Using Measurement System Analysis for the Soft Stuff

We believe that any of the LEAN, Six Sigma or Theory of Constraint principles are all just tools in one big toolbox. As far as we're concerned, any of these methodologies can be mixed and matched to help a team improve their processes. Too many companies decide to implement one **or** the other and end up trying to force fit tools from their chosen methodology to any and all problems they encounter.

It may sound a bit bizarre, trying to use measurement system analysis (also known as Gage R&Rs) for performance appraisals, but this presentation will show you the results that prove, if applied correctly, the results will be just as meaningful as on any piece of equipment tested in your labs.

It all began for me with a project I was asked to facilitate on the performance appraisal system at an aerospace company. It's a fact that I have yet to find any employee that enjoys performance appraisal time – whether it's the giver or the receiver! As you can imagine, we spent a great deal of time on putting together a charter for this project.

> We started with trying to form a business case. With all of the projects the company had underway already and those still in the queue to be done, why did we want to do this project? How could we phrase it that would convince the steering committee to fund this project over all others? Well, we went mining for data, and found the following: "For the past five years, in every one of the 19 employee surveys sent out to all employees, performance appraisals have ranked absolutely last in satisfaction. Five years ago there were 53 employees who refused to sign their performance appraisals, and 22 of those appended comments directly related to the unfairness of the rating process. This past year that number of refusals to sign has jumped to 187, with those who commented on the appraisal process at 156."

Needless to say, our business case was accepted and our project was funded. Next came our goals and objectives – what did we want to achieve? This was really tough, because fixing the performance appraisal process can quickly become trying to solve world hunger. We needed SMART goals: specific; measurable; achievable; realistic and time bound. Our team eventually decided on the following:

> Improve employee satisfaction with the performance appraisal system from the $<10^{th}$ percentile to at least the 45^{th} percentile. As you can imagine, there were team members who thought 40 was such a stretch goal we'd never get there, and those who thought we should go for 95%. 45 was the percentile where we finally achieved consensus. We also want to ensure there was a well thought out process for all managers to use with clear, concise directions on the who, when, how, how often, what, where and why.

So, we put together a team of different levels in the organization, made sure every function was represented, and set to work. In our very first session we mapped the multitude of ways performance appraisals were being done. It probably won't surprise you to find out that every single manager/supervisor did things slightly different. For those of you familiar with current state mapping, just imaging a map from floor to ceiling, around three walls of a very large conference room. That was the map from the time notice was received from Human Resources that it was time for appraisals until Human Resources had received a signed (or not) performance appraisal for every employee. Needless to say, the discussion was great, the wailing and gnashing of teeth even greater and the language not fit for prime time. To say the team went in the ditch – the part where the storming happens - is a gross understatement! I'm pretty sure we had hurricanes and tornadoes in the same room!

I will mention that we used the BMW chart more in that single event than in any event before or after. A BMW chart is only pulled into play when the team has got to the point of endless loops of whining. You can see where that might happen with a team focused on performance appraisals. So, after giving the team some time to vent – they do need to get some of those gripes off their chests - at the appropriate moment the facilitator posts the "Bellyaches, Moans and Whines" chart. Then, much like a parking lot, you simply ask whoever has the next gripe to post it there. It's a way to get out of the whining loop with a little understanding and humor.
It wasn't long after the current state mapping that the team decided to look at the performance appraisal tool being used for the company. One of the interesting exercises we used here was to take paragraphs from the appraisal and, having formed small teams, ask each person on a team to interpret what it meant. Of course, we quickly learned that much of the language was open to cultural differences, educational differences and maybe even mood changes! So the team asked if there was a tool that would test these appraisals in real life to see if any two managers/supervisors interpreted them the same way. After much thought, since the facilitator was in Master Black Belt training and had heard of a similar test happening at corporate, a Measurement System Analysis was proposed. After some training on the tool, the team concurred and the test was set up.

Human Resources provided us with five completed "generic" performance appraisals – meaning all identifying data had been removed – and we managed to find five managers who "volunteered" to do the appraising. We recruited managers who had a large number of employees in their departments. Our premise was that these people were more apt to be ranking employees that were unknown to them personally.

The current appraisal was set up to rank every employee on the following scale:

1 = Outstanding
2 = Excellent
3 = Fully Successful
4 = Acceptable
5 = Unsatisfactory

against the categories of:

Job Knowledge
Performance
Work Practices
Leadership
Communication
and, if appropriate, Supervision

There were some words in each category, under each ranking to help explain. For instance, in the Job Knowledge category, under 2 (Excellent) was the statement, "Has state-of-the-art knowledge in some areas of the discipline and is competent in multiple disciplines". Or, in the category of Supervision, a ranking of 1 (Outstanding) said, "Always find the right balance between empowerment and appropriate boundaries for subordinates. Aggressively and effectively breaks down barriers to performance improvement and counterproductive behaviors……." These then lead to an Overall Rating with the same rankings of 1 to 5, where 1 is Outstanding and 5 is Unsatisfactory. Although all categories were reviewed, this final rating became the basis for the Gage R&R.

