DO's and DON'Ts of Generating Performance Metrics

Michael C. Frick, Warren, MI.

ABSTRACT

The author draws on 10 years of working experience with both executive leadership and operational groups to provide a practical list of the do's and don'ts for generating performance metrics that will drive the desired improvements. A sample of topics covered are (1) the need to separate process improvement metrics (owned by planning) from process execution metrics (owned by operations), (2) the need for alignment of metrics with organizational boundaries and operational job responsibilities, (3) the balancing act between too many and too few metrics on a scorecard, (4) dangers of going to an automated electronic scorecard system, and (5) why medians are a "don't."

INTRODUCTION

It seems that much of the metrics literature is aimed at identifying which set of metrics should be captured and how those metrics are mapped back to key corporate objectives. This work assumes a metric has been identified and focuses on providing a practical set of "do's" and "don'ts" for establishing a measurement process that will help drive process improvement. Its goal is to help others avoid some of my painful learning experiences. As such, the paper is really more of a documentary of real-world events that conspire to break a perfectly laid out measurement process and how one might avoid them. (Yes, people do cheat on their metrics if given a chance!) The paper is organized around a discussion of the nine major pitfalls I have run across over the years, each followed by a table of recommended do's and don'ts.

PITFALL #1: WHO OWNS PROCESS CAPABILITY IMPROVEMENT?

Assume our metric is "average throughput" for a given process, and it has been determined that it needs to be cut by 20% for the company to remain competitive. Who should own the process improvement project? Since the metric is clearly a process metric, one could easily assign responsibility directly to the operations group with a target reduction of 20%. Who better than the folks who live in the trenches? However, unless the operations group is severely underperforming, this can (and has) been a recipe for disaster.

In this type of situation it is imperative that the following steps be performed.

- Construct an estimate of current process capability (expected time if business process is followed.)
    o Compute current performance.
    o Interview the operations group to gain an understanding of how long the process should take, including the day of the week and time each step in the process should occur.
    o Based on information developed in the interviews, reverse engineer the process capability with data. For example, does the weekly process really start on Tuesday mornings?
    o Work with operations to finalize an estimate of current process capability.
- Assign the current process capability to the operations group as their performance target and assign remediation improvement tasks as required if current performance is below standard.
- Assign responsibility for identifying initiatives (including funding and staffing) to improve the current process capability to the planning staff that supports the operations group.
- Assign the task of implementing specific improvement tasks to operations as required.

These steps may seem trivial; but it is imperative that one separates performance against existing capabilities from performance against desired capabilities. The operations group already has its hands full running the current process. Assigning the process improvement target directly to operations with no real support typically leads to either "artificial" improvement based on "creative" book-keeping or simple lack of participation. Assigning improvement responsibility to the planning group assures that leadership is apprised of necessary resources required to achieve the corporate objective and what actions have been put in place to move the needle at the start of the planning cycle. To this day, I am still perplexed by the reluctance of some organizations to establish specific initiatives with measurable outputs within the planning documents for an organization.

DO'S

- Establish a meaningful Current Process Capability
- Hold Operations responsible for performing to current capability
- Hold Operations responsible for implementing specific improvement initiatives
- Hold Planning responsible for identifying improvement opportunities, acquiring the resources necessary to implement them, and the impact a successful implementation will ultimately have on the performance metric.

DONT'S

- Assign process improvement directly to operations without support

PITFALL #2: AGGREGATION OF METRICS ACROSS DIFFERENT BUSINESS MODELS

Many metrics are aimed at measuring the average response time of a process, and it is typical for executive leadership to ask for a single, aggregated metric which captures performance of this metric across all product lines and regions. But what if multiple business models are in use in the company?

Assume that our metric is order response time and that fast-turning product is delivered from a local warehouse while slower moving product is ordered directly from the factory. Assume ½ the product is delivered from the warehouse and takes 2 days and the other ½ is factory ordered and takes 30 days. So, on average our response time is 16 days. But what does 16 days mean? How would one use it? It means nothing to the customer as no 16 day process exists. It cannot be used to contrast regional performance as the business model mix within each region can have a bigger impact on the number than actual performance. Similarly, it cannot be used to track performance over time as the percent delivered from the warehouse will most certainly change over time.

