jmde
Journal of MultiDisciplinary Evaluation

http://www.jmde.com *Articles*

The Logic of Summative Confidence

P. Cristian Gugiu
Western Michigan University

The constraints of conducting evaluations in real-world settings often necessitate the implementation of less than ideal designs. Unfortunately, the standard method for estimating the precision of a result (i.e., confidence intervals [CI]) cannot be used for evaluative conclusions that are derived from multiple indicators, measures, and data sources, for example. Moreover, CIs ignore the impact of sampling and measurement error. Considering that the vast majority of evaluative conclusions are based on numerous criteria of merit that often are poorly measured, a significant gap exists with respect to how one can estimate the CI of an evaluative conclusion. The purpose of this paper is (1) to heighten reader consciousness about the consequences of utilizing a weak evaluation design and (2) to introduce the need for the development a methodology that can be used to characterize the precision of an evaluative conclusion.

One of the principle lessons impressed upon students in introductory methodology courses is that "a weak design yields unreliable conclusions." While this is certainly true, the constraints of conducting research and evaluation studies in real-world settings often necessitate the implementation of less than ideal studies (Burstein, Freeman, Sirotnik, Delandshere, & Hollis, 1985). For example, evaluators and researchers may have no choice but to implement a study that has a small sample size, includes subjects with high heterogeneity, employs instruments with moderately low reliability and validity, implements procedures that produce high measurement error, or utilizes minimal triangulation. In such instances, investigators are left to debate the trade-offs (e.g., time and cost) associated with modifying a weak design (i.e., one that has a moderately high probability of producing an incorrect conclusion) or sacrificing the precision of their conclusions.

The purposes of this paper are twofold: (1) to heighten reader consciousness about the consequences of utilizing a weak evaluation design and (2) to introduce the need for developing a methodology that can be used to characterize the precision of an evaluative conclusion. Careful consideration of the limitations of certain evaluation practices, it is hoped, will sensitize evaluators to the need to include necessary safeguards in planning studies. Moreover, consideration by decision makers of the degree of confidence one may place on an evaluative conclusion, herein referred to as *Summative Confidence*, may alert them to whether they need to take immediate action to correct a serious problem, reward a successful program, or seek further evidence of the merit and worth of the evaluand (i.e., entity under investigation).

It is important to note that this paper will not present the mathematical algorithm that can be used to conduct a Summative Confidence analysis nor provide details on how such an analysis may be conducted. These areas will be covered in future publications by the author, most notably in his dissertation. Instead, the paper will provide a conceptual framework for Summative Confidence. However, the majority of the design factors presented herein and the

claims regarding its usability have been tested in an actual evaluation study by the author. Furthermore, all of the individual design factors have a long history in the research literature. Therefore, the foundation upon which Summative Confidence rests is stronger than just the author's wishes and aspirations.

Statement of the Problem

In scientific circles, research and professional evaluation, herein simply referred to as evaluation, conclusions are persuasive to the extent to which they lack error (i.e., are precise).[1] One method of expressing the precision of a conclusion is through the use of a confidence interval—a practice recommended by leading research organizations, for example, the American Psychological Association (APA)(Wilkinson & APA Task Force on Statistical Inference, 1999). Typically, the method used to determine the precision of a result is the size of the interval. Large intervals suggest that a result is imprecise (i.e., has a large amount of error) whereas small intervals indicate the opposite. Similarly, the confidence level associated with an interval communicates the probability of reaching an incorrect conclusion. Therefore, small intervals that have a low confidence level are not very impressive. As important as confidence intervals (CI) can be for reporting precision and confidence, the analytical method that is used to calculate a CI suffers from one important limitation: It can only calculate a CI for a unidimensional dependent variable. Thus, it cannot be used to calculate a CI for a composite variable. Unfortunately, a large portion of evaluation practice (e.g., summative evaluation) entails the formulation of evaluative conclusions based on numerous criteria or dimensions of merit or worth. Therefore, a significant gap exists with respect to how one can estimate the degree of confidence that should be placed on an evaluative conclusion when such conclusions are the product of a complex synthesis of multiple factors.

Further compounding this problem is the data synthesis dilemma. Evaluation practice often requires the synthesis of qualitative and quantitative data into an overall conclusion, that is, a summative conclusion. However, because different analytical rules and scales underlie each of these approaches, no method has been proposed for calculating the precision of a conclusion. Moreover, the process of transforming one data type into another complicates the ability of calculating the precision of the summative conclusion.[2] For example, suppose a professor needed to assign a final grade to a student who received a C on a term paper and an A on a multiple-choice exam. To what degree is the precision of the student's final grade a function of the weight assigned to each individual grade? In addressing this question, two factors should be considered: the weighting scheme and the two grades. In general, the grades assigned to written assignments are less reliable than those assigned to quantitatively scored exams (e.g., multiple-choice, true-false) because the proportion of

[1] The term "precise" will be used throughout this paper to denote the degree of error with which a variable is estimated. Highly precise estimates have less error (i.e., smaller confidence intervals) whereas imprecise estimates contain more error. From a statistical perspective, this is not synonymous with accuracy, which measures the degree of discrepancy between an estimated value and the actual value. Therefore, a result can be measured with a high degree of precision but produce an inaccurate result. For the purpose of readability, readers who prefer the term "accuracy" may substitute it without significantly altering the intended meaning of the concepts discussed herein.

