doing homework, even though parents may not know these aspects of their children's daily lives. The fact that parents' level of information about these topics might be inadequate does not always stop evaluators from asking about them, and, quite often, that inadequacy does not even stop parents from answering the questions. Survey researchers have found that respondents will often answer a survey question with a complete guess, rather than selecting "I don't know" or leaving the response blank (Groves, 1989).

In summary, assumptions about respondents' honesty, accuracy, motivation, and knowledge are sometimes justified, sometimes not. However, these and other assumptions are frequently taken for granted without critical appraisal. If even a relatively small percentage of the respondents are unreliable in these respects, the overall quality of the dataset will be compromised. Furthermore, concerns about these assumptions should be accentuated when the respondents are children. Depending on their ages, many children's power of recall and their motivation may be insufficient to answer certain questions.

Characteristics of Measurement that Affect the Credibility of Evidence

Given this background context, let us examine several specific factors that, in many cases, might influence credibility. Most of these directly involve data quality and rigor, but others involve the perspectives of different evaluation audiences and the larger context in which the evaluation is conducted.

The Reliability and Validity of Measures

Reliability refers to the consistency of a measurement. If a survey item or scale is found to produce widely varying responses across conditions in which consistency is expected, the accuracy of scores from that item or scale will be suspect. That consistency, or lack of it, will influence the confidence we can place in the scores.

Our expectation of consistency is tied to our understanding about the variable we are trying to measure. For example, we would expect weight, height, and body mass index to be very stable if measured twice within 30 minutes (assuming no eating in the interim), but we would not have

Measurement and Credible Evidence 95

that same expectation for blood pressure, which varies to some degree every time it is taken (Bandalos, 2018). Thus, proper interpretation of reliability requires that we understand our variables, particularly with regard to prior expectations about the consistency and replicability of scores.

The "consistency" of measures can refer to replications across different time periods, different versions of a test, individual questions within a scale, or different individuals who are making judgments to produce the scores. Several of the major categories of reliability, as applied to quantitative forms of data, are the following:

- ***Test-retest reliability*.** This refers to consistency across short periods of time for variables that we expect to be relatively stable. Test-retest reliability is not appropriate for variables that experience change, e.g., indicators of mood or fatigue.
- ***Interrater reliability*.** Some outcome scores are based on judgments by raters. Examples might include essay tests to assess writing ability, athletic efforts to assess mastery of physical skills, or observations of parent-child interactions to assign scores on parenting style. The accuracy of these scores depends on the skill of the rater, and there should be minimal variation in scores based on who is doing the rating. Interrater reliability, often measured with the statistic Cohen's kappa, is a measure of consistency across judges.
- ***Internal consistency*.** This refers to the consistency of items that make up a scale. For example, if we have a brief, 8-item scale to measure leadership style, each item should correlate positively with the other items and contribute toward the overall score. Internal consistency is usually measured with the statistic Cronbach's alpha and is relevant for knowledge tests, attitude scales, and measures of psychological constructs. The statistic can also be used to produce the strongest scale from a set of candidate items.

Several qualitative research theorists have discussed how the concept of reliability can be applied to the analysis of qualitative data, although there is no consensus on this topic. Creswell and Poth (2018) place emphasis on the coding process: "In qualitative research, *reliability* often refers to the stability of responses to multiple coders of data sets" (p. 264). Miles, Huberman, and Saldaña (2014) note, with regard to what they call the "reliability/dependability/auditability" of qualitative data: "The underlying issue here is whether the process of the study is consistent, reasonably stable over time and across researchers and methods. We are addressing issues of quality and integrity: Have things been done with reasonable care?" (p. 312).

Validity refers to the appropriateness of interpretations, judgments, and conclusions that are made on the basis of scores. Validity theory has evolved significantly in the last several decades. The modern conception, developed by Samuel Messick (1989) and others, rejects the previously dominant view that validity is an inherent, identifiable quality of measures and tests. It is not

Journal of Human Sciences and Extension Volume 7, Number 2, 2019

accurate to talk about "a valid test," because for any given test or measure, some uses will be valid while others will not. Instead, validity is a property of the ways that measurement scores are used. Thus, one can talk about valid uses, inferences, or conclusions that are based on the information from one or more measures. (See Bandalos, 2018, or other recent texts on measurement and psychometrics for further discussion.)

