A Better Way to Do R&R Studies

The Evaluating the Measurement Process Approach

Donald J. Wheeler

Last month's column looked at how to fix some of the Problems with Gauge R&R Studies. This month I will show you how to learn more from your gauge R&R data with less effort. Rather than getting lost in a series of computations, the Evaluating the Measurement Process (EMP) approach uses the power of the graph to reveal the interesting aspects of your data so that you can know how to ask the important questions.

AN EMP STUDY

The idea behind an EMP study is both simple and profound. As expressed by my friend and colleague, the late Richard Lyday, "Measurement is a process, and with rational subgrouping you can study any process." An EMP study begins very much like a gauge R&R study, but rather than computing estimates of everything possible, it immediately places the data on an average and range chart in order to discover what is happening in the data.

When we use an average and range chart with experimental data we are doing something completely different from what we usually do with this chart. When an average and range chart is used with data from a continuing process it is properly called a process behavior chart. There the objective is to classify the process as either being predictable or unpredictable. In contrast to this, in an EMP study we are looking at the results of a special type of experiment. Here we trying to determine if we can detect part-to-part differences in spite of the uncertainty introduced by measurement error. This shift in both the nature of the data and the nature of our questions will change the way we interpret the average and range chart of an EMP study.

While the EMP approach can be adapted to many different data structures and data collection schemes, we will illustrate the basic EMP study using the same data collection strategy used in a gauge R&R study. A simple fully crossed experiment is performed where two or more operators measure each of three to ten parts two or three times apiece. For our example we shall use an EMP study where six operators measured each of four parts three times apiece.

The measurement system consists of a manual test stand that measures an electromagnetic property of particular product. Since this manual test stand is used in production for 100 percent inspection, it is crucial to the operation of the plant. Since six operators routinely perform this test, all six were included in the EMP study. The four parts used in the study were selected from the product stream on each of four different days.

With simple and objective measurement systems the EMP study may be performed in a fairly straightforward manner. Richard Lyday would usually collect his data in two or three rounds where each operator would measure each part once in each round. However, with subjective or complex measurements it may be necessary to "blind" the experiment where the operators do not know when they are retesting a given item, and where the order of testing is shuffled or "randomized" in some manner.

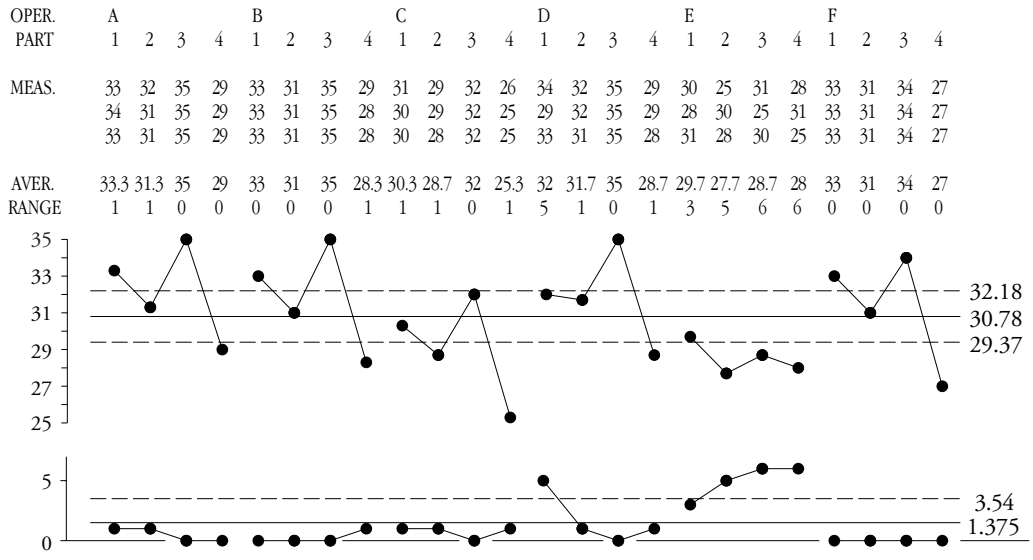

Figure 1: EMP Study for the Manual Test Stand

The key to understanding any average and range chart is to understand what sources of variation are found *within* the subgroups and what sources of variation are found *between* the subgroups. In Figure 1 there are three distinct sources of variation: The operator-to-operator and part-to-part differences which show up between the subgroups and the measurement-to-measurement differences which show up within the subgroups.

