When Are Instruments Equivalent? Part One

Practical answers to an age-old question

Donald J. Wheeler and James Beagle III

As soon as we have two or more instruments for measuring the same property the question of equivalence raises its head. This paper provides an operational definition of when two or more instruments are equivalent in practice.

Dr. Churchill Eisenhart, while working at the U.S. Bureau of Standards in 1963, 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." Before we begin to talk about the equivalence of measurement systems we need to know whether we have yardsticks or rubber rulers. And the easiest way to answer this question is to use a consistency chart.

CONSISTENCY FOR A MEASUREMENT SYSTEM

In order to evaluate a measurement system we will need some sort of study where multiple determinations are made of the same thing or the same group of things. Perhaps the simplest form for studies of this sort is to measure the same thing repeatedly. When these values are placed on an *XmR* chart we end up with a test for the consistency of the measurement system. Figure 1 shows the consistency chart for thirty determinations of a standard using instrument A. (The readings shown on the chart have been coded by subtracting 400 from each observation.)

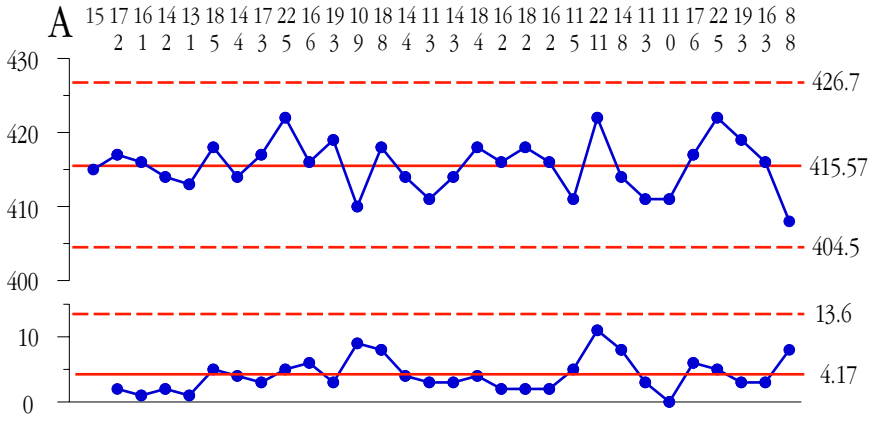

Figure 1: Consistency Chart for 30 Repeated Measurements of a Single Standard with Inst. A

Here we find all of the points within the limits. This is taken as evidence of a consistent

measurement process. If any of the points on a consistency chart fall outside the limits they represent strong evidence that the measurement system is inconsistent.

Any attempt to characterize an inconsistent "measurement system," or to compare it with other measurement systems, will be premature. (A measurement system that gives inconsistent results is not even equivalent to itself, so how can it ever be equivalent to another instrument?) All that follows is built on the assumption that each of the measurement systems being compared has already demonstrated its consistency by means of a consistency chart. (For more about consistency charts see Wheeler "Consistency Charts," ***QDD,*** April 2, 2013.)

THE EFFECTIVE RESOLUTION OF A MEASUREMENT

Since a consistency chart is built upon multiple readings of the same thing, the variation on a consistency chart has nothing to do with product variation, so the moving ranges must be thought of as capturing the essence of measurement error. To turn the average moving range into an estimate of the standard deviation of measurement error, *SD(E)*, we divide by the bias correction factor $d_2 = 1.128$. For the instrument in Figure 1:

$$Est.\ SD(E) = \frac{average\ moving\ range}{1.128} = \frac{4.17}{1.128} = 3.7 \text{ measurement units}$$

When we multiply this value by 0.675 we convert it into the "probable error" of a single reading. This value defines the effective resolution of a single measurement. Here we estimate the probable error to be 2.5 units. Half the time a single value will differ from the average of all possible measurements by this amount or more. Thus, any single value found using instrument A should only be interpreted as being good to within ± 2.5 units.

For more on this topic see the authors article "Is That Last Digit Really Significant?" ***Quality Digest*** Feb. 5, 2018.

THE AVERAGE DISTANCE BETWEEN DUPLICATE VALUES

When comparing instruments we need to consider how two measurements of the same thing differ from each other. A lower bound for this difference is the average moving range (which is the average size of successive differences for measurements made with the same instrument). Thus, by using the formula given above in reverse, we find that the average distance between two measurements of the same thing, using the same measurement process, will be 1.128 *SD(E)*.