[2] Because it is often simpler to reduce greater detail to less detail, rather than the reverse, quantitatively oriented analysts transform qualitative data into binary, binomial, or ordinal data. However, the reverse process is also possible. Qualitatively oriented analysts may convert quantitative data into qualitative data through a process of interpretation and labeling. For example, a quantitative IQ score of 160 may be interpreted and labeled as superior whereas an IQ score of 70 may be classified as below average.

error variance is greater in the scores of the former than in the scores of the latter (Hopkins, 1998). Therefore, the final grade will be more precise if the student is assigned a B+ (multiple-choice exam is given more weight) rather than a B- (term paper is given more weight)—assuming both tests are equally valid measures of the student's academic ability. Unfortunately, no method exists that quantifies the difference in CIs between the two possible grades that could be assigned.

Similar issues arise with regard to sampling. While the family of randomized sampling is widely regarded as the "gold standard" for the purpose of generalizing results from a sample to the population (Cook & Campbell, 1979; Kish, 1995), the impact of sampling error on evaluative conclusions appears to have been overlooked. That is, most researchers and evaluators acknowledge that selecting a small sample (say fewer than 100) out of a much larger population (say more than 1,000) limits the degree to which one may generalize a result to the entire population. However, many fail to recognize that the larger the sampling error (i.e., deviation of the sample estimate from the true score), the lower the precision of their estimates and thus their final conclusions. In another words, sample statistics (e.g., means, variances) are only approximations of true scores (i.e., population parameters). Therefore, unless evaluators wish to confine their conclusions only to the sample, they must account for sampling error in order to reach conclusions about population parameters. Unfortunately, similar to the prior example, no method exists for quantifying the impact of sampling error on the confidence one can place on a summative conclusion.

The process of formulating a conclusion may also require comparison against a known or constructed standard. For example, while the ability of two graduate students, one with a 2.95 GPA and one with a 3.00 GPA, may be nearly identical, the conclusions one would reach about each student would differ when compared against a university's minimum standard of acceptable academic performance (generally set at a 3.00 GPA). In the case of the former student, one would conclude that the student failed to meet the minimum expectation while in the latter case one would conclude the reverse. However, how accurate is the conclusion that the latter student's ability meets or exceeds the minimum expectation? Given their proximity to the standard, it is safe to conclude that one would be less confident that the second student met or exceeded the standard than had they earned a 3.60 GPA. Therefore, the degree of precision of a conclusion is inversely related to the difference between performance and the standard. While methods exist for calculating a confidence interval for such cases (Crocker & Algina, 1986), no method exists for estimating the impact of such cases on composite variables.

Finally, common sense dictates that the more information one knows about an evaluand, the more confident one may be in the conclusion formulated about the evaluand. Similarly, the wider the array of methods used to collect information about the evaluand and the data sources from which information is collected (i.e., triangulation), the greater the precision of one's conclusions. For example, if one wishes to know the weight of an object, one could simply weigh the object on a scale. If the scale was error-free, only one weigh-in would be necessary. However, because scales do not measure weight with perfect precision, one should place more confidence in the estimate provided by the scale with the lowest measurement error (providing this information is available) or the average of all the estimates. In the majority of scenarios, the choice of selecting the instrument with the lowest measurement error is not possible due to the multidimensional nature of latent constructs.[3] In

[3] In other words, because no instrument can measure the entire latent construct, multiple instruments and data sources will need to be combined to measure the construct. For example, a composite variable that

such instances, one must utilize several instruments to measure the construct in its entirety. This, of course, raises the question, should one have more confidence in a conclusion that was formulated from instruments that measured unique dimensions of the latent construct or from instruments that measured highly correlated dimensions? To date, however, no method has been able to express the exact relationship between the precision of a result and the amount of triangulation used to formulate the result.

Background

Logic Underlying Summative Confidence

One may think of Summative Confidence as the mathematical degree of confidence that one can place on an evaluative conclusion that was derived from a synthesis of the performance of the evaluand on multiple criteria of merit and worth (i.e., the product of a summative evaluation). More specifically, it refers to the band of error surrounding an evaluative conclusion given a specified level of confidence. Therefore, if one was to replicate the evaluation ad infinitum, a distribution of sample conclusions would form around the true or correct conclusion. Summative Confidence refers to the band of uncertainty placed around a sample conclusion at a specific probability (i.e., confidence level). Clearly, smaller confidence bands indicate that the evaluative conclusion was estimated with greater precision whereas larger bands indicate the reverse.

Two types of confidence intervals may be calculated: one in which the location of the interval is a function of the estimated parameter (e.g., the mean) and one in which the location of the interval is fixed. In the case of the former, the confidence level is fixed while the location of the interval changes with each replication of the study. Therefore, Summative Confidence refers to the proportion of intervals that would contain the true conclusion if the evaluation was repeated an infinite number of times. For example, a teacher who calculates with 99 percent certainty that the true performance (i.e., ability) of one of her students falls between an A- and an A+ can feel very confident about her grading scheme and according the student an A for the course. However, had there been a 99 percent chance that the student's true performance fell between a C and an A then the teacher should feel less confident about her grading scheme and giving the student a B because of the greater imprecision of her estimate.