The experiment was set up as follows:

- Each of the five managers was given one appraisal with the instructions to note the date and time started and the date and time completed with scoring.
- Every appraisal came with some background material on the "employee" being rated, and certain of the team members were assigned the task of developing in-depth information about each in case one or more managers had questions the background material did not cover.
-

In an attempt to normalize the data, other team members checked on the managers to determine the following:

1) Just returning from vacation?
2) Getting ready to start vacation?
3) Easy day

4) Busy day
5) Bad day
6) What went exceptionally well/bad for them the day of the appraisal

As soon as the first appraisal was completed and returned, the next one was given, and so on. Once all five appraisals were completed, two weeks were allowed to elapse, the cycle started over. The appraisals were given to the managers in random order in every scoring. After the second set was complete, three weeks elapsed, and the final round began.

As you can tell, this took several weeks to complete. Some managers returned the appraisals like clockwork, and some had to be reminded many times. Sounds just like the real think, huh? Actually, the entire experiments took just under 10 weeks, but the team believed the results were extremely useful.
The following is the ANOVA completed using a software package called SPCXL:

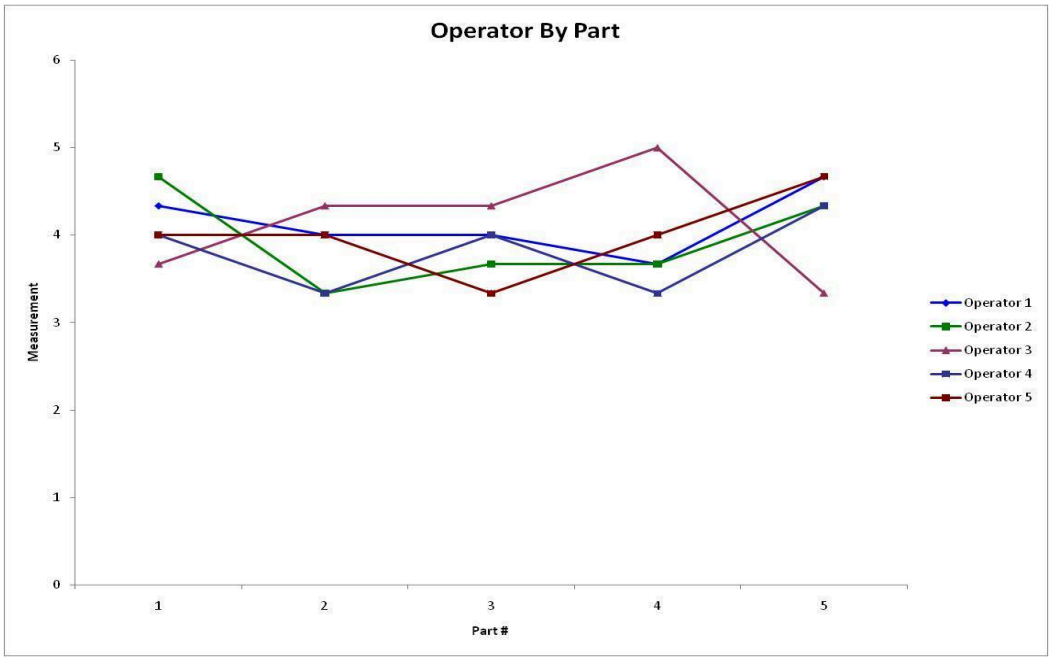
USL	5
LSL	1
Precision to Tolerance Ratio	1.14836841
Precision to Total Ratio	1
Resolution	0.0

MSA ANOVA Method Results

Source	Variance	Standard Deviation	% Contribution
Total Measurement (Gage)	0.58611111	0.765578939	100.00%
Repeatability	0.46666667	0.683130051	79.62%
Reproducibility	0.11944444	0.345607356	20.38%
Operator	0	0	0.00%
Oper * Part Interaction	0.11944444	0.345607356	20.38%
Product (Part-to-Part)	0	0	0.00%
Total	0.58611111	0.765578939	100.00%

For those of you who may have forgotten, if you have a Precision to Tolerance Ratio of $\leq$.10 your measurement system is considered adequate. Greater than .30 means your measurement system is inadequate. The same measures apply to Precision to Total. The rule of thumb is that if your Precision to Total is $\leq$.10, your measurement system is adequate. Another really interesting thing about this data is that Repeatability (the ability of one operator to measure the same instrument multiple times and, on average, get the same measurement) contributed 79.62% of the problem.

This is a look at how each manager did on the five "employees":



You get the idea, I'm sure! The measurement system was so broken you could almost say it didn't exist. Obviously we were beginning to understand why employees didn't like the system.

