Evaluating Destructive Measurements

You can still put an upper bound on measurement error

Donald J. Wheeler

What can be done when a test is destructive? How do we characterize measurement error? How can we determine if a test method is adequate for a given product or application? How can we check for bias?

All of the techniques for assessing the quality of a measurement system require us to make multiple measurements of the same thing. This allows us to isolate the test-retest error from the product variation so that we can estimate and characterize measurement error. However, when the act of obtaining a measurement also destroys the sample being measured it becomes impossible to test the same thing twice. So what do we do?

In those cases where we can obtain paired samples that are thought to be reasonably similar, we test the paired samples with our destructive measurement system, and use the difference between the values obtained as an approximation to the test-retest error. While this difference will also contain some amount of product variation, hopefully it will be small. By doing this repeatedly and averaging the results, it is possible to obtain an upper bound on the test-retest error for the measurement system. Our first two examples will illustrate this approach.

When there is no basis for making a judgment about the similarity of two samples, then there can be no way to isolate the variation in the measurement system from the variation in the product stream. Measurement error and product variation will be inseparable. However, it is still possible to place an upper bound on measurement error. Our third example will illustrate this situation.

In both cases you can still use your data with process behavior charts because they work with less than perfect data—the limits automatically incorporate the uncertainties due to measurement error. This is why you don't have to qualify your measurement process before using a process behavior chart. (See "The Intraclass Correlation Coefficient," ***QDD***, Dec. 2010.)

USING THE UBBELOHDE TUBE

The viscosity of product 20F is routinely measured by obtaining a sample, stirring it, splitting the sample into two subsamples, and then measuring each subsample with a Ubbelohde tube. In this case the U-tube is read after a period of 60 seconds. These two determinations of viscosity are then averaged and the average value is the reported measurement. With such duplicate testing of split samples the differences between the two determinations will provide a reasonable estimate of test-retest error. Figure 1 shows the values found for ten batches of Product 20F and the Average and Range Chart corresponding to these ten subgroups of size two.

The reader is cautioned here that the average and range chart in Figure 1 is not a process behavior chart, but is rather a one-time analysis for a finite data set known as a short EMP study (Evaluation of the Measurement Process). The fact that each subgroup consists of the two

determinations of the viscosity of a sample from a single batch means that the within-subgroup variation does not capture the batch-to-batch variation of the production process. Instead it captures the test-retest error in the use of the U-tube plus any small differences that might exist between the two parts of the samples. Thus, the average range will provide an estimate of the test-retest error. The range chart will check these test-retest errors for consistency, and the limits on the average chart will define that amount of variation in the reported averages that could be due to test-retest error alone. Since we want to be able to detect the batch-to-batch differences above and beyond the effects of measurement error, we want to find points outside the limits on this average chart. The more points outside the limits, and the further outside the limits these points lie, the greater our ability to discriminate between the batches of Product 20F. In this case, while measurement error is present, we can still discriminate between batches. (Those who are not familiar with EMP studies may need to reread this paragraph. For more on this topic see "A Better Way to Do R&R Studies," ***QDD***, February 2011, and "Rational Subgrouping," ***QDD***, June 2015)

Batch	42	43	51	52	59	61	64	76	79	81
U-Tube	29980	28970	29850	29670	29360	29320	30480	29400	29270	29810
Readings	29930	28730	30190	29730	29940	29300	30180	29630	28820	29530
Averages	29955	28850	30020	29700	29650	29310	30330	29515	29045	29670
Ranges	50	240	340	60	580	20	300	230	450	280

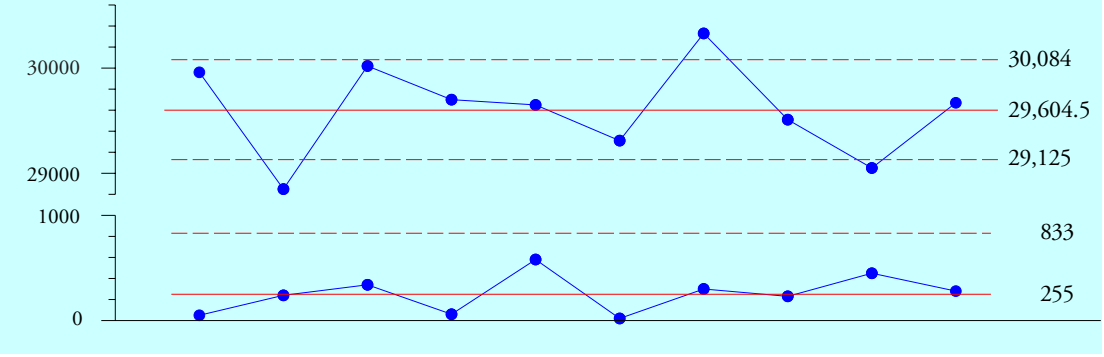


Figure 1: Average and Range Chart for Duplicate U-tube Viscosity Measurements for Product 20F