If duplicate readings obtained from one instrument will differ by this much on the average, we cannot expect readings obtained from two instruments to do any better. When two instruments display the equivalent amounts of measurement error and also have no bias relative to each other, two measurements of the same thing made with the two instruments will also differ by 1.128 *SD(E)* on the average. This minimum average difference between duplicate measurements is shown in Figure 2 as a horizontal line.

EQUIVALENCE BETWEEN TWO INSTRUMENTS

When one instrument has a bias relative to another, this bias will affect every reading made by the one instrument. This will, on the average, inflate the differences between the readings from the two instruments. In the absence of measurement error, the average difference between

two readings of the same thing would be a linear function of the bias. Thus, the expected effect of instrument bias alone may be shown by the 45 degree line in Figure 2.

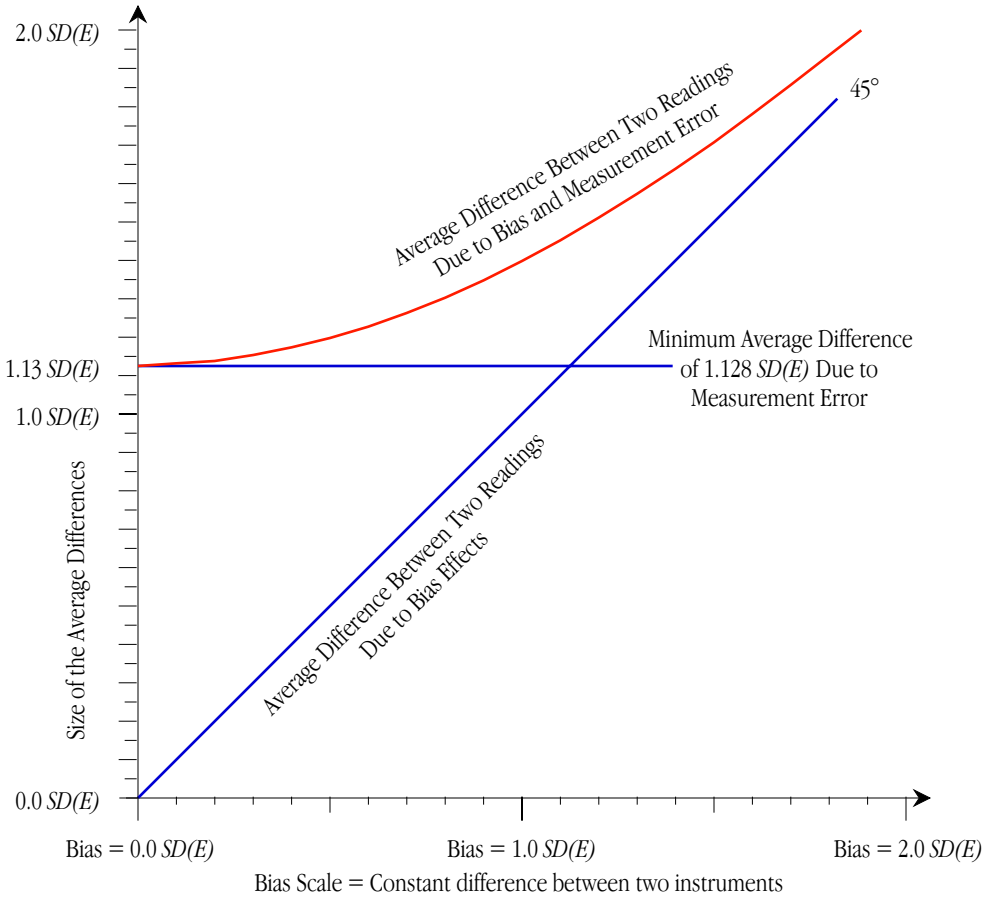

Figure 2: The Average Difference Between Two Readings

When we combine the effects of both bias and measurement error we find that the average difference between two readings of the same thing by two instruments will follow the curve in Figure 2. The points that define the curve of Figure 2 are tabled in Figure 3.