Our next objective was to drill down through the data to find out why. Why would a manager rate an employee a 5 once and a 3 the next time. Why would an employee average 4 from three managers, a 3 from one and a 5 from the last? Every manager was given exactly the same information about each employee. We knew real-life bias such as "my best buddy" couldn't be happening; these people only existed in our minds. This is where the background material of the kind of day, vacation or no, worst/best thing happened became invaluable.
A sub-team tackled all of that data and found the following trends:

1. Higher employee ratings were given by managers who had at least one day off prior to doing the evaluation
2. Managers who were having very bad/busy days tended to rate more employees in the middle (3)
3. If something really good had happened that day, ratings were higher
4. If something really bad had happened that day, ratings were very low
5. Early mornings produced the highest ratings
6. Late afternoons produced the lowest ratings
7. Managers who took the longest to complete each round were the most consistent in their repeatability

As a result of this experiment, the original project team broke into two smaller teams. The first team was focused on precise language. Their work included benchmarking performance appraisals around the world to develop a world class rating system and the phrasing those systems used. Much of their time was spent hammering out the exact language to use in each ranking of each category to ensure specificity. There were many surveys completed – mostly in-person interviews during lunches to find a common language that everyone in the company understood. Some team members did "man in the street" interviews, where they would stop employees in the hallways and spend five minutes clarifying a phrase or sentence currently being worked on. A tremendous amount of time was spent getting as many employees involved as possible.

The second team focused on the training that each manager/supervisor would undergo prior to the next appraisal time. Part of the training involved showing the experiment outlined above to those managers, along with an in-class demonstration using two fictitious employee appraisals. Another part of the training was based on a list of common adjectives used in performance appraisals and a concise, specific meaning for each. A "dictionary of common terms" was developed and provided to each manager/supervisor.

After the new appraisal tool was developed, and all of the training was completed, it was time for the annual performance appraisal round. Two months later (November) the next employee survey went out. To say the team was anxious to see the results would be an understatement. I'm sure the outside group compiling the results had never seen so much interest in one subject on a survey.

As a result of the work of this team, the satisfaction of the performance appraisal system went from the <10th percentile to the 31 – 40th percentile! A little more probing by the Human Resources group shows the actual percentage to be just over 38%. Not only did the satisfaction greatly increase, but the number of refusals to sign had dropped from 187 to 9! Not one person of the 9 refusals cited anything to do with the appraisal process! Project success!!

There were many parts of the project that have not been discussed here – getting the waste out of the appraisal system, improving the method of delivery, more timely feedback rather than waiting until the end of the year, etc. However, this team feels certain that the MSA was the greatest motivator in changing the process and increasing Customer satisfaction.

While all of this was happening at our location, the Master Black Belts from Corporate had completed an MSA on performance appraisals with almost identical results. They also showed tremendous amounts of variability in the measurement system, and Repeatability was the culprit for their study also. As of today, we have now completed a total of five Performance Appraisals Gage R&Rs, and the results show a marked trend concerning the Repeatability vs Reproducibility. Repeatability numbers seem to always be higher.

As a result of these two studies, we decided to rewrite a portion of our Green Belt and Black Belt training to incorporate the "quick and dirty" MSA used in the performance appraisal training. The results of the appraisal Gage R&R in every class since then have continued to prove the theory that as managers and supervisors we need to be much more cognizant of the impacts our day, our timing and our language skills have on performance appraisals.

It is through non-traditional uses of tools such as the Measurement System Analysis that we can continue to improve the "above the factory" processes that cost us the most. Employee satisfaction may not have a place on the profit and loss statements, but we all know how easily it can affect that bottom line.

Two more examples of using tools in non-traditional ways include:

- A hospital House of Quality to improve the reputation and customer satisfaction with the emergency department. This was a county hospital that had a reputation that would send patients driving an extra 45 – 50 minutes to the next closest hospital regardless of how poorly they felt, or how serious their condition. As a result of the work the team did using the QFD this emergency won an award two years in a row for most improved emergency room in the state.

- A screening Design of Experiments in an aerospace company to determine the most significant factors affecting the inability of the Production Department to meet the goals of cost, quality and schedule of the program. It turned out at the interaction between Production and the Engineering Department was a major variation contributor. As a result of the work the team accomplished to control the major factors, schedule was improved by 28%, costs dropped by 22%, and quality improved by 43%.

We are constantly looking through our box of tools to find new ways to solve some of the old problems. Whether the tool is a LEAN principle, a Six Sigma tool, or something from the Theory of Constraints, we're looking at them all with new eyes to stretch the old boundaries.

Sandy Miller
Executive Partner & President
***X*-Stream Leadership Group LLC®**
www.x-slg.com
484-941-3943
A **S**ervice-**D**isabled **V**eteran-**O**wned **S**mall **B**usiness