The test-retest variation found within the subgroups is commonly known as the repeatability. This isolation of the test-retest error within the subgroups, with all of the other sources of variation showing up between the subgroups, is the distinguishing characteristic of an EMP study. Because of this isolation of test-retest error, the limits shown on the average and range chart in Figure 1 depend solely on the test-retest error. Therefore, the limits in Figure 1 specifically show that amount of variation which can be attributed to measurement error alone.

As always, the average chart looks for differences between the subgroups while the range chart checks for consistency within the subgroups. This characteristic of the charts means that the range chart in Figure 1 checks these 24 subgroups to see if there is any inconsistency in the amount of test-retest error shown. The range values that fall above the upper range limit are signals of inconsistency in the test-retest error. Since such inconsistencies represent serious problems with the measurement procedure itself, the causes of these points deserve investigation.

Because the limits for an average and range chart are robust, we can, in spite of the inconsistency on the range chart, also use the average chart to evaluate the part-to-part and operator-to-operator differences. We begin by discussing the part to part variation.

The differences between the parts will depend upon how the parts were selected. Sometimes the parts may be selected at specific intervals from the product stream. At other times the parts may be a simple grab sample, or some other type of haphazard sample, selected from the product stream. And in some cases the parts may be deliberately selected to represent a range of product values. Regardless of how the parts are selected, you will want to detect the part-to-part differences in spite of the uncertainties introduced by measurement error. This means that you will want to find points outside the limits on the average chart. As long as you did not select the parts in such a way that they all end up being alike (such as might happen if you use only

rejected parts in the study) you will expect to find points outside the limits. The average chart allows us to make a visual comparison between the part-to-part variation and measurement error. The part-to-part variation is represented by the width of the band swept out by the running record. The measurement error is represented by the width of the average chart limits. Thus, the wider the band swept out by the running record is relative to the width of the limits, the easier it will be to detect product variation in spite of measurement error.

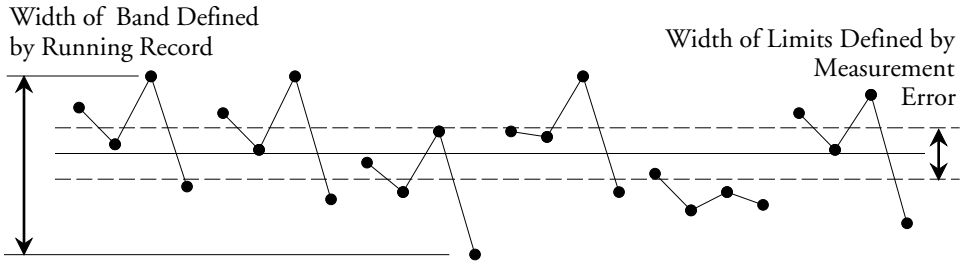

Figure 2: Relative Utility of Measurement System Shown on Average Chart

At the same time that we want to detect the part to part differences, we prefer for there to be no differences between the operators. There are two ways to check for operator differences using the average chart. The first of these uses the *shapes* of the running records and the second uses the *positions* of the running records. In order to facilitate both of these comparisons an EMP study will omit the line segments that would connect dots from one operator to the next.

To see how to interpret the shape of the running record it is helpful to begin by considering what the average chart would look like if there were no differences between operators and also no measurement error. Under these conditions the running records for each operator would be exactly the same. Segment by segment they would be perfectly parallel to each other (rather like the curves for Operators A and B). However, as soon as we introduce measurement error into the picture we will begin to see slight departures from perfect parallelism (rather like the curves for Operators D and F). As long as there is a reasonable degree of parallelism we need not be concerned. Here Operators A, B, C, D, and F all show a reasonable degree of parallelism. Operator E, on the other hand, displays a serious lack of parallelism.