Corresponding Biases and Average Differences in *SD(E)* Units

Bias	Avg. Diff.	Bias	Avg. Diff.	Bias	Avg. Diff.
0.0	1.128	0.7	1.262	1.3	1.572
0.1	1.129	0.8	1.302	1.4	1.638
0.2	1.138	0.9	1.347	1.5	1.708
0.3	1.152	1.0	1.397	1.6	1.781
0.4	1.171	1.1	1.451	1.7	1.856
0.5	1.196	1.128	1.467	1.8	1.935
0.6	1.226	1.2	1.510	1.88	2.000

Figure 3: Bias and Corresponding Average Differences in *SD(E)* Units

Figure 2 shows that there is a region where measurement error is the dominant effect and that there is a region where bias effects dominate. The crossover point between these two regions occurs at the point where the two straight lines cross. Here the bias is equal to 1.128 *SD(E)* and

from Figure 3 the average difference between pairs of readings by the two instruments will be 1.467 *SD(E)*.

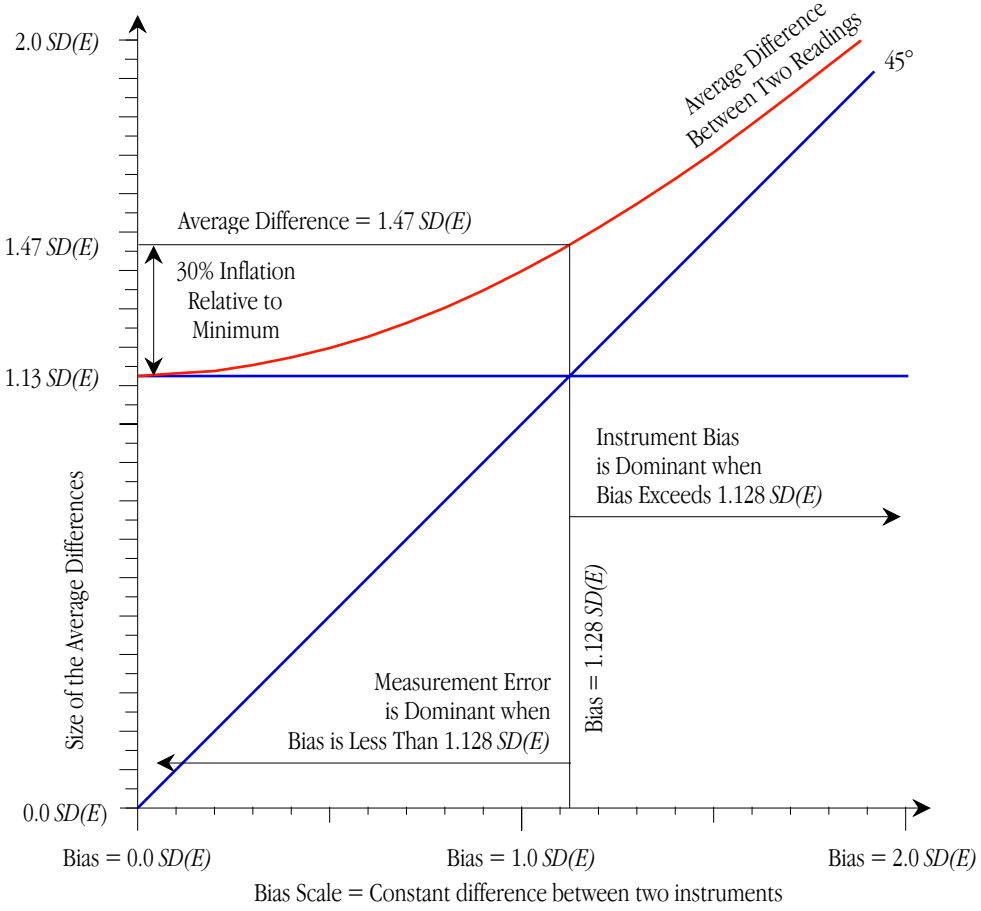

Figure 4: The Average Difference Between Two Readings

On the left side of Figure 4, where the instrumental bias is less than 1.128 *SD(E)*, the average distance between readings of the same thing using the two instruments will always be within 30 percent of the minimum value.

On the right side of Figure 4, where the instrument bias exceeds 1.128 *SD(E)*, the bias will begin to create a systematic difference between the measurements from the two instruments. Here the instruments will no longer be equivalent in practice.

So our criterion for practical equivalence becomes one of having an instrumental bias that is smaller than 1.128 *SD(E)*.

THE THREE INSTRUMENTS

In Figure 1 we have 30 determinations of the value of a standard with instrument A. Figure 5 shows the consistency charts for 30 determinations of the same standard using instruments B and C.

