Performance criteria

Q 7-03. What is the usefulness of performance criteria?

With any complex and continuing problem such as state debt issuance, it is critical to employ performance criteria - set in advance - to guide decision-making and to provide feedback on how well those decisions lead to meeting the debt management goals. The act of selecting performance measures itself can lead to a clearer understanding and statement of the goals of a program and the process by which decisions lead to achievement of those goals.

Q 7-03.01. What should guide selection of performance metrics?

The performance criteria must reflect the management goals and must address the key issues about those goals in a definable way. The criteria should address the questions most asked regarding the management of debt and provide answers to those questions.

To identify possible metrics, begin with the goals that debt issuance is supposed to achieve and any constraints imposed by other policies or organizations. What do these goals mean in terms of outcomes? What specific evidence or results indicate that the objectives have been met? Why do these outcomes reflect the objectives? What is the mechanism that links the outcomes to the objectives? Ask the key questions first, decide what information is needed to answer them, and, only then, select the measures that give that information.

Do not choose metrics simply because the data are readily available. Ease of data collection, although desirable, is not a sufficient answer to the measurement question. The basis for choosing particular measurements is their appropriateness to answering the critical questions.

Q 7-03.02. Why should criteria be set out in advance?

If performance standards are not chosen prior to the activity starting or actions being taken, it is too easy to rationalize a decision after the fact by finding criteria to justify a conclusion already reached. Even if rationalization is eschewed, it will be difficult to avoid a subconscious bias toward measures that confirm the current course. This destroys the usefulness of the criteria and wastes the efforts that went into formulating them. Set out the criteria that are to be used to evaluate actions and policies before beginning to make specific plans.

Q 7-03.03. How many criteria are needed?

There is no basis in theory for the number of measurements to furnish the right amount of information. Each goal may have several performance measures to capture better the meaning of the goal. If it is possible to have multiple goals, it may be necessary to establish the priority among them should they be in conflict. No goal, however, should be without a performance measure.

Q 7-03.04. Should one avoid conflicting measurements?

When several goals are each evaluated by separate performance criteria or when a single goal is evaluated with multiple criteria, it is highly likely that a situation will exist in which a high score along one dimension or criterion can only be obtained by sacrificing a top score along another dimension.

This situation adds nuance and complexity to debt management decision-making, but it need not be taken as a fatal problem. A robust set of criteria is likely to include conflicting results. When this occurs, one cannot do well according to one gauge because one succeeds in another. This outcome is useful in highlighting the tradeoffs among policies and actions.

Despite the challenges presented by the situation, it is generally to be preferred to redefining performance measures so that they agree.

Q 7-03.05. Should all performance criteria be numeric?

That a criterion has a numeric value attached does not mean that it is unambiguous or that it carries more information than any other performance measure. If the criterion does not address the key questions about the program's goals, the fact that it takes a numeric value adds little to its usefulness.

In practice, quantitative judgments must be used with caution because they can give the impression of high precision or exactness that is unwarranted by the underlying information that produces the measurement.[1]

There is also the risk that more information is expected from numeric scores than is actually yielded by them. For example, it is a common practice to transform a Likert scale[2] to a numeric range (numbered 1 through 5, for example to cover the range from "Completely disagree" to "Agree completely") to ease tabulation of

[1] For example, the statement, "the risk of an uncovered auction is 3.7 percent," is likely to result from a confusion of precision with accuracy of information.

[2] A Likert scale is a response format with distinct values for answers. It commonly appears in cases where the respondent chooses among "Agree completely", "Agree"," Neither agree nor disagree", etc.

results. Often, however, this leads to numeric statements that defy accurate interpretation or that provoke misleading conclusions. To the first point, an average score of 3.4 on the five-point scale does not have a meaningful interpretation. This appears to be in our example 40% of the way between "Neither agree nor disagree" and "Agree." Does this mean that the balance of opinion is near the middle? Suppose the actual answers were 40% for "Completely disagree" and 60% for "Completely agree." This result yields the same score, but one is unlikely to suggest that the balance of opinion is near the middle. One cannot take such shortcuts if the responses are not coded into a numeric scale. The results might be forced into a more accurate statement of the situation.