In Figure 1 we find all of the ranges within the limits and so we judge the measurement procedure to be consistent. There we also see that the U-tube readings were recorded to the nearest 10 centistokes (cs), and the averages of the duplicate readings were reported to the nearest 5 cs. But how good are these values?

The average range for the duplicated readings is 255 cs. Dividing by the bias correction factor of 1.128 for subgroups of size two, we obtain an upper bound estimate of the test-retest error for a single determination of:

$$\text{Test-Retest Error for a Single Determination} = \frac{255 \text{ cs}}{1.128} = 226 \text{ cs}$$

Thus, the test-retest error for the average of two determinations of the same batch would be:

$$\text{Test-Retest Error for Average of Two Determinations} = \frac{226 \text{ cs}}{\sqrt{2}} = 160 \text{ cs}$$

This value of 160 cs is the estimate of the standard deviation of the measurement error in the reported values. To determine what this means in practice, we multiply by 0.675 to convert the estimated standard deviation into the Probable Error of a reported value:

$$\text{Probable Error for Average of Two Determinations} = 0.675\ (160\text{ cs}) = 108\text{ cs}$$

This is the demonstrated resolution of the reported values. The Probable Error is the median amount by which a measurement will err. So even though the averages of these duplicate determinations are recorded to the nearest 5 cs, they will err by 108 cs or more half the time, and they will err by 108 cs or less half the time. Thus, these measurements are only good to the nearest 100 cs. Reporting the averages to the nearest 5 cs is an example of writing down more digits than the measurement system will support. Here we have a discrepancy between the measurement increment used in recording the data and the effective resolution of those data. Just writing down more digits does not always increase the precision of the measurements.

In general we want our measurement increment to be about the same size as the Probable Error. Acceptable measurement increments range from twice the size of the Probable Error down to one-fifth the size of the Probable Error. In this case appropriate measurement increments would range from 216 cs down to 22 cs. So, recording each individual determination to the nearest 100 cs, and reporting the averages to the nearest 50 cs would be appropriate here.

If any of the range values in Figure 1 had fallen above the upper range limit it would have been interpreted as an inconsistency between the duplicated readings. Here it would have suggested that either the two subsamples from a given batch were detectably different or else the measurement process was applied inconsistently to the two subsamples. If the measurement process is inconsistent, then the reason for this inconsistency needs to be found so that he measurement process can be fixed. If the subsamples are thought to have been different, then ranges above the range limit may be deleted from the computation of the average range before estimating the effects of measurement error.

THE CONE AND PLATE METHOD

The same ten batches of Product 20F were also used to evaluate a new method of measuring viscosity. This automated method, known as the Cone and Plate method, determines the viscosity of a sample five times and reports the average value. Here the viscosities are recorded to the nearest 100 cs, and so the averages are reported to the nearest 20 cs. The data for our ten batches of Product 20F are shown in Figure 2.

The Range Chart shows no evidence of inconsistency in this measurement process. The average range of 150 cs, when divided by the bias correction factor of 2.326 for subgroups of size five results in an estimate of test-retest error for a single determination of:

$$\text{Test-Retest Error for a Single Determination} = \frac{150\text{ cs}}{2.326} = 64.5\text{ cs}$$

