
Composite Performance Measures – White Paper

Introduction

In the development process for measuring complex outcomes it may be that a single, linear measure is inadequate to the task. For example, when looking at the outcomes of public assistance, a single measure of “cases closed” or “clients served” may be somewhat uninformative or even misrepresentative of the complete process under consideration. Creating a series of individual measures around a single outcome may be unnecessarily complicated and expensive, and still uninformative as a whole.

One potential answer for this problem is *aggregated* measures (where more than one data set is combined to form a more complex representation of a defined complex outcome).

Understanding Composite Measures

Composite measures have two primary advantages; they allow a broader and more complete range of measures of a complex process without resorting to several individual measures, which then have to be correlated ... *they show a big picture at a glance*. Secondly, they allow mixing different types of data together. A composite can be formed of time, quality, and frequency datasets even though each of these individual things is measured differently. An index can also be formed of positive (increase) and negative (decrease) expressions.



Using Composites: Data is not Information – Focusing on Meaning

Everyone is flooded with increasing amounts of data. We are often overwhelmed with it. In the current system it is most common to report “raw” (unprocessed) data in spreadsheet or tabular formats, often with little or no accompanying analysis. Where graphs are used, they often have little if any

Composite Performance Measures

explanation attached to help with interpretation; the results are often largely ignored.

Part of the solution to developing more useful measures of performance is to create measures that are more easily understandable and usable.

Arguably the most important question for a performance measure is, *“What does the data mean?”* This understandable need for meaning in measures is one of the reasons why the use of “dashboards” and other methods for making data more accessible and useful is growing exponentially.

Since public sector performance data should be useful for a wide range of consumers, *how* performance measures are formed and represented becomes very important. Usefulness is often a function of the degree to which complex data is aggregated (or “rolled up”) and displayed. However, since there are generally multiple users, multiple levels of aggregation are logically required (one size does not fit all). For example, the level of data detail required for agency specialists is very different from that required by budget or policy makers.

Unfortunately, this can result in the perceived need for parallel systems, where measures used for operational decision-making exist more or less separately from “reported” data. This is expensive, unnecessarily duplicative and wastes vanishingly thin resources. Instead, imagine a series of “roll-ups” starting with the original, raw operational data, finally culminating in a single, understandable-at-a-glance, composite or index (with points in between driven by different consumer needs for the data). Done properly, however aggregated the *final* expression may be ... it remains possible to de-aggregate the data back to the original raw form, if required or desired.



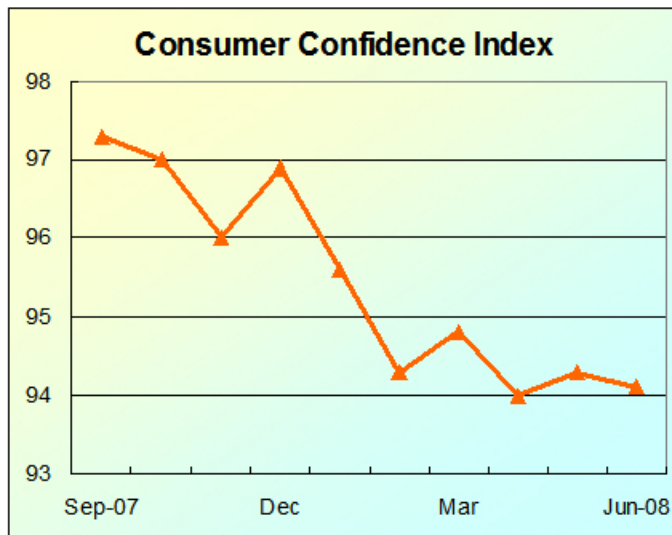
Composite Performance Measures

Primary Performance Outcomes

The initial pass at achieving useful and sustainable performance reporting is for leadership to clearly focus on the critical 20% of enterprise and mission-level outcomes; it is these essential outcomes which must be most carefully and explicitly informed by data. Once these essential outcomes are clearly identified and articulated the next step is to pose two questions:

“How should this outcome be measured?” and ...“What data do we have/need that tells us how well we are doing in reaching our targeted outcome(s)?”

Composite Performance Measures—An Example



People are most accustomed to composite measures in the form of indices such as the Standard and Poors, or Consumer Price Index, in which numerous individual measurements are aggregated into a single number and then tracked over time.