In the case of the second type of Summative Confidence, the location of the interval is fixed and the confidence level varies according to the proportion of sample conclusions that fall within the fixed interval, if the evaluation was repeated an infinite number of times. Therefore, if a high proportion of sample conclusions fall within the specified range, one may take solace in that the evaluation methodology yields a replicable conclusion. Returning to the previous example, if the teacher calculates that there is a 75 percent likelihood that the student's true performance is, at least, a B or higher, she could be fairly confident about the reliability of her grading scheme. Of course, there is a 25 percent chance that her grading scheme erroneously produced an inflated grade. In certain evaluations, such a low level of confidence may call into question the reliability of the entire methodology.

Before going further, it is important to note that Summative Confidence is distinct from existing analytical methods (e.g., meta-analysis, multiple regression) for which one can construct a CI for the dependent variable. First, meta-analysis combines the results of *several* studies that address a set of related research hypotheses, whereas Summative Confidence

considers behavioral, cognitive, and biological indicators derived from multiple sources will be a more accurate measure of depression than any unidimensional measure.

focuses on a *single* evaluation.[4] Second, because a summative conclusion is directly calculated from the criteria of merit and worth (i.e., the independent variables), measurement of the impact of the independent variables on the dependent outcome would result in a model in which $R^2 = 1$. Therefore, the model would be completely predictive of the dependent variable and the CI would be zero. Finally, traditional methods for calculating CIs do not account for design characteristics, such as sampling and measurement error. Therefore, such CIs tend to overestimate the precision of the study design.

One may wonder, however, given the vast methodological variability that exists across studies, how can one hope to be able to develop a methodology that can be applied in every study? The foundation of Summative Confidence rests upon two principles. First, everything can be measured; it is just a matter of precision. For example, a doctor checking a patient for high blood pressure (hypertension) could look for typical symptoms such as severe headaches, fatigue or confusion, vision problems, chest pain, difficulty breathing, irregular heartbeat, and blood in the urine (Chang, 2005). However, since a large proportion of the people afflicted by this disease have no symptoms, a diagnosis of hypertension based on the presence or absence of symptoms alone is likely to be error-prone. A more precise diagnosis of hypertension can be obtained by using a sphygmomanometer. Therefore, evaluands measured with precise instruments and methods yield conclusions that contain less measurement error.

Second, the degree of measurement error in a summative conclusion is a function of the measurement error of all of the elements used to formulate the conclusion. This leads to one of the basic principles of Summative Confidence, which should be familiar to all computer programmers: "garbage in, garbage out." Stated more formally, if the criteria of merit and standards from which the summative conclusion is formulated are measured with a high degree of error, then little confidence should be placed on the summative conclusion. However, some reprieve may be gained from triangulation. While little confidence should be placed in a conclusion derived from data containing a high degree of measurement error, hope does exist for conclusions derived from several indicators that were measured with a small or moderate amount of error. This is because information is cumulative and the composite measure generally provides a more accurate explanation of the construct than its constituent parts.[5] Therefore, a second basic principle of Summative Confidence is that the more information one has upon which to base a conclusion, the more confident one can be in the conclusion.

Factors that Impact Summative Confidence

The Summative Confidence of an evaluative conclusion is contingent upon the measurement error introduced into the study by the choices an evaluator makes regarding sampling scheme, instrument selection, and methodological design. For example, to the extent to which sampling error is largely due to a small sample size or heterogeneity, the degree of confidence that can be placed on the interval surrounding a conclusion will be low (Hays, 1994). To the

[4] This may not always be true, for there is reason to believe that Summative Confidence can be extended to multiple studies. However, to the best knowledge of the author, the reverse cannot be said for meta-analysis.

[5] This is not always the case. For example, if the information provided by a set of indicators is redundant, then a composite measure of these indicators will not be more accurate than any individual indicator. However, this assumes not only that each indicator is perfectly correlated with one another, but that they have identical correlations with the construct. If one of the indicators had a stronger correlation with the construct and was perfectly correlated with the other indicators (think of a Venn diagram in which the other indicators are a subset of this indicator which, in turn, is a subset of the construct), then this indicator would be the most accurate measure of the construct.

extent to which instruments are unreliable or poor agreement is attained between raters of qualitative data, the standard error of the conclusion will be large (Crocker & Algina, 1986). To the extent to which few values are measured or greater weight is assigned to poorly measured values, Summative Confidence will be negatively affected.[6] More specifically, the degree of confidence one can place on an evaluative conclusion depends upon the following 11 family of factors:

1. Family Type I Error (alpha): The probability that the true score of the parameter being estimated (e.g., the summative conclusion) falls outside of the estimated CI.
2. Values: The number, organizational structure, and correlation between the criteria of merit and worth that are used to formulate an evaluative conclusion about the performance of an evaluand.
3. Standards: The variability at which the performance benchmark is set for a criterion—which is deemed to be of critical importance to the overall performance of the evaluand—that demarks acceptable from unacceptable or excellent from less than excellent performance.
4. Effect size: The difference between the performance of the evaluand on a criterion and the standard set for the criterion.
5. Sample size: The size of the sample taken from the population of impactees or decision makers.
6. Heterogeneity: Individual differences between the stakeholders from whom data are being collected.
7. Measurement error: The difference between a measured result and its true score.
8. Sampling error: The difference between the result produced by a sample and the result produced by a population.
9. Construct validity: The correlation between a micro-value and its corresponding macro-value.
10. Triangulation: The amount of information collected from multiple data sources or data collection techniques that are used to measure a micro- or macro-value.
11. Weighting scheme: The amount and variability of the importance accorded to each micro- and macro-value.