To cite an example relevant to Extension, after a series of trainings for volunteers in a food preservation program, participants may be asked to rate the amount they learned with regard to food safety precautions. Self-ratings of this type cannot be considered a rigorous measurement strategy for the assessment of knowledge (as will be discussed further below). Thus, it might be considered a valid use of those ratings to make relatively low-consequence decisions about how the training sessions can be revised to be more interesting and comprehensive. However, it would not be a valid use of the self-rating scores to determine and certify which of the volunteers are sufficiently prepared to provide advice to the public regarding the safety of specific food preservation practices.

For qualitative data, as with the concept of reliability, there are competing perspectives on how the concept of validity can or should be applied. Miles et al. (2014) summarize the debate: "*Validity* is a contested term among selected qualitative researchers. Some feel that this traditional quantitative construct. . . has no place in qualitative inquiry. Alternative terms such as *verisimilitude* and *a persuasively written account* are preferred. But other qualitative methodologists continue to use the term purposefully because it suggests a more rigorous stance toward our work" (p. 313). Goodrick and Rogers (2015) prefer the term *inference quality* in place of *validity* for qualitative data, and they describe multiple strategies for strengthening the quality of inferences from qualitative analyses, depending on the specific analytic approach.

Scaling and Interpretive Clarity

Several principles of effective measurement relate to the way that potential responses are scaled. Consider a behavioral frequency question that may be asked in a nutrition education program, about the consumption of sugar-sweetened beverages (SSBs): "I drink sugar-sweetened beverages," with response options consisting of *Yes* and *No*. Most evaluators would consider that wording to be inadequate for several reasons. First, it is unclear what the dividing line should be between *Yes* and *No* if there is no guidance provided within the question. The respondent might interpret *Yes* to mean either "ever" or "regularly"—two very discrepant meanings. Due to the question's ambiguity, different respondents will probably make different judgments, and it will not be possible to understand precisely what information they have given us. Furthermore, even if the dividing line between *Yes* and *No* is clear, this behavior should be best expressed as a *range* of frequencies, since the two options are insufficient to cature the variability that exists in people's lives. Respondents might drink SSBs every day, once a week, once a month, or never. To be useful, the collected data should be able to reflect real-life variation, to whatever degree is needed for our intended uses.

Measurement and Credible Evidence 97

In addition to considering the number of response options, the wording of those options is important as well. For behavioral frequency, many Extension evaluations use some variation of *Rarely / Sometimes / Often / Always*. Although the inclusion of four options may provide sufficient spread, the labels are vague with regard to what the options actually mean, leading to ambiguity in the resulting responses. Survey research texts (e.g., Dillman et al., 2014) recommend using a set of options that are worded as clearly as reasonably possible. An example might be: *Rarely or never / About once a month / About once a week / More than once a week*.

To summarize, if we ask a question in a way that fails to represent the range of variation that exists among our respondents, our data will lack important information. Similarly, if we do not really understand what our respondents have told us with their answers, our data will be ambiguous. In assessing evidence, less informed stakeholders may not recognize these measurement problems, but more knowledgeable stakeholders will find such results confusing or untrustworthy.

Acknowledging and Compensating for the Limitations of Individual Measures

Every measure used in evaluation has limitations. The use of multiple approaches to measure a single critical construct, known as *triangulation*, can increase our confidence in the findings by exploiting strengths and compensating for limitations in individual measures.

Consider the construct of *healthy eating*, illustrated in Table 1. This can be measured in numerous ways, including direct self-report on survey questionnaires, food diaries, food pantry surveys, and plate waste studies (Braverman, 2013). Self-report is probably the most commonly used due to its convenience, time efficiency, low expense, and capacity to address past time periods. Yet as noted above, survey self-report also entails drawbacks, such as the risk of deliberate misrepresentation and potential problems with memory recall, question comprehension, and/or motivation to respond accurately. Therefore, an evaluator may choose to include additional methods of assessing eating behavior to supplement the information gained from self-report. If the information from multiple measures provides a consistent picture, the strength and credibility of the findings will be enhanced.