While it is straightforward to develop a single-element measurement for something

like tracking the typical time it takes to issue a permit, many of the outcomes of government are very complex and frequently have significant variables which are partially or wholly outside of the direct control of the individual agency.

Most frequently, even in a complex process there are a small number of *critical* variables which can be identified. For example, in a composite or aggregated Taxpayer Assistance measure for the Department of Revenue, three variables are being tracked:

1. Number of people using automated help versus call-center telephone assistance

Composite Performance Measures

2. Average wait time for call-center telephone inquiries, and ...
3. Customer satisfaction levels with both call center and automated assistance.

The high level outcome being measured is “*efficient and effective taxpayer assistance.*” The strategy embedded in the measure is to use the data to make decisions that serve to shift simple taxpayer inquiries from the call center to the automated system whenever appropriate, because it is faster for the customer and cheaper for the agency than having customers waiting on the phone for a call-center employee. It also increases the standardization of response content for the most common inquiries, because each person accessing the automated system is provided with consistent responses to the same questions.



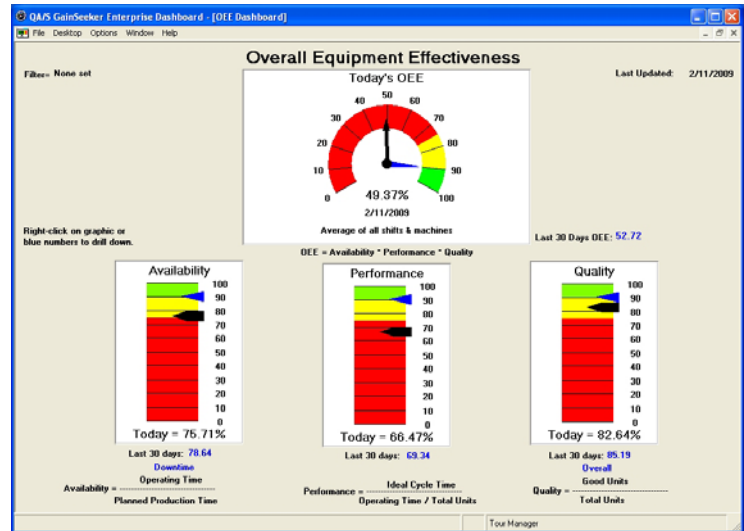
In this specific case, ever shorter call wait-times are not really the desired outcome; rather, the target for this portion of the measure is a targeted *range of wait times*, which (hypothetically) will not be unduly frustrating for people with complex questions (or for those who do not have computer access), but sufficiently long that people with simple inquiries will choose the automated system, rather than wait on hold for a call-center employee.

If the strategy is valid, the measure should show movement from call-center to automated system activity, wait times for the call-center falling within established standards, and reasonably positive customer service ratings. The “check” portion of the measure is to assess customer satisfaction with the process. Another statistical check would be a random sample periodically taken from callers on *why* they are using the phone instead of the automated system.

Composite Performance Measures

Although this is a relatively complex outcome, it can be comprehensively measured using three interdependent variables. Additionally, the three variables can be “weighted” to reflect their relative importance in determining the outcome measure. This measure also has the advantage of being equally useful for management and informational (reporting) purposes; separate measures are not required. It is, for lack of a better description, *elegant*.

The basic idea of a composite performance measure is two-fold: first, to clearly identify the most critical performance variables in a given process and measure them; and, to “compress” or aggregate several variables to form a more complete “picture” of a complex process which can then be reported out more simply.



Aggregating data can produce distortions and obviously reduces the level of initially available detail; there is a trade-off, detail and absolute accuracy for simplicity and usefulness. The case for aggregated measures is most effectively made where absolute accuracy is less a concern than usefulness, and where there are several potential consumers of the information, each of whom with at least somewhat disparate needs.

In such things as medicine and engineering high degrees of accuracy are paramount, but even in these situations reporting aggregates can be useful (such as the relative overall condition of interstate bridges or pavement surfaces within the state, or the overall relationship between variables in a multi-channel blood test). The absolute accuracy of the data is less important than its ability to inform in an adequate and reasonably valid manner.