Illustrative Example: Recommending a Faculty Member for Tenure

To more fully appreciate the complexity of the factors that contribute to Summative Confidence, a more realistic illustration may be helpful. Please note, however, that although the following example focuses on an academic personnel evaluation, the Summative Confidence algorithm can be utilized for any type of summative evaluation (e.g., consumer reports, promotion decisions, parole board decisions, medical diagnoses). Furthermore, for the purpose of this paper, the accuracy of the process described in the following example is unimportant. What is important is the role each of the aforementioned factors play in determining the precision of the final conclusion.

Suppose a university provost was interested in evaluating the university's tenure review process by calculating the Summative Confidence of a randomly selected case. Examination of the case revealed that the decision was reached after an exhaustive deliberation about the applicant's performance on numerous values, including research, teaching, service, professional accolades, academic interests, and collegiality. The provost also learned that prior to the start of the process, a panel of faculty members deliberated about which factors were critically important to the decision, the weight assigned to critically and noncritically important factors, the standards used to judge acceptable performance

[6] In keeping with the evaluation-specific terminology defined by Scriven (1991), the term "values" will frequently be used throughout this paper to denote criteria of merit, rather than to signify a numeric score.

on factors identified as critically important, and the standard used to arrive at a decision based on a synthesis of the data. Finally, to ensure that the ratings of faculty members were not unduly influenced by "stronger" members within the group, all ratings were anonymous.

The tenure review process began with a meeting between five tenured faculty members from within the department and five randomly selected, tenured faculty members from outside the department. During the first meeting, the review panel generated and agreed upon a list of criteria upon which to judge the merits of the candidate. This decision was the first of several decisions that impacted the precision of the final decision of whether or not to recommend the candidate for tenure. Although the process of deliberating over criteria and their importance is common, most evaluators treat the agreed-upon decisions derived from such processes as unequivocal when, in fact, unanimous agreement does not always, or even typically, exist. Clearly, the greater the disagreement over the macro-values that should be considered in the evaluation, the lower the likelihood that the same conclusion could be replicated by a different panel of faculty or even by the same faculty at a different point in time. Likewise, the lower the agreement over which macro-values should be considered critically important, the weight that should be applied to each macro-value, and the level at which a standard for a critically important macro-value should be set, the lower the probability that the final decision could be replicated. From a Summative Confidence perspective, the most accurate procedure would be for the university or department to devise a uniform policy or for the faculty to take steps to increase consensus (i.e., inter-rater reliability) among themselves on these matters. Of the two alternatives, policy decisions are likely to improve Summative Confidence to a greater extent because they place greater limits on rater disagreement.

In addition to the aforementioned factors, the faculty's decisions regarding the number of micro-values selected, the macro-value structure within which these micro-values were organized, and the degree of redundant information shared by the micro- and macro-values influenced the precision of the final decision. The common attribute underlying each of these factors is information. As stated previously, the more information (e.g., number of micro-values examined) one has upon which to base a conclusion, the more confident one can be in the conclusion reached.

Figure 1 presents the list of micro- and macro-values that were used by the faculty to render a decision of whether or not to recommend the candidate for tenure. Macro-values are represented by a square while the micro-values are organized underneath the macro-value with which they are associated. Furthermore, these values were organized into critically and non-critically important values, with greater weight assigned to the former group of values.

As illustrated in the figure, some macro-values were measured by a greater number of micro-values than other macro-values (e.g., Research versus Teaching). Therefore, the precision of the conclusions reached about these macro-values should exceed the precision of conclusions derived from poorly measured macro-values, all other factors being equal. The organization of micro-values also significantly influenced precision because more micro-values were used to measure performance of noncritically important macro-values than were used to measure performance of critically important macro-values. Moreover, the candidate's performance on the latter group of macro-values weighed more heavily on the final decision than the former group of macro-values. It stands to reason that the greater the precision with which a macro-value is measured, the more confident one may be in the conclusions reached about the macro-value. Therefore, one can improve confidence by using micro-values with lower levels of measurement error, increasing the number of