The major disadvantage of using multiple measures is the extra time and effort it requires for data collection. Respondents might lose patience or be confused by what they perceive as redundancy. Since the time allotted for data collection is almost always limited, other variables may be left out if a great deal of attention is devoted to accurately assessing one particular construct. The evaluator must weigh the advantages and disadvantages of using multiple measurements to strengthen the inferences regarding a single component of the evaluation, compared to addressing a broader set of questions.

Journal of Human Sciences and Extension Volume 7, Number 2, 2019

Measurement and Credible Evidence 98

Examining the Implicit Assumptions Associated with Measures and Measurement Strategies

As described earlier, all measures involve assumptions, and credibility is related to the reasonableness of those assumptions, relative to scientific and evaluation standards, or personal criteria of program stakeholders. An example is the use of self-ratings to determine levels of learners' skills and knowledge, a widely-used measurement strategy in Extension program evaluations. Rather than give participants a subject matter test on the content of an educational program (e.g., on gardening, diet and nutrition, parenting, personal finance, etc.), participants are simply asked to report the degree of their knowledge and/or the amount they have learned from the program. The level of self-rated knowledge can be compared pre- and post-program, and if it has increased, the conclusion will usually be drawn that the program has been effective.

Self-ratings are considered to be a form of *indirect* rather than direct measurement (Banta, 2004; Braverman, 2013). With regard to rigor, this approach is a weak strategy for assessing subject matter knowledge because many people will either under- or overestimate their own levels of mastery, and there is usually little or no evidence to support the accuracy of those judgments. The most obvious alternative strategy is to assess knowledge directly with a subject matter test. However, this will typically be more logistically difficult in several ways, which explains why it is not more commonly used. First, it would require more time for measurement, probably involving multiple questions about the subject matter, whereas a self-rating might involve only a single survey item with a 4- or 5-point rating scale. Second, an appropriate knowledge test that closely matches the Extension program curriculum would probably be available only rarely, and thus would often need to be created by the program staff or evaluator. Therefore, one can see why the use of self-ratings might be preferred based on evaluation logistics. However, with regard to the data quality itself, the credibility of the direct test is far superior, and the reliance on self-ratings in Extension program evaluations can result in reduced credibility of the findings.

Another illustration of potentially questionable assumptions is the measurement of behavioral intentions in place of measuring the actual behaviors of interest. For example, following an educational program, an evaluation may measure program participants' *intentions* to engage in regular physical activity or personal financial planning. These intentions can be a valid type of outcome on which to focus in an evaluation, but it is a mistake to assume that positive intentions can be equated with actual behavior changes, which are usually the outcomes of greatest interest in the assessment of program impact. Intentions can be easier to measure than behaviors because they can be assessed immediately following the end of the program, whereas the assessment of actual behaviors often requires the passage of time before those behaviors kick in. This would necessitate additional contacts with the program clients, many of whom might not respond. However, the danger of equating intentions with behaviors was demonstrated by Lohse, Wall, and Gromis (2011), who reported that participants' intentions to increase fruit and vegetable consumption following an Extension nutrition education class correlated poorly with their actual

Journal of Human Sciences and Extension Volume 7, Number 2, 2019

Measurement and Credible Evidence 99

consumption three weeks after the program's end. Yet sometimes these two kinds of variables—behavioral intentions and actual behaviors—are treated interchangeably in making claims for a program's success.

Evaluation Stakeholders: Variations in Sophistication, Judgments, and Priorities

As discussed earlier, the credibility of data is, in part, subjectively and individually determined. Credibility, by definition, refers to believability, and whether someone believes a set of conclusions based on a body of evidence is not entirely under the control of the evaluator, even if the evaluation has been designed and implemented in an exemplary fashion. Credibility depends to a certain extent on the perspectives of stakeholders. And as Miller (2015) notes, "it is not always the case that people will evaluate the credibility of evidence or information through rigorous analytical means. Indeed, the default appears to be *not* to analyze information rigorously and to rely instead on an initial intuitive judgment that is based largely on peripheral informational cues" (p. 49).