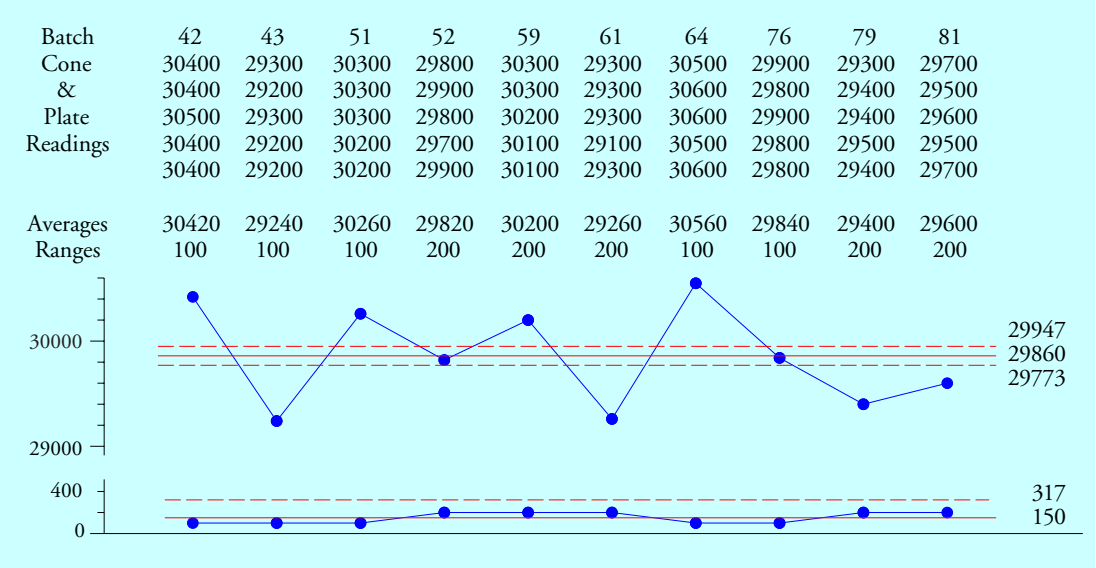
Thus, the test-retest error for the average of five determinations of the same batch would be:

$$\text{Test-Retest Error for Average of Five Determinations} = \frac{64.5\text{ cs}}{\sqrt{5}} = 28.8\text{ cs}$$

This value of 28.8 cs is the estimate of the standard deviation of the cone and plate measurement

process. To determine what this means in practice, we multiply by 0.675 to obtain the Probable Error of a reported value:

$$\text{Probable Error for Average of Five Determinations} = 0.675\ (\ 28.8 \text{ cs}\ ) = 19.5 \text{ cs}$$



Batch	42	43	51	52	59	61	64	76	79	81
Cone	30400	29300	30300	29800	30300	29300	30500	29900	29300	29700
&	30400	29200	30300	29900	30300	29300	30600	29800	29400	29500
Plate	30500	29300	30300	29800	30200	29300	30600	29900	29400	29600
Readings	30400	29200	30200	29700	30100	29100	30500	29800	29500	29500
	30400	29200	30200	29900	30100	29300	30600	29800	29400	29700
Averages	30420	29240	30260	29820	30200	29260	30560	29840	29400	29600
Ranges	100	100	100	200	200	200	100	100	200	200

Figure 2: Average and Range Chart for Cone and Plate Viscosities for Product 20F

Here the viscosities are reported to the nearest 20 cs, and they are good to the nearest 20 cs. They will err by 20 cs or less half the time. The measurement increment and the Probable Error are essentially the same.

The smaller measurement error for the cone-and-plate method results in tighter limits on the average chart and therefore a greater ability to discriminate between batches of Product 20F. While eight averages are outside the limits in Figure 2, only three averages were outside the limits in Figure 1.

CHECKING FOR BIAS BETWEEN METHODS

The Probable Error characterizes a measurement system in an absolute sense by showing the demonstrated resolution of the measurements. The duplicated U-tube readings result in averages that are good to the nearest 108 cs. The automated Cone and Plate averages are good to the nearest 20 cs. Therefore the effective resolution of the Cone and Plate method is about five times smaller than the effective resolution of the U-tube method. But is there a bias between these two measurement systems?

Figure 3 places the two average charts side by side using the same vertical scale. There we see a reasonable degree of parallelism between the two methods as they measure the same ten batches of Product 20F. This is desirable since we would like the two methods to agree as to which batches are high and which batches are low. However, the two methods do not have similar grand averages, and this may be indicative of a bias between the two methods.