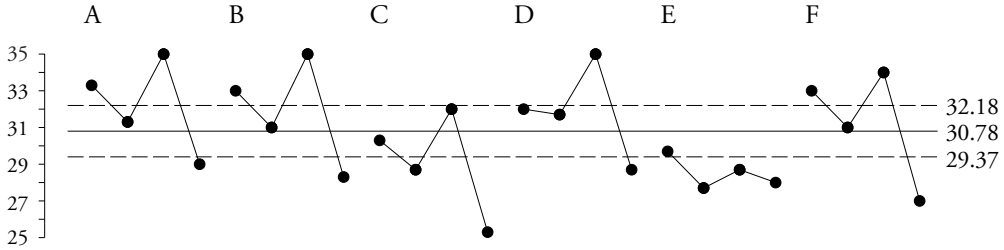

Figure 3: Lack of Parallelism for Operator E

So what does a lack of parallelism represent? Serious nonparallelism is indicative of an *interaction effect* between the operators and the parts. (Algebraically, interaction effects and nonparallelism are one and the same thing: You cannot have an interaction effect without a lack of parallelism, and vice versa.) Here we see that Operator E is measuring these four parts in a

substantially different manner. Since there should be no interaction effects between the operators and the parts, this interaction represents a serious inconsistency in the measurement process that needs immediate attention. Such interaction effects might be due to operators using different techniques, or to some operators skipping a step in the measurement procedure, or simply due to the presence of one or more untrained operators. But whatever the cause, it is a problem with the measurement process that needs to be fixed.

In addition to checking for parallelism, we can also compare the positions of the running records. When we do this we are essentially comparing the operator averages. In Figure 4 we see that both Operator C and Operator E have averages that are substantially lower than those of the other four operators. Such differences between operator averages are potential operator biases.

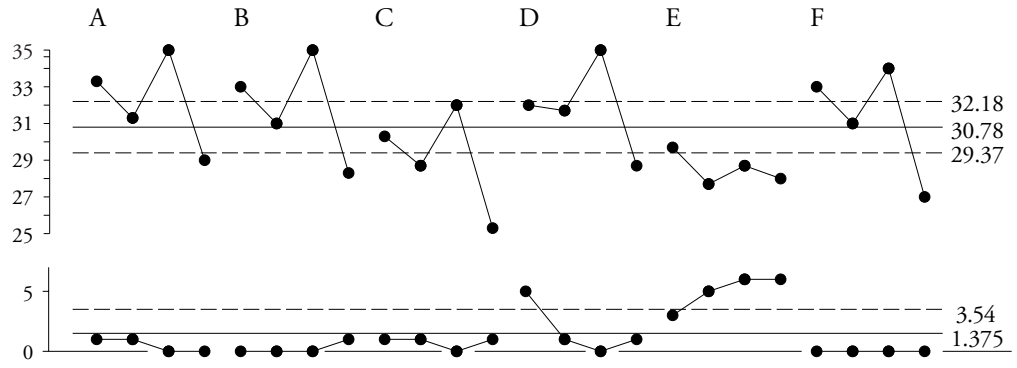

Figure 4: Potential Operator Biases for Operators C and E

So, what can we say is the overall message of the EMP chart in Figure 4? Operators A, B, and F show good parallelism, have similar averages for these four parts, and show consistently small amounts of test-retest error. Comparing the width of the limits with these three running records we can see that the manual test stand can detect product variation.

Operator C shows good parallelism, and a small amount of test-retest error, but he is consistently low on all of his measurements. This is a potential operator bias. The reason for this bias needs to identified so that the bias can be eliminated.

Operator D has reasonable parallelism and a good average, but she has a range point above the upper limit of the range chart. Obviously one of her readings for Part 1 is problematic. While the other ranges and the reasonable parallelism show that she usually does a good job, the reason for this aberrant reading needs to be identified.

Operator E has large ranges, poor parallelism, and the wrong average for these four parts. Whatever else you might say about him, he clearly does not know how to use the manual test stand. While Operators C and D may need a refresher on using the test stand, Operator E needs to be moved to another job until he can be trained in how to use this device and can display a skill level comparable to that displayed by the other operators.

Of course, the first step is getting the operators to measure things alike is to convince them that they are not currently doing so. It is likely that Operators C, D, and E all think that they are measuring these parts in the same way as Operators A, B, and F. Creating Figure 1 is the first step in convincing them that they are not.

SO WHAT HAVE WE LEARNED?