So why report it? The most common response I have been given when questioning leadership is the desire to limit the number of metrics on the scorecard. And yet, I have had to participate in many inter-regional deep-dives because one regional leader does not understand why another region's performance can be "so much better" due to his lack of understanding of how business models were deployed in different regions.

DO'S

- Measure at the appropriate business model/process level where a physical interpretation exists.
- If forced to aggregate, continue to report the disaggregated numbers and their usage rates even if it must occur on a lower regional scorecard.

DONT'S

- Aggregate metrics across business models without a fight.

PITFALL #3: ASSIGNING METRICS THAT SPAN DEPARTMENTAL CONTROL TO INDIVIDUALS AND/OR DEPARTMENTS

As part of the process to evaluate departments and people, there is often a desire to measure the individual's performance against key corporate initiatives. For example, a key corporate initiative might be to reduce inventory expenditures by reducing the mix of warehouse to factory orders from 50/50 to 25/75. Placing this metric on a department's scorecard, such as logistics, who do not contact customers regarding order practices, would be of little value as they cannot be held accountable. One must define departmental metrics that directly measure department output that has an impact on the larger corporate objective. Here, since delivery reliability is a key driver of whether a customer will order from the factory, an appropriate logistics metric for this initiative would be aimed at improving logistics reliability. Note, per our previous discussion under Pitfall #1, logistics would only be responsible for hitting an improved reliability level. The planning organization would be responsible for its ultimate impact on the warehouse to factory order mix.

DO'S

- Individual and department metrics must be tied to department output.
- If necessary, place the higher corporate parent initiative on individual and department scorecards as a memo only item for reference.

DONT'S

- Assign high-level metrics to lower organizations.

PITFALL #4: POTENTIAL DANGERS OF AUTOMATED SCORECARDS

My concerns with automated scorecard systems are two-fold: (1) their lack of functionality and (2) their potential to negatively impact the metrics review process.

From a functionality stand-point, the tool we currently use plus others we have reviewed seem to lack the following desired capabilities:

1. Built-in volume weighted average for quarterly and year-to-date aggregates.
2. Built-in volume weighted glide path through the end of year.
3. Built-in year over year comparison.
4. Roll-up capabilities across scorecards (again volume weighting issues.)
5. Acceptable formatting/printing – extra work needed to dump to Excel.
6. Automated metric entry.

Not only is the lack of functionality annoying, it pushes otherwise rationale business planners into making poor design choices when setting up the metrics template for the tool (formulae, formatting, etc.). Most of the tools I have seen provide manual work-arounds for some of these missing capabilities; but, I have seen quarterly and year-to-date aggregates reported without volume weighting because it was either too much extra work or the planner did not even realize that it was necessary. I have also had to live for years with a flawed glide-path process that does not seem to be able to focus on comparing an expected year to date value against the actual year to date value to determine if one is on track. In our current process, we manually type in our volume weighted year-to-date values and our year-end forecasts. However, since the tool does not allow me to enter a cumulative year-to-date target, it is not really possible for the executive staff to evaluate the presented year-end forecast.

Perhaps worse, my personal observation is that the use of an automated tool accustoms the executive leadership to focus exclusively on point estimates, R-Y-G flags, and one-line variance explanations. Implementing an automated tool can push an organization away from an intense focus on a few key metrics (trend charts, Pareto charts, etc.) towards a superficial focus on a large number of metrics. And it is particularly easy to provide a less than "honest" variance explanation absent any back-ground materials.

DO'S

- Avoid the use of automated scorecards if possible. The temptation to race through the metric review will be too great.
- If your tool allows for links to back-ground materials, insist that it is used.

DONT'S

- Allow the metrics template within an automated system to be set up without a technical review.

PITFALL #5: TOO MANY/TOO FEW METRICS?

Many factors contribute to whether an organization has set up too many or too few metrics. But I am sure of one thing after having sat through and presented at many, many metric reviews; a given body can only concentrate on a limited number of metrics in one sitting (perhaps a dozen.) After that, people tune-out and the rush-to-the-end begins. I believe the primary driver for having too few metrics is the inappropriate aggregation of metrics across business models/processes that was discussed above. I believe the primary driver for having too many metrics is leadership's desire to micro-manage through scorecards.