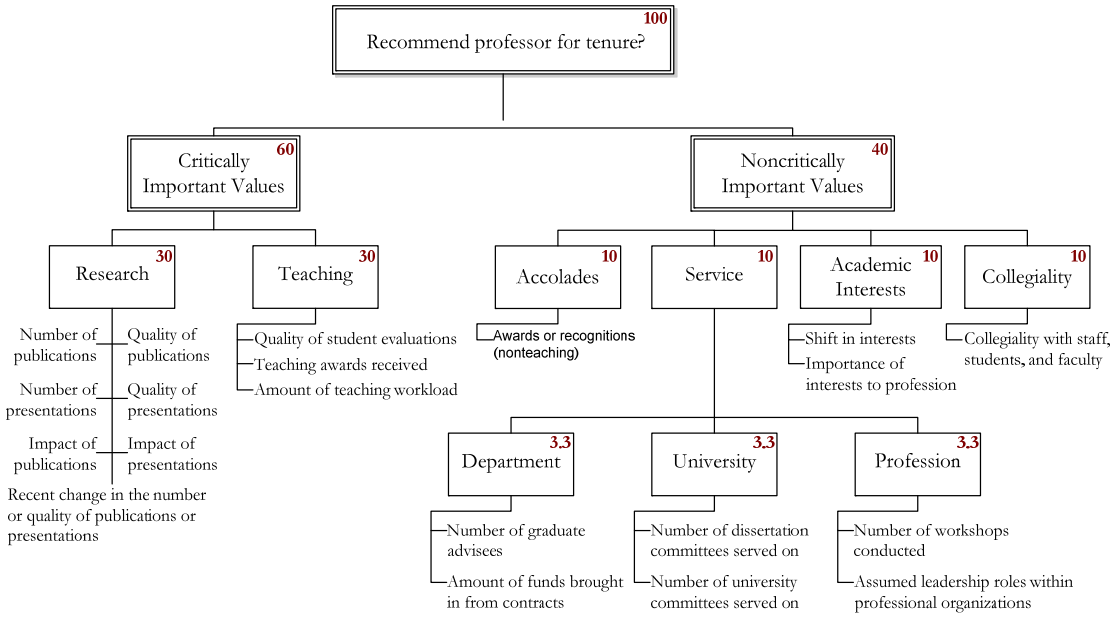

Figure 1. Values Used to Determine Whether the Faculty Member Should be Recommended for Tenure

Note. The numbers in the top right-hand corners of each box represent the weight assigned to the macro-value.

micro-values and methods used to measure a macro-value (i.e., triangulation[7]), and assigning more weight to precisely measured macro-values.

Although somewhat counterintuitive, another method for improving precision is by reducing the amount of redundant information between micro-values and macro-values. For example, if the three collegiality indicators were highly associated with each other—which is entirely plausible—they would have contributed less information to the precision of the conclusion about collegiality than the three teaching indicators—which were probably only modestly related with one another—contributed to the precision of the conclusion about Teaching, all other factors (e.g., weights, measurement error, scaling) being equal. Similarly, a high association between macro-values, e.g., the two critically important macro-values, would have lowered the degree of confidence of the overall group-level conclusion (e.g., all critically important values) compared with unassociated macro-values.

Despite the implicit suggestion embedded in the previous two paragraphs, the relationship between values and Summative Confidence is not linear. If it were, then one could simply improve the Summative Confidence of a conclusion by adding unrelated micro-values. For example, the faculty could have increased the Summative Confidence of their conclusion regarding the candidate's teaching ability by adding micro-values such as showed up to class on time, turned in grades on time, liked by students, failed few students, and so forth. While each of these criteria is related with teaching, none of them are strong indicators of teaching proficiency.

[7] Macro-values are generally composed of multiple dimensions. Therefore, increasing the number of micro-values with which a macro-value is measured will improve Summative Confidence, provided that these micro-values measure distinct dimensions of the macro-value.

An even more extreme example would have occurred if the faculty included completely unrelated criteria (e.g., attractive, well groomed) that would have altered both the conclusion and its precision. Therefore, although micro-values should be unrelated, they must be valid indicators of the macro-value they purport to measure. However, this dual standard is difficult to attain. In many instances, the best one can hope for is a set of indicators that are marginally associated with each other and moderately associated with the macro-value.

Another factor that impacted the precision of the summative conclusion was the panel's decision regarding the weighting scheme. A 100 point weighting scheme was used in which 60 points were allocated to the critically important values and 40 points were allocated to the noncritically important values, where points were redistribute evenly to every macro- and micro-value underneath the two groups. As a result of this decision, critically important macro-values had 1.5 times the impact on the precision of the summative conclusion than their counterparts, individual teaching micro-values had a greater impact on the precision with which critically important values were measured than individual research micro-values (10% for teaching versus 4.29% for research), the accolades micro-value had a greater impact on overall precision than any individual research micro-value due to the distribution of weights among micro-values, and so forth. Finally, because the weighting scheme was not prescribed by the department or university, it would have a profound impact on the replicability of the decision if there was great variability between the weighting schemes each faculty member generated before agreeing to the final scheme.

In addition to their agreement on a weighting scheme, the faculty agreed upon a set of standards for some of the critically important values. Specifically, they decided to not recommend the faculty person for tenure if she or he did not have at least one publication per year in a peer reviewed journal and had not presented at a conference once every two years. The impact of a standard on Summative Confidence cannot be summarized in a single statement. For one, the magnitude of the impact will depend upon the type of standard set. Scriven (2007) and Davidson (2005) have identified three types of standards: soft-hurdle, hard-hurdle, and bar. Essentially, these standards differ in the penalty exerted on the summative conclusion. However, all of them require ignoring some information about the performance of the evaluand on one or more dimensions of merit. For example, if the panel set a soft-hurdle on the frequency of conference presentations and the candidate failed this standard, the faculty would have to ignore all of the candidate's presentations, essentially giving the candidate no credit for their performance on this micro-value. The penalty for failing a hard-hurdle is even more stringent. The faculty would have to ignore all of the candidate's performance on the research macro-value. Likewise, in the case of the last standard, failure of a bar would result in the failure of the entire evaluand, i.e., ignoring the impact of all passing values.

Clearly, these penalties can have a significant impact on a summative conclusion but what effect do they have on Summative Confidence? The impact of failure on a standard on Summative Confidence is similar to the impact that failure has on the conclusion in that the impact of failing a soft-hurdle will be smaller than the impact of failing a hard-hurdle, which, in turn, will be smaller than the impact of failing a bar. More specifically, the Summative Confidence of a value on which the evaluand failed the standard is a function only of the evidence that supports failure (i.e., evidence of positive performance on the dimension(s) impacted by the standard is ignored). Therefore, in the case of soft-hurdles, the confidence level associated with concluding that the evaluand failed a specific criterion is a function of the precision with which that

criterion is measured. Similarly, in the case of hard-hurdles and bars, the confidence level associated with concluding that the evaluand failed the macro-value or evaluand is a function of the precision with which the composite of failed criteria are measured. Therefore, one may be more confident in concluding that the evaluand failed when a greater number of criteria support this conclusion.