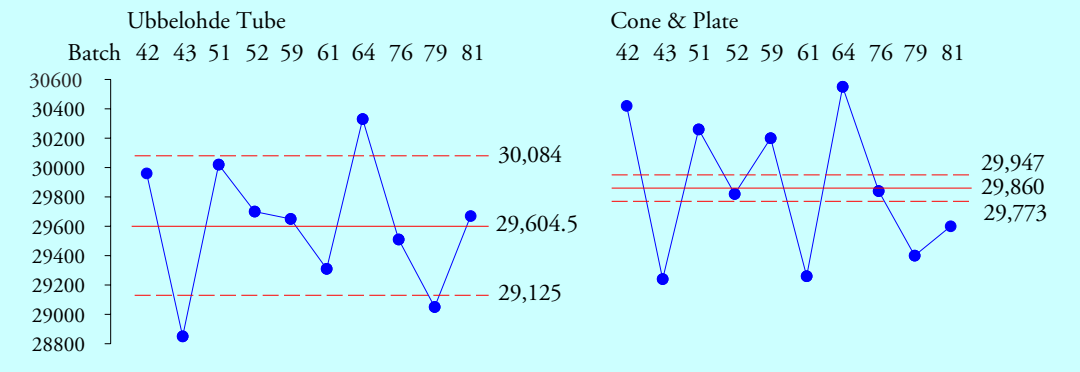


Figure 3: Average Charts for U-tube and Cone & Plate Viscosity Measurements

Since the same batches were tested using each of the methods we can use a paired-t test to check for detectable bias between these two measurement systems. Figure 4 shows the reported values for the viscosities for each of the ten batches as well as the differences found when the U-tube values are subtracted from the C&P values.

Batch	42	43	51	52	59	61	64	76	79	81
C & P Avg	30420	29240	30260	29820	30200	29260	30560	29840	29400	29600
U-tube Avg	29955	28850	30020	29700	29650	29310	30330	29515	29045	29670
Differences	465	390	240	120	550	–50	230	325	355	–70

Figure 4: Differences in Viscosities for Ten batches of Product 20F

The average of the ten differences in Figure 4 is 255.5 and the standard deviation statistic for these ten differences is 206.0. With nine degrees of freedom the 99% Interval Estimate for the expected value of these differences is:

$$\textit{Average} \pm t_{.005}\frac{s}{\sqrt{n}} = 255.5 \pm 3.250\frac{206.0}{\sqrt{10}} = 255.5 \pm 211.7 = 43.8 \text{ to } 467.2$$

This interval estimate for the difference *does not contain zero.* Therefore, *detectable bias exists* between these two measurement systems. The Cone and Plate Method produces values that average about 255 centistokes higher than those obtained from the Ubbelohde Tube.

Before we condemn the Cone & Plate measurements as being biased relative to the standard U-tube readings, it is important to note that the U-tube readings in Figure 1 were obtained after 60 seconds. The standard operating procedure for the U-tube calls for the tube to be read after a period of 300 seconds. Thus, while there is a detectable bias between the automated Cone & Plate method and the U-tube method used, we cannot say what would happen if the SOP for the U-tube had been followed.

In these two examples it was possible to split our samples into subsamples that were essentially the same. This allowed us to evaluate our destructive tests and quantify the effects of measurement error. In the next example we look at a problem where such matched subsamples are not available.

VOLATILE MATERIALS

A company wanted to know how much uncertainty was contained in a standard test for the

volatile material content for product 1109. The test procedure evaporates the volatiles from a 250 gram sample in two or more stages. The sample is weighed following each stage. When the post-stage weight of the sample changes less than 0.1 gram between successive stages the sample is considered to be volatile-free. A comparison of the original weight and the final weight permits a computation of the percentage of volatile material in the sample.

Unlike the previous example where the liquid could be stirred and split into similar samples, the samples here may vary even when they come from adjacent portions of the product stream. So now we will use sets containing multiple samples and will seek to get each set of samples to represent a cross-section of the product variation.

Six technicians who routinely performed the volatile material test were selected to participate in a special study. One bale was selected on each of three days to capture the variation in the product stream. Each bale was divided up into five zones, and each zone was divided up into 250-gram samples. Next, each technician was given one sample from each zone from each bale. This was done so that each technician had samples that reflected the routine within-bale variation for each bale. Each technician tested his five samples from each bale, and these results were organized into an average and range chart having eighteen subgroups of size five.