Ideally, stakeholder judgments about the credibility of evidence will be based on considerations of evaluation rigor. But this implies the presence of a critical, attentive, and knowledgeable audience. Without this orientation, the notion of credible evidence becomes irrelevant: even the least rigorous evaluation evidence will do.

Why would an evaluation audience not be engaged in this process? One reason is that they may feel they lack the necessary expertise, in which case they may relinquish responsibility to the evaluator for understanding and interpreting the results. Indeed, specialized knowledge is often needed, such as in decisions about how best to measure constructs when multiple options may be available. An evaluator must make measurement choices based on time, resources, and other considerations. Nevertheless, interested audiences may want to know the reasoning behind a particular choice and the evaluator's justification for why it was considered best.

A second reason for a lack of stakeholders' engagement would be if they are heavily invested in a particular result, e.g., finding evidence of outstanding program success, which could influence their willingness to be objective. If the evaluation comes back to their liking, they may embrace those results, disregarding considerations about rigor and the relative strength or weakness of the evidence. In such cases, it is up to the evaluator to strive for objective interpretation, recognizing and acknowledging whatever limitations in the measurements may exist. Even if some stakeholder audiences, such as program participants, are willing to take the evaluator's word about the strength of evidence, it would be a mistake to assume that this is true for all key stakeholders. At some point, it is likely that the evaluator will face tough questions about the methodological choices that were made.

Active and engaged stakeholders can help to ensure that evaluation results are used appropriately. To promote this orientation among stakeholders, they should be encouraged to

Journal of Human Sciences and Extension Volume 7, Number 2, 2019

engage in *evaluative thinking*, which is being increasingly recognized as a key component of evaluation use. Evaluative thinking (ET) has been described as "in essence, critical thinking applied to contexts of evaluation" (Buckley, Archibald, Hargraves, & Trochim, 2015, p. 376). It is especially relevant to the goal of building evaluation capacity within organizations (Patton, 2018; Vo & Archibald, 2018).

Evaluative thinking is important because credibility judgments are enhanced by the growing sophistication of stakeholder audiences. Being able to understand the basis for strong evidence leads to more appropriate and effective evaluation use. Evaluation users who demand or value rigorous evaluation methods will be better able to use evidence to build more effective programs. Buckley et al. (2015) make this point in noting that evaluative thinking "is the substrate that allows evaluation to grow and thrive. . . . ET is a protective factor to prevent against the risk of senseless, mindless evaluation" (p. 378). In other words, promoting evaluative thinking among our Extension program stakeholders with regard to measurement and other evaluation elements will eventually result in stronger programs.

The Use of Common Measures to Evaluate Multi-site Programs

Many Extension programs are implemented across multiple community sites. Some of these, such as 4-H Youth Development, are typically delivered in every county within a state. In addition, innovative projects funded by external federal, state, or foundation grants frequently involve wide implementation, sometimes covering sites in multiple states. In many of these programs, the goals, objectives, and target outcomes across sites are highly similar, and it can be logical to seek to measure common outcomes using standardized measurement instruments and strategies. For example, based on this reasoning, the National 4-H Council has developed common measures to be used in evaluating youth programs in the areas of science, healthy living, civic engagement, college/career readiness, and positive youth development (National 4-H Council, 2019), with the expectation that these measures will be used in programs that the Council funds within these topic areas.

Common measures are also used by the National Institute of Food and Agriculture's (NIFA) Children, Youth, and Families at Risk (CYFAR) Initiative (CYFAR Professional Development and Technical Assistance Center, 2018). CYFAR provides grants to every state to implement innovative projects that serve at-risk families. The projects are planned at the state level, and thus there is no expectation of continuity or coordination of specific program activities from one state to another. Further, within a state, there is often diversity in how the program is shaped and delivered across counties, and even between community sites within counties.