An EMP study begins by placing the data from a gauge R&R study on an average and range chart. By doing this we can make several qualitative assessments even before we begin to make any specific computations:

1. The range chart will allow us to determine if the test-retest error is consistent throughout the study, and also to judge if it is consistent from operator to operator. When test-retest error is not consistent we will need to find out why.
2. The average chart will allow us to assess the relative utility of the measurement system by showing whether the measurement process can detect product variation.
3. The average chart will allow us to spot nonparallelism between the operators. Since any appreciable nonparallelism will indicate an interaction effect between the operators and the parts, it will warn of serious inconsistencies in the measurement process.
4. The average chart will allow us to assess the likelihood of detectable operator biases. If such biases exist they will need to be eliminated to get the most out of the measurement process.

By the time you have constructed your EMP chart you will know what is going on in your data. You will know the interesting questions, and you will know if problems exist. One of the fundamentals of data analysis is to always begin with a graph of your data. Computations exist to complement graphs, but they can never replace them. When you depend upon the computed quantities alone, you are likely to miss many of the interesting aspects of your data.

The objective of analysis is insight, and the best analysis is the simplest analysis that provides the needed insight. Moreover, it does no good to discover something when you cannot communicate your discovery to others. EMP studies use the power of the graph to help with both the discovery and the communication.

BUT HOW CAN WE BE SURE ABOUT THE DIFFERENCES?

While the graph in Figure 4 is fairly clear, not all EMP studies are so clear cut. If we think we see an operator bias, or if we think the operators have different amounts of test-retest error, how can we be sure that we are not merely interpreting noise? To answer these questions we will need to rearrange the data for further analysis. The following charts will provide a powerful way to answer all of the questions of interest arising out of the EMP study.

MAIN EFFECT CHARTS

To compare the operator averages we use an Analysis of Main Effects (ANOME). This is a generalization of an average chart that is appropriate for experimental studies (which will have a fixed amount of data). The limits will be computed using the grand average and the average range from Figure 1. In an ANOME the original k subgroups of size n are rearranged into m subgroups, and the idea is to see if any of these m subgroup averages are detectably different from the grand average. The limits for the Main Effect Chart will be:

$$\textit{Grand Average} \pm \textit{ANOME}_{.05}\,\textit{(Average Range)}$$

where $\textit{ANOME}_{.05}$ is the 5% critical value for an Analysis of Main Effects. This critical value will depend upon n, k, and m, and may be found in Table 1.

For the data from Figure 1, the six operator averages are 32.17, 31.83, 29.08, 31.83, 28.50, and 31.50. The grand average is 30.78 and the average range is 1.375. With $n = 3$, $k = 24$, and $m = 6$, our scaling factor from Table 1 is $ANOME_{.05} = 0.415$, resulting in limits of:

$$Grand\ Average \pm ANOME_{.05}\,(Average\ Range) = 30.778 \pm 0.415\,(1.375) = 30.21 \text{ to } 31.35$$

The Main Effect Chart in Figure 5 shows Operators A, B, and D to have averages that are detectably greater than the grand average, while Operators C and E have averages that are detectably lower than the grand average. Since the grand average is a somewhat arbitrary point of comparison, we can look at the width of the limits and conclude that Operators A, B, D, and F have reasonably similar averages, while Operators C and E have averages that are substantially different from the other operators.

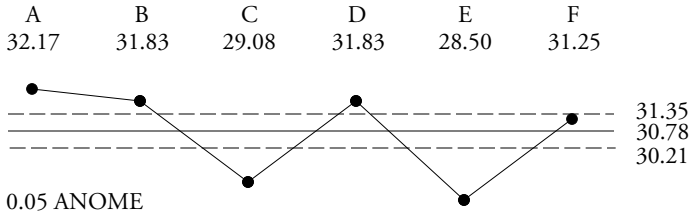

Figure 5: Main Effect Chart for Manual Test Stand Data

Hence, Figure 5 shows that the apparent Operator Bias seen in Figure 4 is real. When you are presenting the results of an EMP study, the use of a Main Effect Chart will make any operator biases easier to see. Moreover, it will make it much harder to dismiss real operator biases as a fluke.