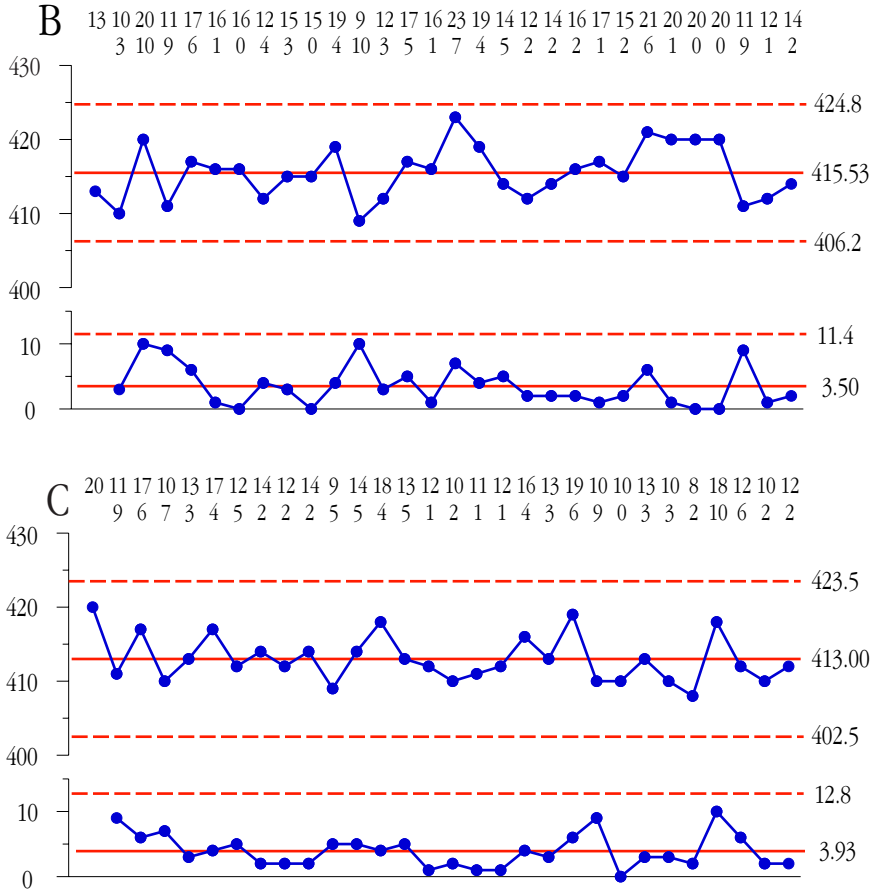

Figure 5: Consistency Charts for 30 Repeated Measurements of a Single Standard with Inst. B and Inst. C

The average range for instrument B is 3.50 units, thus the probable error is estimated to be 2.1 units. The average range for instrument C is 3.93 units, giving a probable error of 2.4 units. Earlier we found a probable error of 2.5 units for instrument A. The similarity of these probable errors suggests that these three instruments can be considered to have equivalent amounts of measurement error. A way to test this idea will be given in next month's column.

For purposes of checking for detectable instrument bias we shall treat these three sets of data as $k = 3$ subgroups of size $n = 30$. (The demonstrated consistency of the instruments justifies this arrangement of these data.) Instrument A has an average of 415.57 and a global standard deviation statistic of 3.151. Instrument B has an average of 415.53 and a global standard deviation statistic of 3.598. Instrument C has an average of 413.00 and a global standard deviation statistic of 3.569. Each of these standard deviations has 29 degrees of freedom.

Thus, the pooled variance estimate of *SD(E)* is found by squaring each standard deviation statistic, averaging these values, and taking the square root. This pooled variance estimate has $k(n–1) = 87$ degrees of freedom and is:

$$\textit{Est SD(E)} = 3.446 \text{ units}$$

Thus, from above, we would expect duplicate readings made with the *same* instrument to

differ by an average of 1.128 * 3.446 = 3.9 units.

DETECTING BIAS EFFECTS

Our methodology for determining if two or more instruments are equivalent in practice shall be the analysis of means (ANOM). Here we have three subgroups of size 30, and want to compare the three averages. We shall use the traditional alpha level for one-time analyses of five percent. In this case, the formula for the detection limits for a 5% ANOM can be written as:

$$\textit{Grand Average} \pm H_{.05}\sqrt{\frac{k-1}{k}}\ \frac{\textit{Est. SD(E)}}{\sqrt{n}}$$

The Grand Average is 414.70.