This broad tapestry of projects, many of which share common aims, presents a daunting challenge for the task of measuring their impact on target outcomes. In response, a national CYFAR evaluation team developed a series of scales to be used as common measures across sites (Payne & McDonald, 2012, 2015). Targeted constructs that are intended for measurement

in every CYFAR project include program quality, youth program participation, and "core competencies" such as caring, decision-making, and social conscience. In addition, common measures are made available for specific program outcomes, such as leadership, parenting, nutrition, physical activity, and workforce preparation, to be used by local projects that target those outcomes (CYFAR Professional Development and Technical Assistance Center, 2019).

The use of common measures across program sites can be enormously valuable, by providing continuity and standardization in the evaluation process (Table 2). Results at different sites can be aggregated to allow for evaluative conclusions at the level of the broad program initiative, while different program delivery options can be compared for relative effectiveness. Thus, the credibility and actionability of evaluation evidence will often be considerably enhanced, especially for program funders and other stakeholders at levels of administrative and policy decision-making. For example, the use of common measures to assess critical program outcomes across states allows NIFA to report about the broad impacts of the CYFAR initiative to its parent agency, the U.S. Department of Agriculture. Common measures can enhance the credibility of evaluation data at the community site level as well, if it is communicated that the data stem from highly regarded, widely used instruments.

But Table 2 presents some cautions as well. The focus and delivery of programming at different sites within a project will usually not be uniform, due to either deliberate design or natural variations based on program personnel, location, and scheduling. Therefore, a commonly used measure may have differing degrees of relevance and importance at different sites. In cases where local site staff are given relatively broad latitude to make decisions about program focus and design, the use of externally imposed measures may be an uncomfortable force-fit. The program staff may be most interested in selecting measures that reflect their own priorities, and a requirement to use common measures may introduce extra time, redundancy, and/or irrelevance into the data collection process.

For example, Lewis, Horrillo, Widaman, Worker, and Trzesniewski (2015) examined the psychometric properties of four 4-H common measures, using exploratory factor analysis on responses from 721 California youth. They found that several scales had significant levels of missing responses due to the limited applicability of some items for their respondents (e.g., "I wear a helmet when riding an all-terrain vehicle," from the Healthy Living scale). Payne and McDonald (2015), in reviewing the CYFAR Initiative's common measures for parenting and youth citizenship, asked program staff from seven states to rate the relevance of each of the scale items to their own state CYFAR programs. Out of 22 items making up the two scales, only 66% were rated "completely relevant"; that is, about a third of the scale items did not closely align with the content of the local programs. In sum, the appropriateness of using common measures depends on the uniformity of goals, objectives, and program content across the program sites.

Table 2. Using Common Measures Across Program Sites: Benefits and Cautions

Potential Benefits
• **Consistent interpretation of constructs.** The consistency of measurement across program sites allows for consistent definition and interpretation of the program's core constructs. This is a strength—***if*** that consistency across sites is an expectation of the project. (See *Cautions* below.)
• **Reduced burden on local site staff.** Providing site staff with ready-to-go instruments makes data collection easier and can encourage evaluation practice at the sites.
• **Quality control in measurement.** A common measure that has been carefully developed or selected can help to ensure that best practices in measurement are being followed. However, evaluators must still pay attention to issues of validity and reliability at the local sites. (See *Cautions* below.)
• **Conclusions about collective impact and overall program accountability.** Using common measures for core constructs allows for evaluation conclusions to be drawn at a higher-order program level, combining results from individual sites. These conclusions can be highly credible and actionable for decision-making by funders, administrators, and legislators.
• **Comparisons across sites.** Community sites may differ, in large and small ways, in how program delivery takes place. Using common measures can allow for the effectiveness of program variations to be compared with a standardized measuring stick.
Cautions
• **Appropriateness for local site circumstances.** Multi-item scales used as common measures might include individual items that are not relevant for a local site, e.g., if they refer to content outside the scope of the site's curriculum. This could result in program sites being improperly evaluated based on irrelevant content.
• **Potential for redundancy.** For a variety of reasons, a local site might need to utilize its own instrument to measure the construct of interest, e.g., to track progress over time with an instrument used in prior years. In that case, the need to use the common measure in addition to the local measure may create redundancy in the measurement process.
• **Psychometric properties.** It cannot be automatically assumed that validity and reliability estimates for the common instrument will be adequate and equivalent at each site. If site participants differ in significant ways from the populations with which the instrument was validated, e.g., in terms of age, aptitudes, cultural background, etc., validity and reliability must be established for the local context.
• **Requirements for administering the measure.** Data collection using a common measure may require uniformity of process, e.g., with regard to time allocation, instructions, observation processes, etc. If those uniform procedures are not followed by site-level evaluators, the appropriateness of the measure may be compromised.

