What Are Good Measurements?

Consistency trumps precision

Donald J. Wheeler

Who could ever be against having good measurements? Good measurements are like apple pie and motherhood. Since we all want good measurements, it sounds reasonable when people are told to check out the quality of their measurement system before doing an experiment or putting their data on a process behavior chart. In this article I shall consider what properties your measurement system does, and does not need to be useful in experimental or observational studies.

THE STRUCTURE OF AN OBSERVATION

Any observation can be visualized as having two parts: one component will represent the value of the thing being measured and the other will consist of the combined errors of the measurement operation.

$$Observation = Product\ Value + Measurement\ Error$$

Once we have this model for an observation, we can begin to discuss the properties of the measurement system by considering the properties of the measurement errors themselves. In most cases the properties of greatest interest will be consistency, precision, and bias. As always, the question of consistency (or homogeneity) is primary.

THE CONSISTENCY CHART

Fifty years ago Churchill Eisenhart, a well known statistician working at the Bureau of Standards, effectively gave us the operational definition of a consistent measurement process when he wrote:

> "Until a measurement process has been 'debugged' to the extent that it has attained a state of statistical control it cannot be regarded, in any logical sense, as measuring anything at all."

If repeated measurements of the same thing are placed on an *XmR* chart they should display a reasonable degree of homogeneity (known as statistical control fifty years ago). When 20 to 50 repeated measurements of the same thing appear to be homogeneous when placed on an *XmR* chart, then the measurement process can be said to be consistent. When *any number* of repeated measurements of the same thing display a lack of homogeneity the measurement process is inconsistent and "cannot be regarded in any logical sense, as measuring anything at all."

As an example of a consistency chart I shall use 100 weekly measurements of a standard at the Bureau of Standards from the time period when Dr. Eisenhart made the statement above. The standard is a ten-gram weight known as NB10. The scale used to weight NB10 measured the weight to the nearest millionth of a gram. Figure 1 shows the *X* chart for 100 weighings of NB10.

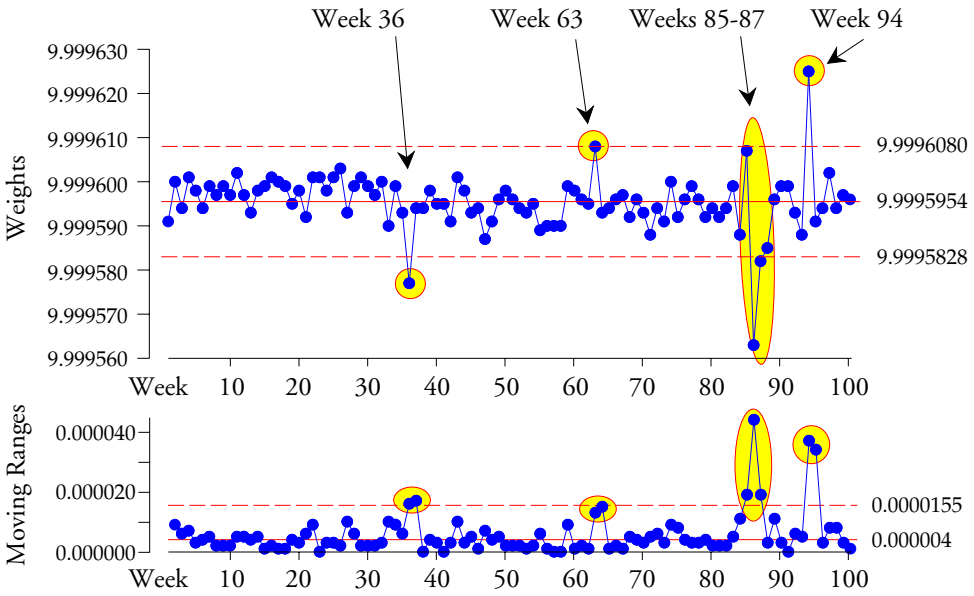

Figure 1: A Consistency Chart for 100 Weighings of NB10 by the Bureau of Standards

This measurement process was operated consistently for extended periods of time, yet during this two-year period there were three weeks and one four-week period when the measurement system was not operated consistently. While measuring a ten-gram weight to the nearest millionth of a gram is a bodacious feat of measurement, the upsets shown tell us that there are some factors that affect this measurement process that are not always being controlled in practice. The reason for each of these upsets needs to be understood in order to get the most out of this measurement process.

If the measurements of a standard at the Bureau of Standards are occasionally inconsistent, what can you say about your measurement systems? The only way to answer the question of consistency for a measurement process is to use a consistency chart. *Anything* else is simply wishful thinking.