A second problem that arises when numeric scores are used is when the ordinal intent of the numbers – the sense of first, second, third that is intended only to denote relative position – is confused with the cardinality of the numbers – the sense that a "4" is twice as great as a "2." This can lead to absurd or meaningless interpretations of the results, particularly if statistical or numeric analysis is carried out on the data. It may be more difficult to know what a statistical regression on such data may mean than it would be to perform the calculations.

In many cases, the appropriate indicator of a goal or an objective's status cannot be easily measured because of its non-deterministic or qualitative nature. The correct action is not to force fit into the performance measurement system artificial and simplifying numeric measures. Often, such substituted measurements are more complicated and costly to create and evaluate.

Q 7-03.06. What should be done about variability in numeric data?

Repeated measurements taken over a period are likely to vary, even in process in control. Although it is simplest to report single data points, e.g., the average value or the last observed value, the complexity added by showing the degree to which outcomes differ from each other is more than compensated for by the information such data contain. The variability itself is an important measurement in evaluating how well objectives are being met and it should be reported.

The variability can be presented in several ways. The variance (or standard deviation) of the data series is the most obvious measurement. Note that absent any statement about the underlying probability distribution, reporting the variance makes no particular claims about confidence intervals about some value. It is equally informative to report previous values in "buckets" or sub-ranges as in a histogram. Finally, if the volatility of the measurements has been changing over time, one may wish to restrict the period over which all results are reported. This is analogous to the n-period moving average used to measure the mean.

This issue becomes more critical with forecasts based upon past observations. Do not use point estimates alone. Every estimate should be accompanied by a well-reasoned estimate of the likely error about the point. It is not necessary to resort to popular approaches such as constructing a two standard deviation range about either side of the mean value and suggesting that this interval captures 95% of the likely variability unless one can be reasonably sure that the data follow a Normal distribution.

Q 7-03.07. Can qualitative criteria be informative?

Clearly stated qualitative criteria, like numeric criteria, are highly useful if they answer the questions posed. Non-binary qualitative judgments, i.e., those with responses that are more than "yes" or "no," actually are powerful because of their fuzzy nature. Words can convey nuanced meanings that humans understand more clearly than concepts tied to crisp definitions to delineate in meaning.

For example, the response to the criterion "risk of an uncovered auction" being limited to "low, bearable, or high" has an appealing simplicity of concept. It is also cognizant of the absence of a sharp boundary between categories. The transition from "low" to "bearable" or from "bearable" to "high" happens over a range of shades or values before becoming definitely in one class or the other. That process reflects human judgment patterns. Any given value that might be cited as the boundary between two of these categories is likely to be seen as an arbitrary cut-off. It may be safe to say that an individual who was 6 feet 3 inches in height may be considered tall. Trying to specify that only individuals at least 6 feet 3 inches in height are tall will provoke suggestions that persons slightly shorter should still be considered tall. When concepts (or measurements) are subject to some reasonable interpretation over a range, it is better not to use crisp cut-off values.

Q 7-03.08. What are the issues with macroeconomic-based criteria?

Many sovereign borrowers select performance measures that incorporate macroeconomic criteria, such as national income accounts data (particularly GDP), into their calculation, or as ends in themselves.

One significant problem arises when the wrong types of measurements are applied. For example, it is sometimes argued that debt sustainability can be measured by controlling the ratio of debt to GDP. As stated, the argument blurs the distinction between a flow (GDP) and a stock (debt outstanding) and is inappropriate as a measurement. The unstated justification is that the stock of debt will generate an expenditure flow (debt service) that will strain the income flow of GDP. Argued from this perspective, the measurement has some validity, but the criterion must be stated far more carefully actually to compare flows with flows.

If the criteria are carefully stated to address the flow-to-flow comparison, a second issue derives from the use of statistical data such as GDP. A criterion based on statistical estimates such as these requires a broader tolerance on outcomes. It must be remembered that macroeconomic measures are generally derived by sampling or other statistical means that leaves them subject to error and, often, subsequent revision. Measures that are derived from macroeconomic estimates will reflect the inherent variability that comes with sampled data. This will require that targets not be rigidly enforced as they may be subject to change as the statistical measurements are revised.