MEAN RANGE CHARTS

To see if the operators display different levels of test-retest error we can use an Analysis of Mean Ranges (ANOMR). Beginning with the original k subgroups of size n, we use the k subgroup ranges to compute an average range for each operator. These m average ranges will be compared in a Mean Range Chart. As before, the limits for an Analysis of Mean Ranges will be based on the original average range for the k subgroups of size n. Since these limits can be nonsymmetric, we will need two scaling factors. The limits for a Mean Range Chart will be:

$$Upper\ ANOMR\ Limit = UMR_{.05}\,(Original\ Average\ Range)$$
$$Lower\ ANOMR\ Limit = LMR_{.05}\,(Original\ Average\ Range)$$

where $UMR_{.05}$ and $LMR_{.05}$ are the 5% critical values for an Analysis of Mean Ranges. These critical values will depend upon n, k, and m = number of mean ranges to be compared, and may be found in Table 2.

For the data from Figure 1, the $m = 6$ operator average ranges are 0.50, 0.25, 0.75, 1.75, 5.00, and 0.00. The original average range is 1.375. With $n = 3$, $k = 24$, and $m = 6$, our scaling factors from Table 2 are $UMR_{.05} = 1.679$, and $LMR_{.05} = 0.438$, resulting in limits of:

$$Upper\ ANOMR\ Limit = 1.679\ (1.375) = 2.31$$
$$Lower\ ANOMR\ Limit = 0.438\ (1.375) = 0.60$$

Table 1: Scaling Factors for Main Effect Charts

$ANOME_{.05}$		$n = 2$	$n = 3$	$n = 4$	$n = 5$
$k = 4$	$m = 2$	0.833	0.384	0.261	0.202
$k = 6$	$m = 2$	0.610	0.299	0.206	0.162
	$m = 3$	1.084	0.519	0.356	0.276
$k = 8$	$m = 2$	0.501	0.253	0.176	0.139
	$m = 4$	1.157	0.568	0.392	0.305
$k = 9$	$m = 3$	0.814	0.408	0.283	0.221
$k = 10$	$m = 2$	0.435	0.224	0.156	0.123
	$m = 5$	1.202	0.599	0.414	0.324
$k = 12$	$m = 2$	0.389	0.203	0.142	0.111
	$m = 3$	0.678	0.346	0.242	0.190
	$m = 4$	0.884	0.448	0.313	0.245
	$m = 6$	1.233	0.622	0.432	0.338
$k = 14$	$m = 2$	0.357	0.186	0.129	0.091
	$m = 7$	1.258	0.639	0.444	0.345
$k = 15$	$m = 3$	0.592	0.306	0.215	0.160
	$m = 5$	0.928	0.477	0.333	0.254
$k = 16$	$m = 2$	0.331	0.172	0.114	0.083
	$m = 4$	0.741	0.383	0.264	0.205
	$m = 8$	1.272	0.650	0.452	0.354
$k = 18$	$m = 2$	0.309	0.163	0.101	0.071
	$m = 3$	0.531	0.278	0.186	0.140
	$m = 6$	0.959	0.495	0.341	0.263
	$m = 9$	1.288	0.663	0.459	0.358
$k = 20$	$m = 2$	0.292	0.140	0.094	0.066
	$m = 4$	0.650	0.332	0.232	0.174
	$m = 5$	0.782	0.401	0.280	0.214
	$m = 10$	1.304	0.667	0.466	0.365
$k = 21$	$m = 3$	0.485	0.245	0.167	0.125
	$m = 7$	0.980	0.504	0.349	0.272
$k = 24$	$m = 2$	0.264	0.123	0.076	0.055
	$m = 3$	0.451	0.226	0.151	0.115
	$m = 4$	0.585	0.300	0.202	0.155
	$m = 6$	0.811	0.415	0.287	0.223
	$m = 8$	1.000	0.516	0.357	0.278
	$m = 12$	1.327	0.680	0.475	0.374