Even when the evaluand does not fail a performance standard, its effect on Summative Confidence may be observed in its impact on the effect size. According to one of the principles of measurement theory, the reliability of a criterion test is a function of the discrepancy between one's performance and the cutoff score (Crocker & Algina, 1986). The closer one's performance is to the cutoff, the lower the reliability of the decision reached based on the test. Therefore, it stands to reason, the closer one's performance is to a standard (i.e., the smaller the effect size), the lower the Summative Confidence. This invites the possibility of setting really low standards so as to increase the Summative Confidence of a conclusion. However, the gain in confidence would occur at the expense of validity. Therefore, such sacrifices should never be made.

Another factor, and perhaps the most important one, that impacts Summative Confidence is measurement error. Although this factor has been mentioned on several occasions, the nature of its relationship with Summative Confidence has yet to be specified other than to state that the two concepts are inversely related. Measurement error refers to the discrepancy between a measurement and the true score of the entity being measured. It is expressed either as the standard error of measurement (or mean), the standard error of estimate, or as the reliability of a measure or method. The standard error of a measurement is an estimate of the average discrepancy between a measurement score and the true score. Similarly, the standard error of estimate refers to the average discrepancy between a measurement score and the predicted score on a parallel measure. Reliability, on the other hand, is the degree to which a method consistently reproduces the same result. Therefore, lower standard errors and higher reliabilities are each indicative of greater measurement precision.

Returning to the tenure review example, the measurement error of the collegiality macro-value is the discrepancy between the candidate's true collegiality and the degree of collegiality that they possessed in their interactions with staff, students, and other faculty. Considering that individuals may interact with people in a variety of ways, it would not be surprising if the estimate of a candidate's collegiality had a modest amount of measurement error. Furthermore, despite the fact that only one indicator exists to measure the candidate's professional accolades, this micro-value is likely to produce a more accurate estimate of the respective macro-value than the synthesis of the three collegiality micro-values. This is because the measure of the candidates' accolades will only require the counting of their awards—a list of which would not be difficult to obtain and verify. Therefore, no measurement error should exist, assuming agreement exists on what an accolade is. However, if the panel wanted to consider the prestige of each award, measurement error would be introduced into the estimate due to potential disagreements over the prestige of each award.

In general, measurement error is likely to exist whenever interpretation is necessary to transform data from one type to another. Measurement error in these instances refers to the degree of agreement over a set of interpretations or ratings. Two types of errors appear in the literature. Inter-rater reliability refers to the degree of consistency in the ratings of the same entity made by several raters, whereas intra-rater reliability, commonly called test-retest reliability, refers to the degree of consistency in the ratings of the same entity made by a single rater. Both estimates presume

that the conditions under which ratings are made are as similar as possible; otherwise, the degree of consistency between ratings and raters would be a function of the precision of the instrument or method that produced the rating as well as any contextual factors that might influence the rating. Unfortunately, while it is fairly common practice for evaluators to report the reliabilities of their instruments and methods, no evidence exists that these estimates are utilized to adjust the confidence intervals of the parameters they calculate. This is particularly true whenever qualitative analysis is conducted because there is no valid method of estimating a confidence interval around a conclusion, unless the data are quantified and statistical analyses are performed. One can, however, say that lower instrument and method reliabilities produce a lower Summative Confidence.

Several factors influence the amount of error with which a variable is measured. The heterogeneity of stakeholders influences the amount of random error that is introduced into performance estimates. For example, the candidates' performance on collegiality is likely to contain more measurement error than their performance on student evaluation ratings, in part, because a sample of staff, students, and faculty is undoubtedly more heterogeneous than a sample of only students, all other factors (e.g., sample size) being equal. Another often overlooked factor is sampling. Sampling not only dictates the degree to which a result can generalize beyond the sample, but also the amount of sampling error that it incorporates. In the faculty tenure review example, sampling error was likely very high due to the small number of faculty who were randomly sampled from the university's faculty population.

One way of combating heterogeneity, sampling error, and virtually any other factor that weakens Summative Confidence is by increasing the sample size. With a large enough sample size, virtually any level of precision or confidence can be attained. However, while theoretically one can improve confidence up to 100 percent by adding to the sample, practical limits (e.g., cost) make this level virtually impossible to attain. Another limitation is the need to distribute the sample to both the measurement of performance and the construction of standards.[8] Even if the faculty obtained an accurate measure of the candidate's teaching ability from student evaluations—presumably due to a large sample size—the Summative Confidence may still be low if only the 10 faculty members decided where to set the standard. In another words, the measurement error associated with the standard may be so large as to offset the precision of the measure of the candidate's teaching ability.

Finally, one of the most important factors to consider in a Summative Confidence analysis is alpha—the confidence level at which the Summative Confidence analysis is conducted. Setting alpha to 10 percent indicates that the analysis will calculate the confidence interval for the summative conclusion such that if the study were conducted ad infinitum, 90 percent of the calculated intervals would contain the true evaluative conclusion. However, alpha and the width of the confidence interval are inversely related. The lower the alpha, the wider the confidence interval will be. Inversely, if one would like a "tight" CI, one would need to accept a lower probability that the evaluation methodology could produce a correct estimate (i.e., assume a lower confidence level).