Tech	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6
Part	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
	1.2	1.2	1.1	1.1	1.4	1.7	0.7	1.3	1.3	0.7	0.7	1.1	0.9	1.0	1.1	0.8	2.0	0.7
	0.4	1.4	1.1	1.4	2.1	1.7	0.9	1.1	1.4	0.4	0.8	1.2	1.1	1.1	0.9	0.9	1.4	0.9
	0.7	1.4	1.2	1.5	1.9	1.4	0.7	1.2	1.1	0.7	1.2	1.2	1.0	1.3	1.3	0.9	0.9	0.9
	0.9	1.9	1.0	1.2	1.7	1.5	1.1	0.9	1.3	0.7	1.4	1.2	1.1	1.1	1.3	0.9	1.7	0.9
	1.0	1.3	1.4	0.9	1.1	1.4	0.9	0.7	1.3	0.6	1.7	0.9	1.1	1.1	1.1	1.2	1.5	1.1
Average	0.84	1.44	1.16	1.22	1.64	1.54	0.86	1.04	1.28	0.62	1.16	1.12	1.04	1.12	1.14	0.94	1.50	0.90
Range	0.8	0.7	0.4	0.6	1.0	0.3	0.4	0.6	0.3	0.3	1.0	0.3	0.2	0.3	0.4	0.4	1.1	0.4

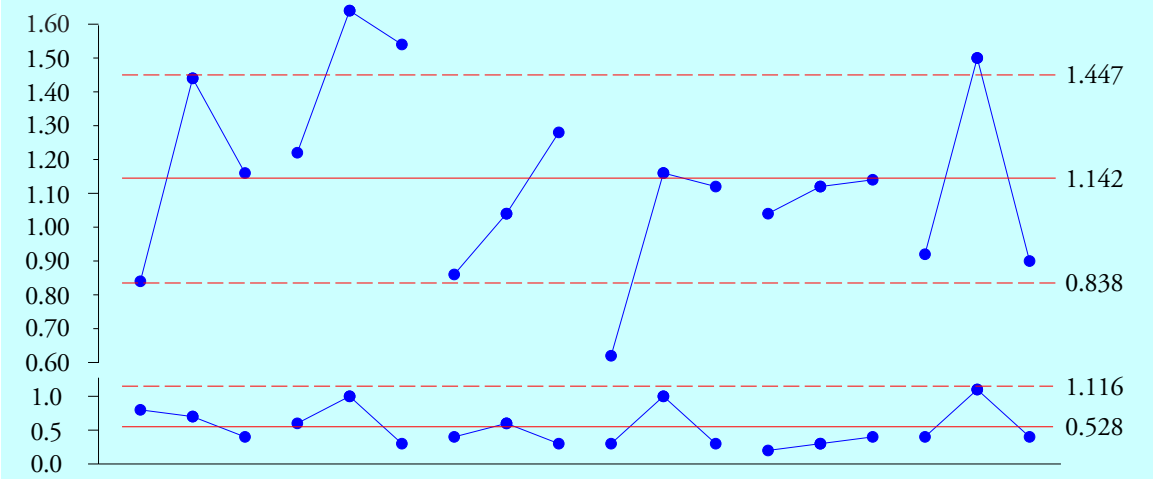


Figure 5: Volatile Materials Content Test

Here the range chart shows both the test-retest error of the measurement system and the routine within-bale variation seen in product 1109. Since no points fall above the upper limit we have no evidence of inconsistency within any of the subgroups. The average range can be used to place an upper bound on the test-retest error. When we divide the average range of 0.528 by the

bias correction factor of 2.326 for subgroups of size five, we obtain an estimate of the standard deviation of the measurement process plus within-bale variation. This value is 0.227 percent. Multiplying by 0.675 we estimate the Probable Error of these measurements to be 0.153 percent. This estimate includes both the test-retest error of the test method and the within-bale variation found in product 1109.

Unfortunately, the average chart tells us that there is more to the story. The limits on the average chart define that amount of variation in the subgroup averages that can be obscured by measurement error. The fact that some averages fall outside the limits is encouraging. The fact that the curves for each technician seem to shift up and down suggests that there may well be a technician bias effect present. The fact that the curves for each technician do not all have approximately the same shape suggests that there is a technician by part interaction present in these data.