Table 2: Scaling Factors for Mean Range Charts

ANOMR.05		$n = 2$		$n = 3$		$n = 4$		$n = 5$	
		LMR	***UMR***	***LMR***	***UMR***	***LMR***	***UMR***	***LMR***	***UMR***
$k = 4$	$m = 2$	0.271	1.729	0.481	1.519	0.578	1.422	0.633	1.367
$k = 6$	$m = 2$	0.395	1.605	0.575	1.425	0.656	1.344	0.701	1.299
	$m = 3$	0.136	2.133	0.333	1.775	0.445	1.620	0.512	1.539
$k = 8$	$m = 2$	0.475	1.525	0.635	1.365	0.703	1.297	0.741	1.259
	$m = 4$	0.109	2.317	0.292	1.881	0.405	1.703	0.474	1.606
$k = 9$	$m = 3$	0.246	1.915	0.442	1.625	0.539	1.502	0.596	1.436
$k = 10$	$m = 2$	0.530	1.470	0.672	1.328	0.733	1.267	0.770	1.230
	$m = 5$	0.092	2.432	0.268	1.952	0.381	1.759	0.451	1.655
$k = 12$	$m = 2$	0.569	1.431	0.703	1.297	0.758	1.242	0.789	1.211
	$m = 3$	0.329	1.784	0.511	1.534	0.596	1.434	0.647	1.374
	$m = 4$	0.210	2.052	0.404	1.705	0.504	1.567	0.563	1.490
	$m = 6$	0.082	2.520	0.253	1.998	0.363	1.798	0.433	1.691
$k = 14$	$m = 2$	0.603	1.397	0.724	1.276	0.777	1.223	0.822	1.178
	$m = 7$	0.074	2.591	0.239	2.043	0.350	1.829	0.424	1.712
$k = 15$	$m = 3$	0.388	1.701	0.559	1.476	0.637	1.387	0.696	1.320
	$m = 5$	0.189	2.142	0.378	1.762	0.479	1.613	0.549	1.513
$k = 16$	$m = 2$	0.630	1.370	0.743	1.257	0.799	1.201	0.836	1.164
	$m = 4$	0.288	1.898	0.476	1.605	0.570	1.481	0.626	1.410
	$m = 8$	0.068	2.65	0.228	2.080	0.339	1.851	0.413	1.730
$k = 18$	$m = 2$	0.649	1.351	0.757	1.243	0.819	1.181	0.856	1.144
	$m = 3$	0.436	1.637	0.599	1.436	0.680	1.339	0.728	1.283
	$m = 6$	0.171	2.213	0.361	1.805	0.468	1.634	0.534	1.536
	$m = 9$	0.063	2.70	0.220	2.107	0.331	1.874	0.406	1.744
$k = 20$	$m = 2$	0.668	1.332	0.781	1.219	0.832	1.168	0.866	1.134
	$m = 4$	0.347	1.797	0.528	1.528	0.617	1.424	0.669	1.351
	$m = 5$	0.265	1.976	0.452	1.644	0.550	1.510	0.608	1.432
	$m = 10$	0.059	2.742	0.213	2.128	0.323	1.890	0.399	1.762
$k = 21$	$m = 3$	0.478	1.585	0.638	1.389	0.707	1.308	0.753	1.253
	$m = 7$	0.159	2.261	0.348	1.833	0.456	1.659	0.522	1.553
$k = 24$	$m = 2$	0.696	1.304	0.803	1.197	0.857	1.143	0.886	1.114
	$m = 3$	0.505	1.547	0.658	1.360	0.731	1.277	0.772	1.234
	$m = 4$	0.398	1.723	0.573	1.478	0.656	1.373	0.701	1.318
	$m = 6$	0.248	2.028	0.438	1.679	0.537	1.530	0.595	1.451
	$m = 8$	0.150	2.309	0.338	1.857	0.447	1.674	0.512	1.570
	$m = 12$	0.053	2.803	0.203	2.158	0.312	1.913	0.386	1.782

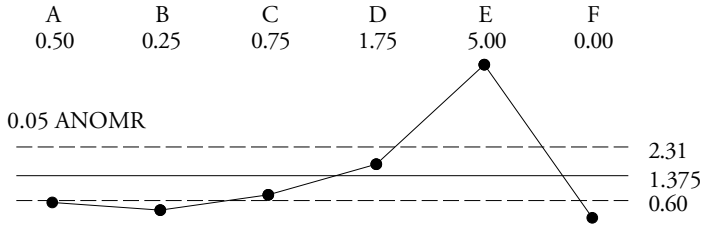

Figure 6: Mean Range Chart for Manual Test Stand Data

Figure 6 shows the Mean Range Chart. With an overall risk of a false alarm of five percent, we can say that Operators A, B, and F have average ranges that are detectably smaller than the original average range, while Operator E has a an average range that is detectably greater than the original average range. Once again, based on the width of the limits, it is reasonable to say that Operators A, B, C, and F show similar amounts of test-retest error, and that Operator E is in a class of his own. While we found both Operators D and E to have points outside the limits in Figure 1, the Mean Range Chart will make the case even clearer.