PRECISION

When your measurement process is consistent you can use the XmR chart to answer the question about the precision of the measurement. Since the measurement process in Figure 1 is intermittently consistent we will use it as our example. The median two-point moving range is 4.0 micrograms. Dividing by the bias correction factor known as $d_4 = 0.954$ results in an estimate of the standard deviation of measurement error of 4.2 micrograms.

To make sense of this value of 4.2 micrograms it is helpful to convert it into the probable error of a measurement. The probable error of a measurement is 0.675 times the estimated standard deviation of the measurement errors. Here we get a value of 2.8 micrograms. This value is the median error of a single observation. Half the time a measurement will err by more than 2.8 micrograms, and half the time a measurement will err by less than 2.8 micrograms. Thus, while these measurements are recorded to the nearest microgram, the measurements are not really good to one microgram. In the words of Nenad Sarcevic, a student of mine, the probable error of

2.8 micrograms is the *demonstrable resolution* of the measurement. While these measurements are recorded to the nearest microgram, they are actually good to within three micrograms.

BIAS

When you use a known standard to create your consistency chart you can use that chart to evaluate the bias of a consistent measurement system. Bias refers to the average value for the measurement errors. A consistent measurement system that produces measurement errors having a zero mean is said to be unbiased. When the average value for the observed measurement errors is detectably different from zero the measurement system is said to be biased.

Clearly, bias has to be defined relative to some master measurement system. In the case of Figure 1 we are looking at values obtained using the master measurement system with a standard having an accepted value of 10 grams. However, since the central line of the consistency chart is 9.999594 grams, we can say that either the scale is biased by approximately 406 micrograms, or NB10 is 405 micrograms light, or some combination of these factors is present. Without reference to other data involving other standards, or without weighing this standard using other scales we cannot say which is the case here. However, whatever happened in Week 36 affected this bias. As may be seen in Figure 2, the average weights obtained for NB10 were about 3 micrograms lighter following the upset in Week 36.

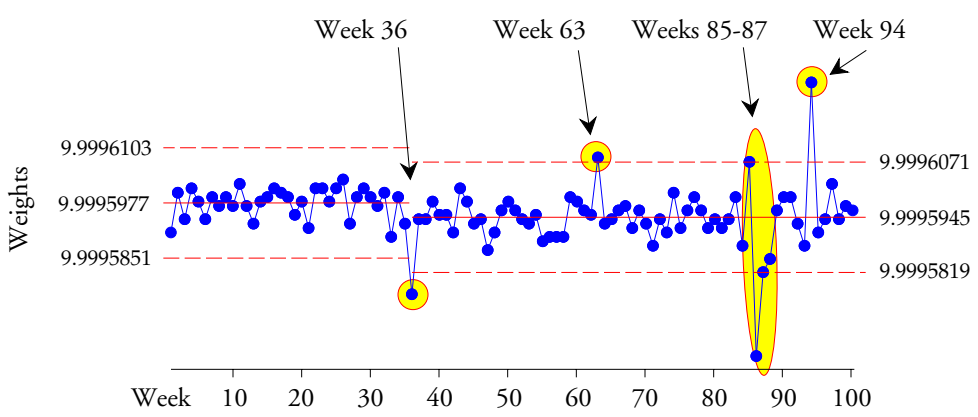

Figure 2: The Bias in NB10 Changed After Week 36

PRECISION WITHOUT CONSISTENCY ?

Can't we talk about precision and bias without having to worry about consistency?

Many have done so only to discover that their work was about as good as an annual weather forecast from the Farmer's Almanac. For example, a traditional way of looking at the precision of a measurement process is to measure a single item 30 times and then compute the standard deviation statistic as the estimate of measurement error. The next example will illustrate the problem with this approach.

The parts being measured were the inserts for the rims of steering wheels. These consisted of steel rods bent into a circle and welded to form rings. A new, high-tech, electronic measurement

system had been installed to measure these inserts. An insert would be placed on a backlit plate where it would be held in place by positioning pins. A camera would take a picture of the part and a computer would count the pixels within the image of the ring to get the effective area. Next the effective diameter of the insert would be computed and displayed. All nice and neat. Put down the part, push the button, get your number.

Richard Lyday wanted to know how good this new measurement system was, so he took a single part, and repeatedly placed the part on the plate, pushed the button, and recorded the diameters. After 30 measurements of the same part obtained over the course of one hour he had an average of 13.495 inches, and a standard deviation statistic of 0.114 inches. If we take the traditional approach and use this value as an estimate of measurement error we would get an erroneous probable error of 0.08 inches. This value would suggest that this fancy, high-tech, automated, electronic measurement system should be good to about one-twelfth of an inch. This would be about as good as measurements obtained using a yardstick. However, this computation does not consider the question of consistency.