Implications and Recommendations for Extension Evaluation Practice

Based on the preceding discussion, a number of recommendations can be offered for Extension practice with regard to credible measurement in evaluation.

1. Prioritize measurement quality in evaluation planning, but know why you are doing it.

No evaluation or research study has unlimited resources. Data quality comprises many components, and even the most carefully conducted studies require decisions about the best that can be done under the circumstances. All study designs have weaknesses and limitations, and

Measurement and Credible Evidence

decisions about data quality involve trade-offs between reasonable options. Design weaknesses can be minimized, but they cannot be eliminated completely.

Measurement choices should be made with the goal of making the strongest possible case for a credible study. In many cases, there will not be a single correct choice. The evaluator needs to make decisions about how each of the available options affects the study's overall strength of evidence. In many cases, each option in a decision context will have its own strengths and weaknesses.

2. Monitor and communicate indicators of quality and rigor in the data.

Researchers have developed procedures and standard practices to help us understand the quality of our data. Assessing reliability, in its various forms, is one important strategy to accomplish that. Evaluators should also track and report the response rate for surveys and, if relevant, the degree of program attrition between pretest and posttest (that is, the number or proportion of participants who were included in the pretest sample but later dropped out of the program and were not part of the posttest sample). Those indicators provide background on how well our intended sample has been captured. Information should also be provided about missing data on individual measures, including the extent of missing scores and how that phenomenon has been handled in the analysis of data.

3. Engage with the Extension program's stakeholder audiences early in the evaluation planning process, to determine the factors influencing their judgments about the credibility of evaluation evidence.

To the extent possible, we should bring our Extension audiences into the evaluation planning process, where methods-related issues can be discussed, digested, and decided upon. There are several advantages to this approach. First, involvement in evaluation planning will increase their buy-in to the evaluation and their interest in the results. Second, they will provide perspectives and details that the evaluator can use in weighing options about evaluation design. Third, stakeholders involved in planning and measurement decisions will be more likely to understand and accept the evaluation results, even if those results are disappointing.

4. Educate your Extension evaluation audiences and promote evaluative thinking.

Extension audiences need to be reasonably educated and attentive about evaluation. If our audiences uncritically accept our evaluation conclusions, without attempting to understand the basis for our claims regarding the data, then our Extension clientele are not being well served, and neither are our programs or our organization. The stakeholders of Extension programs need to be partners in the decision process following a program evaluation. That is the basis for their need to think *evaluatively* about what has been learned and what actions are suggested by the evidence.

Journal of Human Sciences and Extension Volume 7, Number 2, 2019

5. Recognize and communicate that decisions about best practices with regard to measurement are often not clear-cut. Make thoughtful trade-offs between the need for high data quality and the availability of resources.

In many contexts, trying to decide how best to ensure quality can be an uncertain, ambiguous, and unpredictable process. Recommendations from the research literature or lessons from practitioner experience may be inconclusive. We often need to make decisions based on partial information. In these cases, decisions must be based on knowledge of the relevant research literature, our understanding of the target population, the lessons learned from our past practices, and the kinds of data quality limitations that we are most willing to tolerate. Well-informed evaluation stakeholders will be more willing to accept and be supportive of limitations in measurement rigor if the evaluator can demonstrate that the choices made were reasonable and logical in light of the reality of available resources.