With properly constructed aggregated performance measures, one never loses the ability to “de-aggregate” and provide more specific (granular) information, unless there is a deliberate decision not to do so.

Composite Performance Measures

The Hypothetical Nature of Performance Measures

Once one moves beyond the obvious (measures of accuracy, timeliness of a simple processes, etc.) there is always a certain degree of uncertainty in any performance measure. For example, measuring the impact of increased numbers of patrol officers on the highway includes hypothetical or partially understood relationships between law enforcement activities and a variety of outcomes. The



assumption (hypothesis) is that increased enforcement will result in a reduction of bad things and an increase in desirable things; therefore, it is worthy of funding.

Traffic patrol is **one** factor in reducing accident and fatality rates on the highway, but it is *only* one. If

this single measurement is to be used as the only way of evaluating the impact of increased patrol presence the result will be incomplete, and perhaps even misleading. An alternative approach might be regression analysis to try to show the *relative* contribution of enforcement activities to overall highway safety. Another would be a more comprehensive measurement model, which might include other outcomes such as increased levels of capture directly related to patrol contacts for various kinds of crime, reduced fatalities as a result of reduced response times to injury accidents, or by adding complexity to the measure (such as including auto construction safety rating improvements, highway safety improvement activities, etc.).

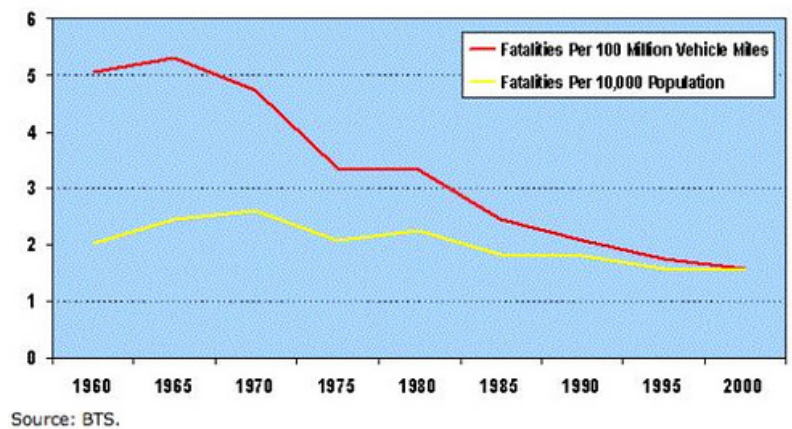
As you can see ... it can get very complicated, very quickly, hence the proliferation of simple, one-element performance measures. Unfortunately, the picture these simple measures paint may be misleading or incomplete and since they are used to make policy and budget decisions, this is problematic. Equally

Composite Performance Measures

often they are a source of frustration on the part of agencies because agencies perceive they are being held “accountable” for things over which they may not have reasonable control.

There are also “confounding” variables. If automobiles are being manufactured to be increasingly safe, the impact patrol activities have on highway safety may be “lost” in the improvements arising from safer automobiles, better roads, etc. Simply counting highway fatalities is an incomplete measurement of the impact of increased patrol presence.

Historical reductions in injuries and fatalities in highway crashes are positively linked to higher rates of seat-belt usage (and safer cars). Reductions in more severe highway crashes are positively linked to reductions in the number of people operating vehicles under the influence of alcohol or drugs. Both



of these variables are *somewhat* impacted by patrol and enforcement activities.

Let us assume that there are two primary desired hypothetical outcomes from increased patrol activities that are more specific and manageable than just highway fatalities (which is really a statistic, not a performance measure); these are:

1. The reduction in “targeted” highway crashes (those involving identified behaviors such as DUI, speeding, etc.), and ...
2. Increased rates of “capture” for targeted illegal activity on the highways (DUI, transporting narcotics, apprehending stolen vehicles, etc.).

Composite Performance Measures

Base-Lining

The initial methodology would be to baseline historical data on these variables, if this is at all possible. Even a single year's data would be very useful. Multiple years of data allows an understanding of historical "trends" which can help inform the measure more accurately, such as, being able to show that while highway fatalities in the state have declined in number over the past three years, the actual number of "targeted" crashes has increased slightly; or that the number of stolen vehicles transported out of state has been steadily increasing over the past four years. In this model, data for targeted enforcement outcomes will be base-lined for the previous three years, and then tracked as the increased enforcement is rolled out.