The ANOM scaling factor $H_{.05}$ is found in the tables at the end of this paper. It depends upon the alpha level of 5%, the number of averages being compared, $k = 3$, and the degrees of freedom for the estimate of dispersion, 87 d.f. Rather than interpolating for 87 d.f., we round the 87 down to the table entry of 60 and use $H_{.05} = 2.394$.

Our estimate of the standard deviation of measurement error was 3.446 units, and the subgroup size is $n = 30$. Thus our 5% ANOM detection limits are:

$$\textit{Grand Average} \pm H_{.05}\sqrt{\frac{k-1}{k}}\ \frac{\textit{Est. SD(E)}}{\sqrt{n}}$$

$$= 414.7 \pm 2.394\ (0.8165)\ \frac{3.446}{5.477} = 413.5 \text{ to } 415.9$$

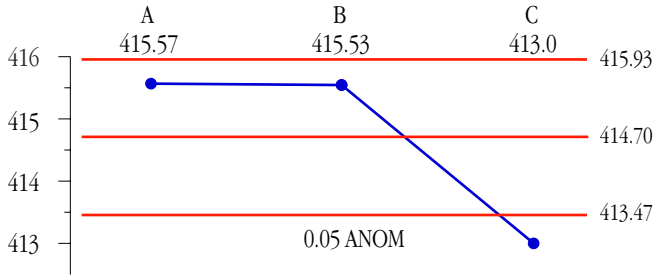

Figure 6. ANOM for Instrument Bias

Here we find that there is a detectable bias effect between Instrument C and Instruments A and B. This gives us a license to estimate this bias. The average for Instruments A and B is 415.5 so readings from Instrument C average 2.5 units less than those from Instruments A and B. This bias is very likely to be real, but is it of practical importance?

ASSESSING THE PRACTICAL IMPORTANCE OF THE BIAS

As we found earlier, the average difference between two readings made with the *same* instrument is 1.128 *SD(E)* = 3.9 units. Our detected instrumental bias of 2.5 units is smaller than this minimum average difference. We can use the table in Figure 3 to characterize the effect of this bias. The estimated bias of 2.5 units is only 73% as large as the estimate of *SD(E)* = 3.446.

$$\text{Bias} = 2.5 \text{ units} = 0.73\ \textit{SD(E)}$$

Interpolating in Figure 3 we find that this bias corresponds to an average difference between a reading from instrument C and a reading from one of the other instruments of 1.274 *SD(E)* = 4.4 units. Since these readings are recorded to whole numbers of units the instrumental bias will usually be lost in the round-off. (Two readings from the same instrument will differ by an average amount that rounds off to 4 units, while two readings from different instruments having a bias of 2.5 units will also differ by an average amount that rounds off to 4 units.

When your measurement systems fall on the left side of Figure 4 measurement error will dominate any bias present, even though you may have sufficient data to detect and estimate instrumental biases. Just because a bias is detectable does not make it of practical importance.

SO WHAT SHOULD WE DO ABOUT THIS BIAS?

Would recording the values to a tenth of a unit help? No. With an estimated standard deviation of 3.446 units the probable error for these instruments is estimated to be 2.33 units. The range of appropriate sized measurement increments will depend upon this probable error:

$$\textit{Appropriate Measurement Increments} = 0.22\,PE \text{ to } 2.2\,PE = 0.51 \text{ units to } 5.1 \text{ units.}$$

Therefore recording values to a tenth of a unit would be an exercise in recording noise. These readings are recorded to a whole number of units, and they could even be rounded off to the nearest multiple of five units without causing any serious degradation in the quality of these readings.

But could we adjust the readings from instrument C? As outlined above, such an adjustment would have a minimal impact upon the quality of the readings. Adjusting the readings from instrument C would minimally reduce the average difference between two measurements of the same thing by different instruments by 13% (from 4.4 units to 3.9 units). So, if the adjustment is easy and economical, and if you are facing a tight tolerance, it is not incorrect to adjust the readings from instrument C after the fact by this known bias of 2.5 units. However, in most cases it is not necessary to do so.