6. Consider using common measures for the evaluation of programs with multiple sites.

The availability and use of common measures can offer significant benefits for the data quality and credibility of multi-site program evaluations, ***if*** there are basic commonalities and emphases across the sites that justify the standardization of measurement processes. Common measures can allow for broad statements about overall program impact using data that are aggregated across sites, and can also allow for direct comparisons between those sites. However, if there is significant variation in the goals and/or the programming content of those local sites, the use of common measures may not be appropriate and may present difficulties for understanding program outcomes.

Conclusion

In this paper, I have examined factors that contribute to the credibility of measurement in evaluation. The measurements, design, and implementation of an evaluation are largely under the control of the evaluator, but credibility also depends on the mindset, understanding, and sophistication of stakeholder audiences. Credibility has sometimes been discussed as an objective phenomenon, a stable attribute of a body of evidence, without consideration of the variability in people's perspectives. However, since credibility—or "believability"—implies that a judgment is being made by one or more audiences, it is a characteristic that exists, at least in part, in the eye of the beholder. Program stakeholders, including staff, administrators, funders, planners, professional associates, and participants, will play a critical role in determining how an evaluation gets interpreted and how it influences the fate of the program. In sum, the credibility of evaluation evidence depends on the intersection of how evidence has been created and the responses of the humans who determine what that evidence means.

References

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. New York, NY: Guilford.

Banta, T. W. (Ed.). (2004). *Hallmarks of effective outcomes assessment*. San Francisco, CA: Jossey-Bass.

Bazeley, P. (2017). *Integrating analyses in mixed methods research*. Thousand Oaks, CA: Sage.

Braverman, M. T. (1996). Sources of survey error: Implications for evaluation studies. In M. T. Braverman & J. K. Slater (Eds.), *Advances in survey research. New Directions for Evaluation, 70,* 17–28. doi:10.1002/ev.1032

Braverman, M. T. (2013). Negotiating measurement: Methodological and interpersonal considerations in the choice and interpretation of instruments. *American Journal of Evaluation, 34*(1), 99–114. doi:10.1177/1098214012460565

Braverman, M. T., & Arnold, M. E. (2008). An evaluator's balancing act: Making decisions about methodological rigor. In M. T. Braverman, M. Engle, M. E. Arnold, & R. Rennekamp (Eds.), *Program evaluation in a complex organizational system: Lessons from Cooperative Extension. New Directions for Evaluation, 120*, 71–86. doi:10.1002/ev.277

Buckley, J., Archibald, T., Hargraves, M., & Trochim, W. M. (2015). Defining and teaching evaluative thinking: Insights from research on critical thinking. *American Journal of Evaluation, 36*(3), 375–388. doi:10.1177/1098214015581706

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Thousand Oaks, CA: Sage.

Creswell, J. W., & Poth, C. N. (2018). *Qualitative inquiry and research design: Choosing among five approaches* (4th ed.). Thousand Oaks, CA: Sage.

CYFAR Professional Development and Technical Assistance Center. (2018). *CYFAR annual report 2017: Promoting the well-being of Children, Youth and Families At-Risk*. St. Paul, MN: University of Minnesota. Retrieved from https://cyfar.org/resource/2017-cyfar-annual-report

CYFAR Professional Development and Technical Assistance Center. (2019). *CYFAR approved common measures*. St Paul, MN: University of Minnesota. Retrieved from https://cyfar.org/ilm_common_measures

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method* (4th ed.). Hoboken, NJ: Wiley.

Donaldson, S. I. (2015). Examining the backbone of contemporary evaluation practice: Credible and actionable evidence. In S. I. Donaldson, C. A. Christie, & M. M. Mark (Eds.), *Credible and actionable evidence: The foundation for rigorous and influential evaluations* (2nd ed., pp. 3–26). Thousand Oaks, CA: Sage.

Donaldson, S. I., Christie, C. A., & Mark, M. M. (Eds.), *Credible and actionable evidence: The foundation for rigorous and influential evaluations* (2nd ed.). Thousand Oaks, CA: Sage.

Goodrick, D., & Rogers, P. J. (2015). Qualitative data analysis. In K. E. Newcomer, H. P. Hatry, & J. S. Wholey (Eds.), *Handbook of practical program evaluation* (4th ed., pp. 561–595). Hoboken, NJ: Wiley.