The Hypothesis as a Foundation for Developing a Measure

The hypothesis contained in a "measure" might be: *Increasing patrol activities by X% over previous levels will result in an increase in capture rates of targeted criminal activities and a "Y" reduction in highway crashes where excessive speed and/or alcohol-drugs were found to be primary contributing causal factors.*

The null hypothesis would be, *"Increasing patrol activities will **not** result in an increase in capture rates of targeted ... etc."* In other words, enforcement is *not* a critical variable in reducing these un-

wanted outcomes. The null is assumed to be true unless proven otherwise; this is why data is required.

We have to set up measures of the activities which will allow us to test our hypotheses. This requires we develop a measurement methodology that reasonably encompasses our hypothesis. This, in turn, becomes the rationale for developing a composite, or index performance measure of several elements, which turn, are tested see what the data shows.



Composite Performance Measures

In order to do this we first need to “normalize” the different datasets ... which in our specific illustration means two things; first to remove “outliers,” and secondly to convert the datasets to *same* unit of expression.

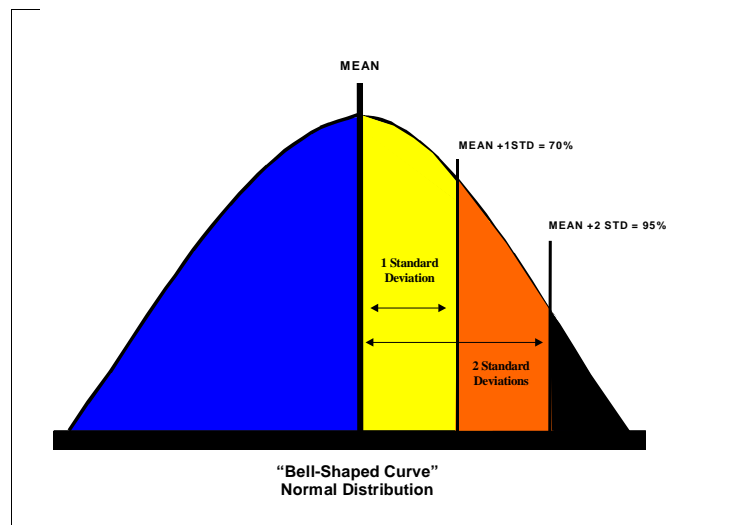
Normalizing Data - Removing Outliers & Converting Data

There are two specific, sequential methods for normalizing data to be used in an index or composite performance measure. First, raw data for each proposed component of the index should be mathematically normalized before being compared to any specific target by excising datapoints outside of mean \pm 2 standard deviations and the “excised” data addressed as “special causes of variation.”

This should be done for most performance measure datasets, single or composite. These statistical “tips and tails” can have a highly distorting impact on mean-based results because they lie outside normal process variation (thus they are not reflective of process performance in the true sense of the word).

Agencies protest that often they are held accountable for things outside of their direct control; removing statistical outliers (special causes of variation) is one way of mitigating this. Identifying these outliers also provides a source of important information about unusual occurrences in the process ... because the question is always, “*Why?*” If the median time for issuing a license is 2 days, and yet there are cases taking more than a year ... *why?*

Identifying the statistical outliers in a given dataset is most easily accomplished by something like the inexpensive Excel add-on QI macros® (which will easily produce a useful distribution histogram directly from an Excel spread-



Composite Performance Measures

sheet), or simply by using Excel to calculate the mean and standard deviation in a specific dataset and then manually inserting a cutoff at the appropriate points (multiply the Standard Deviation value by two and add the product to the mean value to set the cut-off) and recalculate the result.

Any data point that lies outside the mean \pm 2 Standard Deviation range is 'removed,' and the dataset recalculated.

For example, the mean of a hypothetical dataset is 200 and the standard deviation is 20; mean plus/minus two standard deviations would include everything between 160 and 240; anything above 240 or below 160 is removed from the dataset and the mean re-calculated on the remaining data. This will result in a "normalized" distribution, which can then be used for more accurately evaluating variation within the process.

Normalizing Dataset Expressions: Targets within Targets:



Often composite measures are initially suspect, partially because of their perceived complexity, but also because of the concern that important detail will be lost through aggregation.