When Richard placed his 30 measurements on a consistency chart he got the chart in Figure 3. There we see that this one steel part grew a quarter of an inch in diameter over the course of an hour and that it would occasionally have a very small diameter!

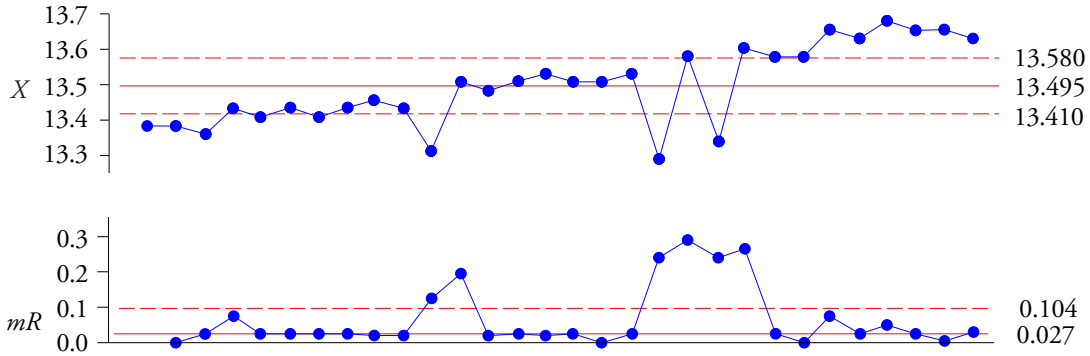

Figure 3: Consistency Chart for 30 Determinations of the Diameter of a Single Insert

The trend in Figure 3 was a problem with the camera. As the camera warmed up the pixel size got smaller, the computer would count more pixels, resulting in a larger diameters.

The small diameters would occur whenever the 800-ton press cycled while the computer was attempting to count the pixels in the video image. When the image vibrated the computer would lose count, and a small diameter would be the result. With the camera mounted on the roof truss, these upsets were an inherent feature of the way this measurement system was installed.

So how good was our description of this measurement system using the traditional approach? What we have here is a fancy, high-tech, electronic, rubber-ruler. As it turns out, we can do much better using a yardstick. A measurement process that does not display consistency "cannot be regarded, in any logical sense, as measuring anything at all." While we may compute the descriptive statistics for a set of inconsistent measurements, those statistics have no meaning in terms of the measurement process because an inconsistent measurement process does not possess a demonstrated precision or a consistent bias.

So the first three aspects of a measurement system are consistency, precision, and bias. However, without consistency you cannot estimate the precision or the bias simply because you do not actually have a measurement system. You merely have a way of generating random numbers. Do not mistake these random numbers for measurements.

PRECISION WITH CONSISTENCY

With consistency you can use your measurements in virtually any statistical analysis. The heart of every modern statistical analysis technique is the separation of signals from noise, and measurement error (or precision) is simply part of the noise that gets filtered out before we identify our potential signals. Of course, if the measurement error is large it will increase the noise level and make it harder to find any potential signals, but in any case, those potential signals that we do find are large enough to show up in spite of the amount of noise present.

If you place your data on a process behavior chart and find signals you do not need to worry about your measurements having sufficient precision—you need to identify the assignable causes of the exceptional variation. If you use your measurements to conduct an experiment and you have detectable signals, then you do not need to worry about the precision of your measurements, but you need to understand what the signals within your data represent and take action accordingly. So with regard to a statistical analysis, the question of precision becomes important only when you are no longer finding any potential signals.

On the other hand, when you are using a measurement to describe a specific item, or to characterize a specific item with respect to specifications, the precision of the measurement will affect your interpretation and use of the value obtained. The probable error will define the demonstrable resolution of the value, and may be used to define appropriate guard-bands around the specifications. For more about this topic see "Is the Part in Spec?" ***QDD,*** June 1, 2010.

USING IMPERFECT DATA

So, rather than having to have nearly perfect data before you do anything, the filtration that is part of any statistical analysis will allow you to use less-than-perfect data as long as they come from a consistent measurement process. How much less than perfect? To answer this question we will need a measure of relative utility. For the process behavior chart this measure is known as the Intraclass Correlation Coefficient:

$$\textit{Intraclass Correlation Coefficient} = \frac{\textit{Variance of the Product Values}}{\textit{Variance of Product Observations}}$$

A consistent set of measurements may be used on a process behavior chart whenever the Intraclass Correlation Coefficient exceeds 20%. This guideline is justified and explained in my article: "The Intraclass Correlation Coefficient" ***QDD***, December 2, 2010. The following summaries come from this article.