Should we recalibrate instrument C? No. Since measurement error dominates the bias effect it is unlikely that a recalibration would actually result in a smaller bias. It is worth noting that all three consistency charts show that measurements of the same standard can vary by up to ± 10 units. Attempting to recalibrate when the bias is less than 1.128 *SD(E)* will be an exercise in frustration because measurement error is likely to result in the creation new biases as you seek to remove the old biases.

SUMMARY

An operational definition has to have three parts: a criterion to be used, a method for testing compliance to the criterion, and a way of interpreting the test results. The criterion for practical equivalence for instruments having similar amounts of measurement error is for biases to be less than 1.128 *SD(E)*. (We will consider a way to check for having similar amounts of measurement error next month.) The methodology for detecting bias is the traditional ANOM. And the decision rule is contained within the criterion and technique. A detectable bias is a problem to be solved only when that bias substantially exceeds 1.128 *SD(E).*

The eighth axiom of data analysis is that you must detect a difference before you can

legitimately estimate that difference, and only then can you assess the practical importance of that difference. All statistical procedures are designed to detect differences in spite of the noise present in the data. Given enough data we can detect differences that are too small to have any practical importance. The criterion offered here provides a way to assess the practical importance of instrumental biases.

POSTSCRIPT

Those who are interested may verify the entries in Figure 3 as follows. Generate a set of standard normal random observations and consider these to be repeated measurements of the same thing (so that the variation represents nothing but measurement error, making $SD(E) = 1$). Now pair these observations up and think of the first member of each pair as "observation A" and the second member as "observation B." The average of the absolute values of the differences [A–B] of all of these pairs should be very close to 1.128. This corresponds to the no bias condition. Now add a fixed amount to each observation B (the bias amount) and recompute the absolute values of the differences between A and (B + bias) and average over all such pairs. The average absolute values for a given bias amount will correspond to the entries in Figure 3.

TABLES FOR ANOM

The following tables are excepted from more extensive tables given in ***Analyzing Experimental Data*** by Donald J. Wheeler and are used with permission. These values are appropriate for use with any *unbiased within-subgroup estimate* for the standard deviation of the k averages being compared. The general formula for computing the ANOM detection limits is:

$$\textit{Grand Average} \;\pm\; H_\alpha \;[\text{ Unbiased Est. } SD(\bar{X})\,]$$

Common unbiased estimators of the standard deviation of the subgroup averages are given in Figure 7 for data arranged into k subgroups of size n.

Statistic	**Estimator**	**Degrees of Freedom**
Ranges	$\sqrt{\frac{k-1}{n\,k}}\;\frac{\textit{Average Range}}{d_2}$	$0.88\,k\,(n-1)$
Std. Dev.	$\sqrt{\frac{k-1}{n\,k}}\;\frac{\textit{Average Std. Dev.}}{c_4}$	$k\,(n-1) - 0.2\,k$
Pooled Variance	$\sqrt{\frac{k-1\;\;\textit{Average Variance}}{n\,k}}$	$k\,(n-1)$

Figure 7: Unbiased Estimators for SD(Averages)

ANOM Critical Values H_α for $\alpha = 0.05$

$H_{.05}$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$	$k = 9$
2 *d.f.*	3.811	5.220	5.849	6.292				
3 *d.f.*	2.930	3.822	4.238	4.524	4.745			
4 *d.f.*	2.613	3.351	3.656	3.873	4.042	4.183		
5 *d.f.*	2.445	3.096	3.356	3.544	3.692	3.806	3.911	
6 *d.f.*	2.350	2.943	3.176	3.344	3.475	3.578	3.674	3.751
7 *d.f.*	2.277	2.841	3.057	3.211	3.332	3.432	3.513	3.590
8 *d.f.*	2.239	2.769	2.972	3.116	3.230	3.323	3.402	3.470
9 *d.f.*	2.198	2.715	2.908	3.045	3.153	3.242	3.317	3.383
10 *d.f.*	2.175	2.673	2.859	2.990	3.094	3.179	3.251	3.313
11 *d.f.*	2.151	2.639	2.819	2.946	3.046	3.128	3.198	3.258
12 *d.f.*	2.135	2.612	2.786	2.910	3.007	3.086	3.154	3.213
13 *d.f.*	2.119	2.590	2.759	2.880	2.974	3.052	3.118	3.175
14 *d.f.*	2.102	2.571	2.737	2.855	2.947	3.023	3.087	3.143
15 *d.f.*	2.095	2.554	2.717	2.833	2.924	2.998	3.061	3.116
16 *d.f.*	2.087	2.540	2.701	2.815	2.903	2.976	3.038	3.092
17 *d.f.*	2.079	2.528	2.686	2.799	2.885	2.957	3.018	3.072
18 *d.f.*	2.071	2.517	2.674	2.784	2.870	2.941	3.001	3.053
19 *d.f.*	2.063	2.507	2.662	2.771	2.856	2.926	2.985	3.037
20 *d.f.*	2.064	2.499	2.652	2.759	2.843	2.912	2.972	3.022
24 *d.f.*	2.039	2.471	2.620	2.723	2.805	2.871	2.927	2.977
30 *d.f.*	2.023	2.445	2.588	2.689	2.767	2.830	2.885	2.932
40 *d.f.*	2.007	2.419	2.557	2.653	2.729	2.791	2.842	2.888
60 *d.f.*	1.992	2.394	2.527	2.621	2.693	2.753	2.802	2.845
120 *d.f.*	1.976	2.362	2.492	2.582	2.650	2.707	2.755	2.797
inf. *d.f.*	1.960	2.344	2.468	2.555	2.621	2.676	2.721	2.761