By converting each dataset to a percentage of target expression the datasets are normalized, but it is still relatively easy to see how each element is performing against its *individual* target. This tends to reassure initial skeptics. Whatever the final level of aggregation, they can always unroll the measure back to the raw data, if they wish.

A common scale of some sort is required for a composite or index measure. For purposes of performance measures, setting targets for individual elements intended for the composite measure as a *percentage of target figure* is perhaps the simplest and most familiar way of achieving the second level of "normalization." This converts individual dataset values to a common scale based on 100.

Composite Performance Measures

Using our earlier illustration of a composite for increasing highway patrol frequency, we will illustrate how this is accomplished:

1. Dataset "A" has a target for at least 50 weekly traffic stops which result in a defined "enforcement" action. Actuals are 55. Percentage of target = +110%
2. Dataset "B" has a target for no more than 25 highway crashes on targeted routes where alcohol and/or drugs are primary contributors. Actuals are 20. Percentage of target = +120%
3. Dataset "C" has a target for 500 enforcement 'contacts' made per week on targeted routes. Actuals are 700. Percentage of target = +140%

Arguably the relationship among the example variables is significant, e.g. more contacts *should* lead to more enforcement, which *should* in turn lead to reduced numbers of impaired drivers by apprehending more of them. Increased enforcement *should* also reduce the frequency of targeted driver behaviors. In effect, a "logic" model is constructed as a result.



Dataset "A" is expressed as numbers of weekly traffic stops resulting in a defined "enforcement" outcome; "B" as the number of highway crashes on targeted routes where excessive speed, alcohol and/or drugs are primary contributors, and "C" is the total number of enforcement 'contacts' made per week on targeted routes.

Obviously, each of these datasets will produce disproportionate counts (the total number of contacts will vastly exceed those resulting in enforcement actions, which will vastly exceed the number of crashes where alcohol and/or drugs are causal factors). Simply combining the raw data and computing the mean will not produce anything useful.

Composite Performance Measures

In building the composite, the three results noted above (converted to percentages of target) are then used as whole numbers, 110, 120 and 140, resulting in an **un-weighted total of 370 against a base target of 100 for each of the three measures, or 300.**

In practical terms both actuals and targets have been “normalized” by converting them into expressions based on some ratio of a base 100.

Since the output from each component is normalized by conversion to a single ratio expression (% of established target) before it can be used for the index, the logic/evidence supporting both the original individual component and aggregate target methodology is critical to the integrity of the index. If the targets are arbitrary, so will be the output of the composite measure based on them. It's like defining 70 as passing for a test with no real validation or rationale. Legitimate targets can be based on statistical or other criteria (such as Federal Requirements). There are two places that indices get pushed on the most, and targets are one of them.

Not all Components are Created Equally – Weighting Elements within a Composite

Weighting is perhaps the most controversial issue in composite measures. The ideal situation is an index of equivalent elements (each presumed to have an equal impact on the ultimate outcome being measured). This is rarely possible. Minimally, there are at least two levels of elements: critical, and contributory. Frequency is also always an issue. An index may contain an element that has a high rate of occurrence but low criticality, or a very low rate of occurrence, but high criticality.



In the case of the example measure, targeted-cause highway crashes would be considered critical in terms of impact, and the criticality factor is increased within the index because the “n” (number of events) is low. Adding weighting

Composite Performance Measures

to the measure differentiates among the variables and identifies those components which are seen as contributing more to the ultimate outcome being measured (either by criticality or frequency).

But presumed impact is only one factor in weighting. Some elements of a measure may be more prone to confounding variables (longer time period, more impacted by things outside of the program confines, etc.). When this is the case, it might be argued that element is *less* critical to the actual measure, because it is intrinsically more random and variable in nature; therefore less of a “performance” issue. Always remember, the emphasis in this application is on measuring *performance*.

For most performance composites, elaborate weighting schemas are not justified. Much simpler is to develop a logic model with one or two weighting steps (x2, x3) and then test the composite with real data to see how it behaves.

Example; assume there are four components in an index, two of which are weighted at par (no multiplier); one is X2, and one X3 of par. The normalized scale of each is 0-100 (simply converting percentage of target to a whole number). The resulting index has a potential value of: 100, 100, 200, and 300 for a total of 700. Because the elements are weighted, element four makes up nearly half the entire potential value of the index.