Whenever the Intraclass Correlation Coefficient is greater than 80% your consistent measurement process can be said to be a first class monitor. First class monitors will detect a three standard error shift more than 99% of the time using detection rule one of the Western Electric Zone Tests. With a first class monitor any signal coming from the production process will be attenuated by less than 10% due to the effects of measurement error.

Whenever the Intraclass Correlation Coefficient is between 80% and 50% your consistent measurement process can be said to be a second class monitor. Second class monitors will detect a three standard error shift more than 88% of the time using detection rule one of the Western Electric Zone Tests, and they will detect a three standard error shift 100% of the time using all four of the Western Electric Zone Tests. With a second class monitor any signal coming from the production process will be attenuated by less than 30% due to the effects of measurement error.

Whenever the Intraclass Correlation Coefficient is between 50% and 20% your consistent measurement process can be said to be a third class monitor. Third class monitors will detect a three standard error shift more than 91% of the time using all four of the Western Electric Zone Tests. With a third class monitor any signal coming from the production process will be attenuated by up to 55% due to the effects of measurement error. With third class monitors a consistency chart on the measurement process is imperative.

Thus, we have approximately a 90% chance or better of detecting a three standard error shift using the Western Electric Zone Tests whenever our Intraclass Correlation Coefficient exceeds 20%. As may be seen in Figure 4, the probabilities of detecting a three standard error shift rapidly drop once the intraclass correlation goes below 20%, making any use of a fourth class monitor an act of desperation.

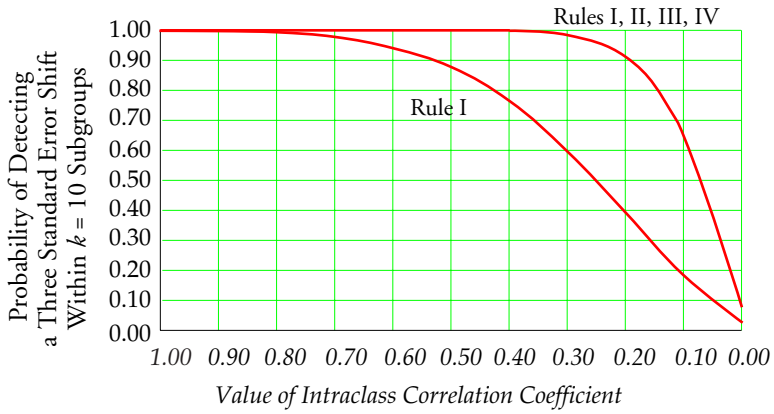

Figure 4: The Effect of Intraclass Correlation upon Detecting a Three Standard Error Shift

SUMMARY

While it may sound reasonable to begin any improvement effort with an evaluation of the measurement process, and while no one can be against having good measurements, it is a fact that you can use less-than-perfect measurements to improve your production process. As long as you have a consistent measurement process, any modern statistical analysis will filter out the noise before identifying any potential signals. This means that automatically performing a gauge R&R study is not the right starting place for your improvement activities. (While a consistency chart would be a better starting place than an R&R study, the curves in Figure 4 show that this is not absolutely necessary in every case.)

When you have a first or second class monitor the effects of measurement error are so slight that they can usually be ignored. Any signals on the process behavior chart are more likely to have come from the production process than from the measurement system. This makes it

possible to simply start off with a process behavior chart without having to pre-qualify your measurement system. If you find no signals, then it is appropriate to turn to a consideration of the measurement system.

If you have a third class monitor you will probably already know that you are pushing the limits of your measurement system. Here it is appropriate to start with a consistency chart. Any signals found on a consistency chart represent opportunities to improve the measurement system, which, in turn, will improve the Intraclass Correlation Coefficient. If your measurement system is consistent, any signals on a process behavior chart for the production process are logically interpreted as having come from the production process.

A good measurement is one that has utility for the user. You can detect process improvements whenever the Intraclass Correlation Coefficient is greater than 20%. Consistency is more critical than having a small measurement error or a large Intraclass Correlation. So quit worrying about the quality of your measurements, and start placing your measurements on process behavior charts to learn how to improve both your production processes and your measurement systems. This approach will not only simplify your improvement projects but it will also tend to yield better results more quickly than any other approach.