Groves, R. M. (1989). *Survey errors and survey costs*. Hoboken, NJ: John Wiley & Sons.

Höglinger, M., Jann, B., & Diekmann, A. (2016). Sensitive questions in online surveys: An experimental evaluation of different implementations of the randomized response technique and the crosswise model. *Survey Research Methods, 10*(3), 171–187. doi:10.18148/srm/2016.v10i3.6703

Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. In M. T. Braverman & J. K. Slater (Eds.), *Advances in survey research. New Directions for Evaluation, 70*, 29–44. doi:10.1002/ev.1033

Lewis, K. M., Horrillo, S. J., Widaman, K., Worker, S. M., & Trzesniewski, K. (2015). National 4-H common measures: Initial evaluation from California 4-H. *Journal of Extension, 53*(2), Article 2RIB3. Retrieved from https://joe.org/joe/2015april/rb3.php

Lohse, B., Wall, D., & Gromis, J. (2011). Intention to consume fruits and vegetables is not a proxy for intake in low-income women from Pennsylvania. *Journal of Extension, 49*(5), Article 5FEA5. Retrieved from https://joe.org/joe/2011october/a5.php

Mark, M. M. (2015). Credible and actionable evidence: A framework, overview, and suggestions for future practice and research. In S. I. Donaldson, C. A. Christie, & M. M. Mark (Eds.), *Credible and actionable evidence: The foundation for rigorous and influential evaluations* (2nd ed., pp. 275–302). Thousand Oaks, CA: Sage.

Mertens, D. M. (2018). *Mixed methods design in evaluation*. Los Angeles, CA: Sage.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). Washington, DC: American Council on Education.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). Thousand Oaks, CA: Sage.

Miller, R. L. (2015). How people judge the credibility of information: Lessons for evaluation from cognitive and information sciences. In S. I. Donaldson, C. A. Christie, & M. M. Mark (Eds.), *Credible and actionable evidence: The foundation for rigorous and influential evaluations* (2nd ed., pp. 39–61). Thousand Oaks, CA: Sage. doi:10.4135/9781483385839.n4

National 4-H Council. (2019). *Common measures*. Retrieved from https://4-h.org/professionals/common-measures/

Patton, M. Q. (2018). A historical perspective on the evolution of evaluative thinking. In A. T. Vo & T. Archibald (Eds.), *Evaluative thinking. New Directions for Evaluation, 158*, 11–28. doi:10.1002/ev.20325

Payne, P. B., & McDonald, D. A. (2012). Using common evaluation instruments across multi-state community programs: A pilot study. *Journal of Extension, 50*(4), Article 4RIB2. Retrieved from https://joe.org/joe/2012august/rb2.php

Measurement and Credible Evidence 107

Payne, P. B., & McDonald, D. A. (2015). Common evaluation tools across multi-state programs: A study of parenting education and youth engagement programs in Children, Youth, and Families At-Risk. *Journal of Extension, 53*(3), Article 3FEA5. Retrieved from https://joe.org/joe/2015june/a5.php

Perinelli, E., & Gremigni, P. (2016). Use of social desirability scales in clinical psychology: A systematic review. *Journal of Clinical Psychology*, *72*(6), 534–551. doi:10.1002/jclp.22284

Schwandt, T. A. (2015). *Evaluation foundations revisited: Cultivating a life of the mind for practice*. Stanford, CA: Stanford University Press.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press.

Vo, A. T., & Archibald, T. (Eds.). (2018). *Evaluative thinking. New Directions for Evaluation, 158*. doi:10.1002/ev.20317

Marc T. Braverman is a Professor in the School of Social and Behavioral Health Sciences and an Extension Specialist in the Family and Community Health program at Oregon State University. His research interests include evaluation theory, measurement, applied research methods, community program development, tobacco control policy, and adolescent health.

Acknowledgments

I am grateful to Ben Silliman for his suggestion of an earlier version of the model that appears on page 2. I also thank my Oregon State University colleagues Kathy Gunter and Shauna Tominey for their very helpful suggestions about sections of Table 1.

Journal of Human Sciences and Extension Volume 7, Number 2, 2019