Understanding the confidence level of a decision potentially has great implication. To illustrate this point further, consider the implications of setting the standard for the summative conclusion to 70 percent. In other words, the tenure review panel would recommend the candidate for tenure if, and only if, the candidate's overall performance score was 70 percent or higher. If the candidate received a score of 77 with a 90 percent CI that ranged from 74 and 79 percent, the provost

[8] Distribution of the sample to the construction of standards is only necessary when such standards cannot be derived from the literature, policy, or logical inference.

could feel reassured in the reliability of the tenure review process. However, what if the Summative Confidence analysis produced a 90 percent CI that ranged from 60 to 85 percent?[9] In this situation, the provost would have reason to question the reliability of the process. An alternative, and equally valid, method of interpreting Summative Confidence is to calculate the probability for a given confidence interval. For instance, the provost may not be as interested in knowing the confidence interval around the performance estimate as much as knowing the probability that the decision reached is correct. In this example, a decision to recommend for tenure is correct if the candidate's true performance is 70 percent or higher. Therefore, the provost may wish to know, "what is the probability that the candidate deserves to be given tenure (i.e., has a performance score of 70% or higher)?" If the probability turns out to be more than 90 percent, then the provost may conclude that the tenure review system is very reliable; otherwise, she will need to take steps to improve the system or face the possibility of lawsuits claiming that the process produced arbitrary or biased results.

So what has Summative Confidence taught the provost about her university's tenure review process? One would imagine quite a lot. This case demonstrated that at every step during an evaluation, evaluators are faced with choices that affect the precision of their conclusions. Even without actual data to compute the Summative Confidence of the case she reviewed, the provost would be able to gain insight into how she might improve the process. The biggest obstacle toward attaining an accurate summative conclusion is variability. Therefore, one method of reducing variability is to standardize as much of the tenure review process as possible. For example, the university or each department could develop a tenure review policy and enforce the implementation of this policy. The policy should regulate which values would be examined in a review, the structure of these values, the validity of the structure, the organization of values into critically and noncritically important groups, the degree of association between values, the standards that would be used to judge acceptable performance for critically important values, the rubric that would be used to grade performance on values, the number of internal and external faculty that would serve on the panel, the methods and data sources that would be used to measure performance on each value, and the weighting scheme that would be used in data synthesis. Additionally, the provost should recommend that the policy address measurement error. For instance, the policy could require that the panel undergo training in coding qualitative data derived from documents, interviews, observations, etc. Finally, she should recommend that the panel gather as much input (i.e., increase the sample size) as possible, using a systematic and reliable data gathering process for subjective values (e.g., collegiality).

Relevance of Summative Confidence to the Discipline of Evaluation

The purpose of a summative evaluation is to examine the performance of an evaluand on a set of values and to compare this performance with relevant standards to render a summative conclusion. However, without knowing the amount of measurement error that impacted the conclusion, an evaluator cannot gauge the precision of the conclusion nor can a decision maker determine whether actions are warranted to address the issues that produced the conclusion. Furthermore, in situations in which funding allocation or the viability of the evaluand is in question, it is reasonable that

[9] Readers should note that confidence intervals do not have to be symmetrical around the summative conclusion. In fact, the only time a confidence interval is symmetrical is when the summative conclusion is equal to the median of the underlying distribution of conclusions.

decision-makers would need and want to consider the quality of the evaluative conclusions prior to forming a decision. Thus, the position advanced by this paper is that evaluators must begin to report the precision of their conclusions and to the extent possible, take steps during the planning phase to ensure that adequate confidence will be attained for each conclusion.

In fairness to the profession, it is important to mention that a large number of evaluators and evaluation firms take great care in planning and conducting evaluations as well as in neither being overconfident or underconfident in reporting results. However, despite this level of care, no studies have ever been published, to the best knowledge of the author,[10] that report the CI of a summative conclusion despite the fact that the formula that forms the foundation of Summative Confidence has been known since at least 1918.[11] Similarly, the impact of measurement error, sampling error, and inter-rater reliability, to name a few of the relevant factors, on evaluation and research results appears to be ignored. At most, evaluators may include such limitations in their narrative. However, they do not assess the mathematical impact of these errors on either their results or ability to generalize beyond their sample. Finally, researchers and evaluators are often told that they should triangulate their results to improve the validity of their conclusions. However, while it stands to reason that more information is better, a question that has yet to be addressed is "How much data are enough?"

The potential relevance of Summative Confidence to the discipline of evaluation cannot be overstated. In a world in which billions of dollars are spent annually on conducting evaluations, the need to maintain high standards can be overwhelming. To date, poor evaluations can only be unearthed through a metaevaluation—an evaluation of one or more evaluations for the purpose of determining the merit and worth of the original evaluation(s), as opposed to the evaluand(s). However, the cost and time of properly conducting a metaevaluation can be considerable, at times even comparable to the cost and time of the original evaluation. Furthermore, few evaluators have the necessary expertise to conduct such studies. Consequently, the proportion of metaevaluations conducted is incredibly low.[12] Although a Summative Confidence analysis cannot replace a metaevaluation, it can act as a barometer of the quality of the evaluation. Even better, it is considerably more cost-effective than a metaevaluation.