Consider the measurement process for a single corporate metric like average order response time, which we will assume is a function of the individual response times of the sales, production control, manufacturing, and logistics departments. If leadership asks to see response time of each, we have grown from 1 to 5 metrics. But the company has 5 regions across the globe each with their own set of metrics yielding a new total of 30. Also, it employs two distinct business models, sell from stock and build to order so the total increases to 60. The number quickly becomes overwhelming when you consider we have not yet considered the need to see the data by origin-destination combinations

My question is simple: "Why does the central executive staff need to see all of this detail?" Their concern should be on whether each region's overall performance is to standard, sufficiency of that standard, and progress made on improvement initiatives. Details, which would still be readily available on a regional lead's scorecard, should only be presented to explain performance issues. If one simply aligns reporting detail with area of responsibility, I firmly believe the issue of too many metrics solves itself.

Providing details on the central scorecard also can have a significant downside. Due to cultural differences, geographical differences, and sophistication of operations in different countries; each region is likely to have

a different organization structure which implies a different internal measurement process. For example, a large manufacturing company may have facilities in the U.S., Europe, and Thailand. For the larger operations, one may have a central production control group which performs both the scheduling and sequencing activities while one or both of these activities may be distributed directly to the manufacturing facilities in the smaller operations. As it is critical that metrics be aligned along organizational boundaries, each region will define their internal metrics based on their organization. However, if the global team insists on reporting details on the central scorecard, they will also insist on a common calculation. This generally means that the smaller regions have to report metrics per the organizational structure of the home region for the global team. The smaller regions will end up keeping two sets of books.

DO'S

- Utilize the organizational structure and report sub-metrics at the appropriate level
- Monitor process improvement initiatives based on the metrics established for the implementation of the initiatives, not the sub-metrics that may be directly impacted by the initiative.

DONT'S

- Micro-manage via scorecards.
- Aggregate metrics across business models/processes without a fight

PITFALL #6: GLOBAL

Language, cultural, and time-zone differences will cause issues when attempting to establish a global metric reporting process. Direct oral communication is a must to insure both comprehension and agreement with those whose first languages are different. Although written communications are imperative for pre-read materials and post-meeting documentation, there is no way for the author to get a read on whether the group has interpreted the materials correctly without direct verbal communication. Even so, one still needs to be very careful in global conference calls. Silence on the other end of the phone can mean:

- It is 11:00 p.m. local time and my child is crying so I didn't hear.
- I totally disagree so I will just do it my own way
- I do not understand but do not want to ask a question.

A weak positive response could also indicate a lack of understanding. Even a strong positive response does not truly mean understanding. After hosting a year's worth global conference calls to initiate a global performance measurement process, I am convinced that many folks around the world speak English better than they understand it - leaving the facilitator with a false sense of the other's comprehension level.

DO'S

- Be aggressive probing for understanding in conference calls.
- Be aggressive following up off-line with the silent majority.
- Write clear, simple communications free from slang, acronyms, and idioms.
- Push for at least one face-face meeting. It does wonders for later phone communications.

DONT'S

- Accept silence as agreement or understanding.

PITFALL #7: USING ONLY ONE TEMPORAL VIEW FOR LONG LEAD-TIMES

As manufacturing companies go global, supply chains lengthen and order response times for both parts and finished goods go up. If one is calculating an order response time metric which is the accumulation of several independent, sequential activities, it may be appropriate to calculate the segment metrics using two different time views.

The first view must be tied to the end of the last process, for example – delivery to customer. To calculate the metrics, one would collect all orders that were delivered to the customer in a given reporting period and then calculate the end-end and sub-process metrics on this set of orders. Assuming you have a good understanding of process capability in each area, this view allows you to assess which sub-processes may have negatively impacted the end-end performance.

Unfortunately, root causing the issue with this type of report for long lead-time order responses is difficult. Operations folks tend to live in the now and have short memories concerning what might have impacted them 2-3 months ago. Further, if the issue occurred in one of the early sub-processes, you may now have an entire pipeline of product impacted by the issue.

The fix is to calculate each sub-process metric a second time focusing only on the orders that exited that sub-process in the previous reporting period. This allows you to see and fix issues as they occur as well as get them documented. One might be tempted to only report this second calculation; however, since each sub-process calculation is based on a different set of orders they will not sum to the total which opens up a potential opportunity for reporting errors. Reporting errors are discussed in more detail in a later section.

DO'S

- Calculate sub-process metrics based on multiple time views for long end-end response times.
- Assign to operational groups a metric which is based on when the orders exited their sub-process.