PROBABLE ERROR

Using the test-retest error shown by Operators A, B, C, and F we find a revised value of the overall average range to be 0.375. This gives us an estimate of repeatability of:

$$\textit{Repeatability} = \frac{\bar{R}}{d_2} = \frac{0.375}{1.693} = 0.22 \text{ units}$$

Since this average range is based on 16 subgroups of size 3 it is said to have 29 degrees of freedom. This means that when this manual test stand is operated consistently the measurements will have a probable error of:

$$\textit{Probable Error} = 0.675\ (\textit{Repeatability}) = 0.15 \text{ units}$$

An appropriate measurement increment will be no larger than twice the Probable Error, and no smaller than one-fifth of the Probable Error. Here we find that twice the Probable Error is 0.30 units. Inspection of Figure 1 will show that they have been recording the values to the nearest whole number. This means that in rounding off their values they have been needlessly discarding useful information. (This also explains why eleven of the subgroups in Figure 1 had a zero range.) Instead of rounding everything off to the nearest whole number, they need to record these values to the nearest tenth of a unit.

INTRACLASS CORRELATION

While the Repeatability and the Probable Error describe the quality of the measurements in an absolute sense, there remains the question of whether or not these measurements can be used to detect product variation. To answer this question we will need an estimate of the product variation. Due to the small number of parts used, EMP studies (and gauge R&R studies) are not the best place to obtain such estimates. In general, estimates of the product variation should be obtained from a process behavior chart. However, failing this, we can still obtain some rough idea about the relative utility of the measurement system for a given application from the EMP study. Here we delete the values for Operator E and use the remaining values to find part

averages of 32.33, 30.73, 34.2, and 27.67. The range for these four part averages is 6.53. The d_2^* bias correction factor for one group of four values is 2.237. Dividing and squaring we find:

$$\textit{Product Variance} = \frac{6.53^2}{2.237^2} = 8.521$$

This estimate of the variance will only have 2.9 degrees of freedom. Nevertheless, we can still estimate the Intraclass Correlation to be:

$$\textit{Intraclass Correlation} = \frac{\textit{Product Variance}}{\textit{Repeatability Squared + Product Variance}} = \frac{8.521}{0.049 + 8.521} = 0.994$$

Given the relative sizes of the two numbers involved (0.049 and 8.5) our uncertainty in the product variance is not going to have an appreciable impact upon the Intraclass Correlation statistic. So while this number may be soft due to the small number of degrees of freedom, we can still see that this measurement system will provide a First Class Monitor for measuring this product.

This example comes from my book *EMP III: Evaluating the Measurement Process and Using Imperfect Data* where further examples and explanations may be found. Tables 1 and 2 were excerpted from my book *Range-Based Analysis of Means.*

SUMMARY

An EMP study uses the power of the graph to reveal the interesting aspects of our R&R study. Here we:

1. identified serious problems with one operator;
2. found two more operators that need some retraining; and
3. identified three operators that are getting the most out of the measurement device.

In addition, with a couple of basic computations, we discovered that we need to record one more digit in these data, and established that this measurement system is a First Class Monitor for use with this product.

While it has been said that "the average and range chart technique will not allow you to estimate the interaction effects," this statement conveys a false impression. When there is an operator by part interaction present, estimation is moot. The real question is "Who is different?" The EMP approach shows any and all interaction effects that may be present and tells you who is different. Without the insight to the operator differences provided by the average and range chart, and confirmed by the ANOME and ANOMR analyses, we might not have removed Operator E's values. This would have skewed both our computations and our interpretation of the data.

The use of an average and range chart to evaluate the measurement process has been around for quite some time. It was briefly described and illustrated in the Western Electric Statistical Quality Control Handbook in 1956. Primarily due to the efforts of Richard Lyday, it was also included in the AIAG Measurement Systems Analysis handbook. However, in both instances, very little guidance was provided on how to interpret the charts and make sense of the analysis. Since this use of the charts with experimental data is substantially different from the traditional use of the charts with observational data, this lack of guidance resulted in a lack of use.

Hopefully, this article has shown how the EMP approach can be used to make the qualitative assessments that are needed to make sense of many R&R studies, and how ANOME and ANOMR can be used to confirm the nuisance components of measurement error as a first step in getting rid of those nuisances.