Furthermore, a Summative Confidence analysis has no data restrictions. It can be used with quantitative, qualitative, and mixed method designs. Undoubtedly, many qualitative evaluators will express a level of discomfort at the idea of quantifying qualitative data. Some may even think that quantification can be imposed on some forms of qualitative inquiry only with a machete. While it is not the author's wish to open up past debates that often led nowhere, from the perspective of a statistician, qualitative data may be transformed easily into binary, binomial, or ordinal data.[13] In fact, the

[10] Several key word searches in 44 scholarly databases to which Western Michigan University subscribed—at the time of this writing—did not produce a single article in which an author calculated a confidence interval for a composite variable or evaluative conclusion. Nor did these searches net a single article in which the mathematical algorithm underlying Summative Confidence was discussed or proposed.

[11] Please note that a distinction is being made between the CI of an individual variable (a plethora of studies report such statistics) and the CI of a summative conclusion—a conclusion synthesized from two or more other variables, which typically include qualitative and quantitative data.

[12] According to Michael Scriven, the author of metaevaluation, significantly fewer than 1 percent of evaluations can be classified as metaevaluations (personal communication, February 28, 2007).

[13] Whenever qualitative data can be ranked on a dimension, nonparametric analyses can be performed. Even when qualitative data are not ordinal, they still can

transformation is simply the extension of the definition and classification process that emerged out of qualitative analysis. After all, to define or classify an object, feeling, experience, and so forth requires the imposition of definitional boundaries. For example, the experience of feeling depressed is different than the experience of having a specific phobia. Therefore, by extension, the indicators of depression (e.g., change in mood, ability to derive pleasure, appetite, sleep, irritability, ability to concentrate) differ in a meaningful way from the indicators of a specific phobia (e.g., unreasonable fear by the presence or anticipation of the phobic stimulus, an immediate anxiety response in the presence of the phobic stimulus, recognition by the individual that his or her fear is excessive, avoidance of the phobic stimulus) (American Psychiatric Association, 2000). However, the moment one establishes a definition or classification rubric, one also establishes the process by which qualitative data may be quantified. How many patients in a psychiatric hospital meet the American Psychiatric Association's criteria for depression? What is the proportion of prison inmates who have committed a violent crime? In both of these instances a qualitative rubric must be used to classify individuals into one or more categories. Once categorized, the quantification process requires one to count the number of individuals in a group or to calculate the proportion of individuals who fall within each category. In the author's experience, this process is routine, even among qualitative evaluators.

The only additional step necessary for incorporating qualitative data into a Summative Confidence analysis is to measure the reliability of interpretations and ratings (i.e., the reliability of the coding rubric). If the coding rubric is reliable, then two or more evaluators examining the same qualitative datum should generally code it in the same way. The degree to which they do not interpret the same datum similarly is an indication of an unreliable coding rubric. From the perspective of Summative Confidence, if two or more evaluators examining the same qualitative data reach different interpretations, then the reliability (and by extension, validity) of the summative conclusion is also called into question. In other words, if the coding rubric was unreliable and different evaluators used it to replicate the evaluation, they would likely reach different conclusions. Obviously, this would present a major obstacle for any decision maker, evaluation consumer, or policymaker who needs reliable and valid results to form appropriate decisions.

Delimiters

It is important to note that Summative Confidence is a method for determining the probability that a result will replicate given parallel conditions. Consequently, it is related to validity because replicability is a necessary but not sufficient condition for establishing validity. That is, if identical design conditions are established, then one should expect to observe similar results. The results produced by a Summative Confidence analysis, however, do not imply that the data and/or methodologies used to collect the data were valid or complete. Nor do they imply that the list of values and standards were valid or complete for addressing the purposes of the evaluation. It also does not suggest that the weighting scheme and scoring rubric were appropriate. These are all factors that must be validated independently by the evaluator. Summative Confidence simply calculates the precision of a conclusion within a specific probability given that the same methodologies, data sources, and procedures are used to replicate the evaluation. If the methods used to formulate a conclusion are invalid, then the confidence level surrounding the conclusion is also invalid. Therefore, a necessary precondition to conducting a

be treated as binary or binomial data (e.g., true/false, yes/no, present/absent).

Summative Confidence analysis is the validation of the methods employed by the evaluation.

References

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (DSM-IV-TR) (4th ed.). Washington, DC: American Psychiatric Press, Inc.

Burstein, L., Freeman, H. E., Sirotnik, K. A., Delandshere, G., & Hollis, M. (1985). Data collection: The Achilles heel of evaluation research. *Sociological Methods Research*, *14*(1), 65-80.

Chang, L. (2005, October). *Hypertension: Symptoms of high blood pressure*. Retrieved February 17, 2007, from http://www.webmd.com/content/article/96/103784.htm.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field-settings*. Boston, MA: Houghton Mifflin Company.

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory*. Fort Worth, TX: Holt, Rinehart, & Winston.

Davidson, J. (2005). *Evaluation methodology basics: The nuts and bolts of sound evaluation*. Thousand Oaks, CA: Sage.

Hays, W. L. (1994). *Statistics* (5th ed.). Fort Worth, TX: Harcourt Brace College Publishers.

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation* (8th ed.). Boston, MA: Allyn and Bacon.

Kish, L. (1995). *Survey Sampling*. New York: John Wiley & Sons Inc.

Scriven, M. (1991). *Evaluation theasuraus* (4th ed.). Newbury Park, CA: Sage Publications.

Scriven, M. (2007, February). *Key evaluation checklist (KEC)*. Retrieved April 15, 2007, from http://www.wmich.edu/evalctr/checklists/kec_feb07.pdf.

Wilkinson, L., and American Psychological Association Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.