DONT'S

- Calculate only the sub-process metric based on when the orders exited the sub-process.

PITFALL #8: MEDIANS

If the metric is stand-alone and does not need to be aggregated or summed, then feel free to use the properties of the underlying distribution to decide which statistic is appropriate. But even then, I am not so sure I would ever use medians again as the need to aggregate or sum may change down the road.

In all other cases I strongly recommend that you run away from medians as fast as you can. Despite their obvious advantages if the distribution is skewed, their disadvantages in real world metrics are much greater as I learned after having to live with a median based metric system for several years.

TRUE STORY NUMBER 1:

Our executive director was out ill and his boss had to present the metrics to the senior manufacturing leadership. Despite the fact that he had been involved in the original mean versus median discussions and regularly participated in metrics reviews, I was called to his office to explain the difference between means and medians. After 5 minutes, I was dismissed under a hail of expletives and the review was postponed.

TRUE STORY NUMBER 2:

One of our key performance metrics was initiated by a weekly process that always started on a Tuesday. As the final process rarely occurred on a Sunday, throughput times (and hence a median throughput) of 5, 12, 19, 26, 33, or 40 days etc. were all but impossible. Unfortunately, our target was set to one of these multiples and we spent the year jumping between 1 day over and 2 days under target (Saturday end dates were more common than Sunday but still unlikely). Try explaining that to the executive described above.

TRUE STORY NUMBER 3:

For two years, I had to report the weighted average of the mean for one division with the median of the other divisions because I did not have access to their divisional data because their leadership wanted to participate and our leadership recognized their inclusion would help us hit our numbers. At the time, I was not that removed from my research days and fought this vigorously. What does the average of a mean and median represent? Upon later reflection, I am not so sure it was the wrong thing to do. Performance on a critical initiative was improved at all divisions. Off course, if we had all been reporting means and volumes, there would not have been a mathematical issue with the inclusion of their data.

TRUE STORY NUMBER 4:

Despite the fact that we used medians for several years and reports were clearly labeled with "WARNING – MEDIANS ARE NOT ADDITIVE", my staff at the time had to continually respond to manager and executive assertions that the numbers were wrong because they did not add.

DO'S

- Use medians when the distribution is skewed and the metric is stand-alone.
- Bend theory if the practical solution drives the correct behavior???

DONT'S

- Use medians if sub-processes are involved or cross-regional metrics must be aggregated.

PITFALL #9: DISTRIBUTED METRICS REPORTING
OR WHY "GREEN" + "GREEN" +"GREEN" CAN EQUAL "RED"

Many metrics, such as "average customer response time", cross organizational boundaries. For example, overall response time might be a function of the individual response times of the sales, production control, manufacturing, and logistics departments. In such cases, it is imperative that each department be assigned a process metric and associated target which, if achieved, drives success in the end-end target.

This sounds and should be simple. But unless all of the departments fall under the same executive leadership, each will have a great deal of latitude in how their scorecard metrics are defined and calculated. Plus, each department will push to calculate its own metrics. And for a global company, it may not even be feasible to build an all encompassing data warehouse.

Hence, for a variety of reasons, you may find yourself in the situation where metrics reporting is distributed across departmental and geographic boundaries. If there is one golden rule of metrics that I have learned in 10 years it is this "If metrics are defined and calculated independently, the sum of the parts will portray much better performance than the whole performance". I have personally witnessed each of the following:

1. The global definition for a metric was defined as the "average" response time. Three out of 4 reporting areas reported the arithmetic mean. One decided to report medians (giving them a distinct performance advantage due to the shape of the distribution). Six months later when the difference was uncovered, they produced a statistical reference that listed Median as one way to compute an average and claimed they were following the global definition.
2. A global metric was defined as the total elapsed time to complete a process which was made up of three sub-processes.
   a. One business unit reported the sum of the three sub-processes rather than actually calculating the total process. As we were using means, this should have been fine. However, there was about a 2 week buffer between the first and second process which neither process claimed as their responsibility – so neither side reported it.
   b. Another business unit responsible for reporting one of the sub-metrics decided that orders placed on hold for any reason were no longer under their control and simply decided to drop these orders from the metrics calculations.
   c. A third business unit decided to report the minimum of the mean and median (response time for their sub-process was short and less skewed than the end-end distribution.)
   d. Another business unit decided that it would only calculate metrics for orders where all key dates in the order's life cycle were known and consistent (e.g. build date after the release to manufacturing date). In their report, they listed volumes along with the segment and end-end response times. Unfortunately, they listed total number of orders found, not total number of orders used in the calculations. Turns out they had a reporting problem inside the database for one key event. Although they reported their calculations as if they had nearly 100% of the orders, they actually were basing their calculations on less than 5%.
3. Another global metric was defined as percent delivered to customer on time.
   a. As "customer" could be either a wholesale customer who intends to resell the product from his shelf or a retail customer who direct orders, the intention of this metrics was to measure reliability based on the promise date from the factory. Despite very explicit language in the global metric definition to this effect, one business unit decided that 100% of the product sold off the shelf should be considered delivered on-time as the retail customer took it off the shelf.
   b. Another business unit had the ability to satisfy a retail order from a locally stored warehouse; however, they did not have the ability to measure when the order was delivered to the customer so they sampled from a distribution to obtain a pseudo "actual" delivery date to the customer. Interestingly, they predicted the "estimated" arrival date for the order by sampling from the same distribution. Needless to say, their delivery reliability for orders delivered from the warehouse was quite good.
4. Perhaps the most creative game I witnessed was the business unit who applied a series of filters (which they called data cleaning routines) to the data as they pulled it from their main-frame systems into their metrics data mart. They used these data cleaning routines to drop any order whose response time had been affected by outside influences over which they had no control (e.g. shipments delayed due to strikes). Since the offending data never made it into the data mart, they could be very open with their data query and calculation methods used inside the data mart without anyone in the global team being aware of the amount of legitimate data being dropped.

The bottom line is that motivated people will find a way to achieve their targets – by hook or crook. The Dos and Don'ts listed below are my best advice as to how to combat the "crook".

6

DO'S

- Assign short term targets that are achievable based on current process capability and defined and staffed improvement initiatives. If you assign someone an unachievable target, they will either ignore it or creatively reach it.
- Assign the job of metric's cop to someone with a technical background. If the local business planner does not know he should be taking weighted averages, he will surely not be able to keep up with a creative analyst.
- Write explicit metrics definitions. Try to think like a lawyer and minimize the opportunity for creative interpretation.
- Use scenario testing. Identify each type of business model and distribution process that might be employed. Ask specifically how each case is handled – both how the data is extracted and exactly how the metric is calculated. Ask about specific types of process interruptions that might occur – like dock strikes. Over time, we found it was much harder for people to mislead you when they were asked direct questions.
- When calculating response time for a metric that is the result of several sequential processes,
    - Report arithmetic means – you cannot police any other statistic
    - Report both total volume and volume used in the calculations
    - Calculate total process time and sub-process times independently, but on the same set of orders. Compare the sum of the sub-process times to the total time.

DONT'S

- Assume honest reporting
- Enable the slippery slope by allowing data to be dropped from the calculations for any reason other than bad or missing fields. Better to be "red" with an explanation than have all the individual segments be "green" and the overall process be "red".
- Allow individual departments to define their own metrics
- Allow individual departments to calculate their own metrics and then expect the sum to equal the overall total.

CONCLUSION

As I said in the Introduction, my purpose in pulling together this documentary on my metrics experiences was to help others avoid some of the real-world pitfalls which can derail a performance measurement process. If I sounded cynical at times, I can assure you I did not start out that way.

Altogether, I have listed close to 40 do's and don'ts in the text. Below are my top 5.

- If you need to report sub-process metrics or aggregate across regional boundaries, use means as your statistic regardless of the shape of the distribution. It is the only central tendency statistic that can be double checked for errors; hence it is the only one that can be policed.

- Always report volumes, both total record count and quantity used in the calculations.

- Avoid the slippery slope – do not drop data for any reason unless the record clearly has bad or missing fields.

- Use scenario testing at least once a year to limit deviation from the global definitions over time.

- Assign targets that are based on a realistic view of current process capability and clearly defined process improvement initiatives. These initiatives must have measurable outputs that can be translated to a specific level of expected improvement in the performance metric.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Mike Frick
Enterprise: Retired, October 1, 2009
Address: 30238 Underwood Drive
City, State ZIP: Warren, MI 48092
Work Phone: N/A
Fax: N/A
E-mail: mcf_for_mwsug@yahoo.com
Web: N/A