ANOM Critical Values H_α for $\alpha = 0.05$ (continued)

$H_{.05}$	$k = 10$	$k = 11$	$k = 12$	$k = 13$	$k = 14$	$k = 15$	$k = 16$	$k = 17$
7 *d.f.*	3.657							
8 *d.f.*	3.528	3.586						
9 *d.f.*	3.443	3.492	3.540					
10 *d.f.*	3.369	3.423	3.462	3.502				
11 *d.f.*	3.263	3.360	3.402	3.441	3.480			
12 *d.f.*	3.265	3.311	3.354	3.389	3.428	3.457		
13 *d.f.*	3.226	3.272	3.313	3.350	3.384	3.414	3.443	
14 *d.f.*	3.193	3.238	3.278	3.314	3.349	3.379	3.409	3.438
15 *d.f.*	3.165	3.208	3.247	3.284	3.316	3.348	3.373	3.403
16 *d.f.*	3.140	3.183	3.221	3.257	3.289	3.320	3.347	3.377
17 *d.f.*	3.118	3.160	3.199	3.233	3.266	3.295	3.323	3.348
18 *d.f.*	3.100	3.141	3.179	3.212	3.244	3.273	3.301	3.326
19 *d.f.*	3.082	3.124	3.160	3.194	3.225	3.254	3.281	3.306
20 *d.f.*	3.067	3.108	3.144	3.178	3.209	3.237	3.263	3.289
24 *d.f.*	3.020	3.059	3.094	3.126	3.156	3.183	3.208	3.232
30 *d.f.*	2.973	3.011	3.045	3.075	3.104	3.130	3.155	3.177
40 *d.f.*	2.928	2.963	2.995	3.025	3.052	3.078	3.101	3.122
60 *d.f.*	2.884	2.918	2.949	2.978	3.003	3.027	3.049	3.070
120 *d.f.*	2.833	2.866	2.896	2.923	2.947	2.970	2.991	3.011
inf. *d.f.*	2.796	2.827	2.855	2.881	2.904	2.926	2.946	2.965

ANOM Critical Values H_α for $\alpha = 0.05$ (continued)

$H_{.05}$	$k = 18$	$k = 19$	$k = 20$	$k = 24$	$k = 30$	$k = 40$	$k = 60$
15 *d.f.*	3.432						
16 *d.f.*	3.397	3.426					
17 *d.f.*	3.370	3.400	3.419				
18 *d.f.*	3.350	3.373	3.393	3.440			
19 *d.f.*	3.330	3.352	3.375	3.430	3.450		
20 *d.f.*	3.311	3.333	3.354	3.410	3.440	3.500	
24 *d.f.*	3.255	3.276	3.296	3.366	3.430	3.490	3.550
30 *d.f.*	3.198	3.219	3.238	3.305	3.387	3.470	3.540
40 *d.f.*	3.143	3.162	3.180	3.245	3.322	3.421	3.520
60 *d.f.*	3.089	3.108	3.125	3.187	3.260	3.354	3.483
120 *d.f.*	3.030	3.047	3.064	3.121	3.191	3.280	3.402
inf. *d.f.*	2.983	2.999	3.015	3.070	3.136	3.220	3.334