If the curves for each technician had shown reasonable parallelism we could easily check for a technician bias effect using the Analysis of Main Effects (ANOME). However, when interaction effects are indicated by substantial non-parallelism I prefer to use the Analysis of Variance (ANOVA). In this case we have a two-factor, fully-crossed layout, and the standard ANOVA table for these data is shown in Figure 6.

Source	Sum of Squares	df	Mean Squares	F-ratios
Bale	1.3476	2	0.6738	12.28 *
Technician	2.8729	5	0.5746	10.47 *
Interaction	1.8471	10	0.1847	3.37 *
Within	3.9520	72	0.0549	
Total	10.0196	89		

Figure 6: ANOVA Table for Volatile Material Content Test

Here we find a detectable difference between the bales, which is the product variation we want to be able to detect in spite of the effects of measurement error. Next we find a detectable difference due to technicians, which would be our license to estimate the technician biases if we didn't have the interaction effect.

But the monkey wrench is the detectable interaction effect. This means that the non-parallelism in Figure 5 actually represents different technicians getting different results when measuring the same product. In short, this is an inconsistency in the measurement process.

With destructive testing, an interaction effect like this might occur due to differences in the samples tested by each technician. If we had simply assigned samples to technicians in a haphazard manner we would not know if the interaction effect was due to inconsistencies in the measurement process or differences in the samples tested. This is why we were careful to make sure that each subgroup contained samples from each zone of a bale. By working to assure that each technician tested samples that covered the whole spectrum of product variation we minimize the opportunity for spurious interaction effects. Since in this case the range chart shows no evidence of inconsistency within the subgroups, it is unlikely that the interaction effect comes from differences in the samples tested. Therefore we interpret this interaction as a real effect indicating an inconsistency in the measurement process.

Thus, it appears that this measurement process contains biases and inconsistencies. What do these biases and inconsistencies do to the precision of the measurements? To answer this question we reanalyze these data by rearranging them into three subgroups of size 30 according

to the bale sampled.

The Average and Standard Deviation Chart for these three subgroups of size 30 is shown in Figure 7. There we see that the measurement system that currently exists can detect the variation in the product stream for product 1109. However, if these three bales represent the full range of variation in the product stream, then the width of these limits suggests that this measurement system is likely to be no more than a Third Class Monitor.

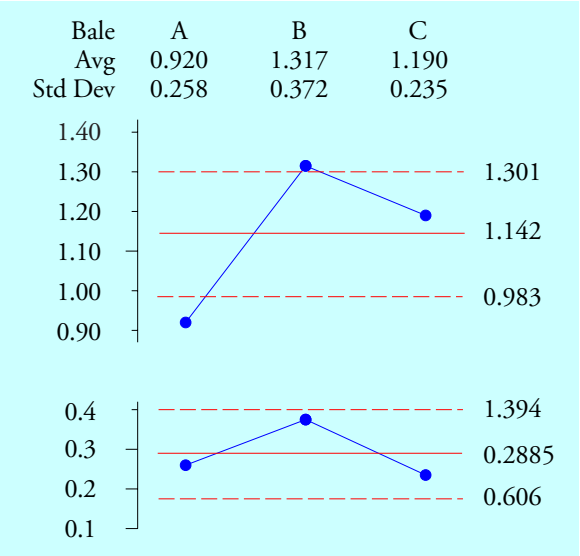


Figure 7: Average and Standard Deviation Chart for Bales

The average standard deviation statistic of 0.2885 is an estimate of the standard deviation of the test method that includes the effects of the technician biases and the inconsistency represented by the interaction effect. Multiplying by 0.675 results in a estimated Probable Error of 0.195 percent.

When compared with the inherent Probable Error of 0.153 percent found earlier, we see that the biases and inconsistencies found here do not greatly degrade these measurements. The values are recorded to the nearest 0.1 percent, and they are good to within 0.2 percent at least half the time. Thus, when using this test, the values obtained should not be interpreted more precisely than plus or minus 0.2 percent.

(While all of these examples are real, in the last example the data were transformed for confidentiality. So while the results shown are illustrative, they are not actually indicative of the actual precision of the test method discussed.)

SUMMARY

When the test method is destructive it becomes more difficult to characterize measurement error. Two approaches have been illustrated that allow you to place an upper bound on the size of measurement error. Having such an upper bound will let you do business with some assurance about the quality of the measurements being used.