Of course, one can use smaller increments of weighting. The most important point is that the weighting must make sense either statistically, or be the result of some formal process, such as an Subject Matter Expert (SME) panel. In an SME panel three to five subject matter experts independently set the weighting for each element with an open scale between 1 and 3X and use the median value as the multiplier. What *can't* be true is that the weighting is simply “picked out of the air.”

One of the reasons why this type of performance measure has fallen into disrepute is “cooking” the index by weighting elements in a way that is intended

Composite Performance Measures

to produce a predetermined outcome. This is called 'gaming' and when identified calls into question the credibility of all of the measures from a given source, not just the tainted index. Agencies are best advised to use a formal process, and have their approach to weighting the elements validated externally to avoid this issue.

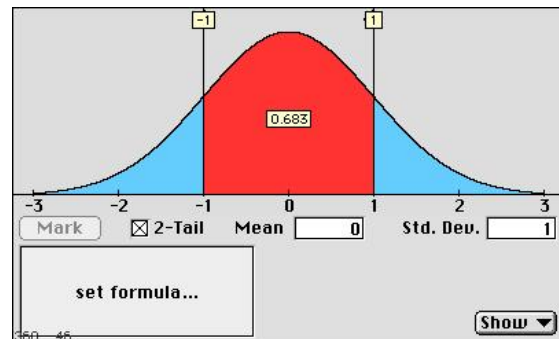
Normal Ranges – Intervals

While composites/indices are generally *expressed* as a single number, they are all about "movement." Very much as with the Consumer Confidence Index, it is the *movement* of an index over time that becomes ultimately meaningful. One of the most profound ways in which this is true, is to be able to plot movement of the index against known significant external events (such as economic downturns, or resource increases/decreases, etc.). Properly designed, this makes it possible to see the broad impact of various events and decisions on the performance of a given process or program.

This means it is necessary to define what constitutes a "significant" variation for the index to make sense. There will always be fluxuations ... normal process variation ... but it is important to understand when the fluxuations are meaningful.

For the level of precision required for most performance measures one can initially apply a customary thumbnail range of variation, such as 5% (people are very used to seeing 95% confidence intervals or 5% error rates, even if they don't understand them) and assert that fluxuations in the index that fall within

the 5% tolerance are assumed to be process "noise" and are thus not significant. Over time a statistician can calculate the real value for you, or if you are very conversant with Excel, make sure the statistics add-on package is installed, and use the following URL, <http://home.ubalt.edu/ntsbarsh/excel/excel.htm>, for help in performing the required calculations.



Composite Performance Measures

Shared Measures

One of the primary challenges in performance measurement is when aspects of an specific outcome are distributed through more than one department within an organization, or among multiple organizations. In government settings this can encompass multiple branches of local, state and federal government and various contractors or vendors of services.

In this application, either each entity has the same measure (which may not be practical, or each “partner” has a somewhat different measure, which is aggregated to measure the ultimate outcome of the process.

For example, permanency of placement for foster children cuts across federal, state and local operations. Piecemeal measures of the activities undertaken to achieve safe, timely, and effective placements may be relatively uninformative of the whole.

There are significant potential benefits to developing shared performance metrics for the broad outcome, which subsume specific elements housed in various organizations, not the least of which is the communication, cooperation and coordination this kind of a process can engender.

Conclusion

The primary function of performance measures is just that ... *measuring performance in some understandable and methodologically valid way*. In most cases, absolute accuracy is less important than usability. Frankly, most consumers of the data (including the agency itself) need answers to a few basic questions:

1. Do I understand it?
2. Do I believe it?
3. Is it getting better, staying the same, getting worse? And why?
4. Where is the optimal level of investment-resourcing? Where do we get the best return for the least amount of money?

Composite Performance Measures

5. Is it being done in an effective and efficient manner?
6. Are the stakeholders reasonably happy?

Simple, linear measures are sometimes uninformative and misleading. More complex measures, properly designed and constructed, can provide a richer source of information to a wider user group.

Rick Gardner
Performance Management Coordinator
Department of Administration / Budget & Management
503-378-3117