When Are Instruments Equivalent? Part Three

Comparisons Using Multiple Standards

James Beagle III and Donald J. Wheeler

In Parts One and Two we defined the equivalence of instruments in terms of bias and measurement error based on studies using a single standard. Here we look at comparing instruments for differences in bias or differences in measurement error while using multiple standards.

When we use multiple standards, either known or designated, to compare instruments we can see how the instruments work over a range of values. For our example we shall use three production parts as designated standards (parts A, B, and C). Four inspection fixtures used to measure precision guide rollers will be compared. Fixture 1 was located at the product designer's plant. Fixtures 2 and 3 were located at the production plants, and Fixture 4 was located at the system integrators plant. Due to critical nature of the parameter being measured it was essential that these four fixtures should be equivalent in terms of bias and uncertainty. In this study the same engineer repeatedly measured each of the three parts in the same way using each of the four fixtures. With five measurements of each part on each fixture we end up with sixty measurements. These sixty measurements are organized into twelve subgroups of size five, and limits are computed for each fixture. The resulting average and range charts are shown in Figure 1. (The values shown have been coded for simplicity of presentation.)

	Fixture 1			Fixture 2			Fixture 3			Fixture 4		
Part	A	B	C	A	B	C	A	B	C	A	B	C
	69	46	57	68	49	59	75	55	67	77	56	69
	65	47	59	71	50	60	76	54	67	77	57	69
	70	48	59	69	48	61	75	56	68	76	56	69
	69	48	60	70	49	61	76	55	67	77	57	70
	69	49	60	71	50	61	75	55	67	78	57	69
Average	68.4	47.6	59.0	69.8	49.2	60.4	75.4	55.0	67.2	77.0	56.6	69.2
Range	5	3	3	3	2	2	1	2	1	2	1	1

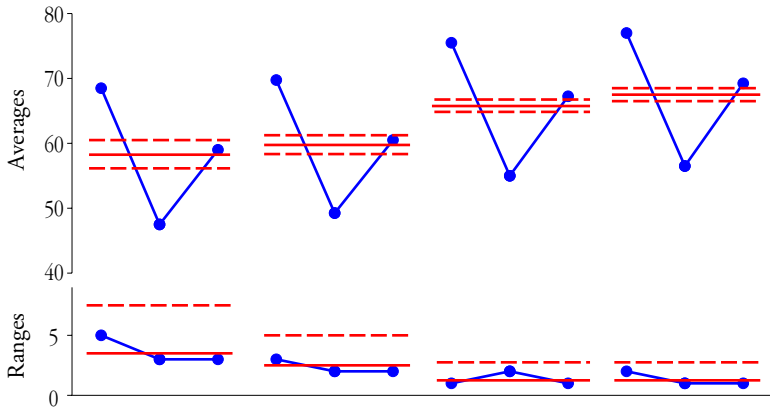

Figure 1: Average and Range Charts for the Four Fixtures

Figure 2 shows the grand averages, average ranges, and limits for each of these four charts. While we could have used ANOM for the initial analysis of these experimental data, in this case the average and range charts are roughly equivalent. The limits shown in Figures 1 and 2 were computed using the standard formulas for average and range charts.

The average and range chart in Figure 1 provides us with the ability to quickly understand the issues involved. We can check for internal measurement consistency and perform some simple eye tests for external consistency and bias between instruments.

	Grand Average	Average Range	Upper Average Limit	Lower Average Limit	Upper Range Limit
Fixture 1	58.33	3.67	60.45	56.22	7.8
Fixture 2	59.80	2.33	61.15	58.45	4.9
Fixture 3	65.87	1.33	66.64	65.10	2.8
Fixture 4	67.60	1.33	68.37	66.83	2.8

Figure 2: Chart Limits for Figure 1

THE "EYE TESTS"

Since no range chart in Figure 1 has a point above the upper range limit we may say that each of these fixtures displays internal consistency. However, the differences between the four range charts suggest that these four fixtures may have different levels of measurement error. So we will begin our follow-up analysis by considering if these differences are large enough to justify taking action.

If the instruments are measuring the parts the same way the graphs on the average charts should show reasonable parallelism. Here they do. All four fixtures agree that part A is high, part B is low, and Part C is in the middle. Any failure for the average charts to show the same patterns can be an indication of a serious problem where different instruments measure the same parts in different ways.

At the same time, while we can see that while the four fixtures show the same pattern, they also have different grand averages. These differences suggest possible bias effects between instruments, and later we will evaluate this to see if it is significant enough for action to be taken.

CHECKING INSTRUMENTS FOR DIFFERENCES IN MEASUREMENT ERROR

We will use the Analysis of Mean Ranges (ANOMR) to check for detectable differences in measurement error between the four fixtures. Here we want to compare the $m = 4$ average ranges. The original study had $k = 12$ subgroups of size $n = 5$. The average of all four average ranges will be our central line. This value is known as the overall average range, and in this case it is 2.167. From Table 2 at the end of this article, with an overall alpha level of 5 percent, and with $n = 5$, $k = 12$, and $m = 4$, our ANOMR scaling factors are 0.565 and 1.487. These values are multiplied by the overall average range to obtain the limits shown in Figure 3.

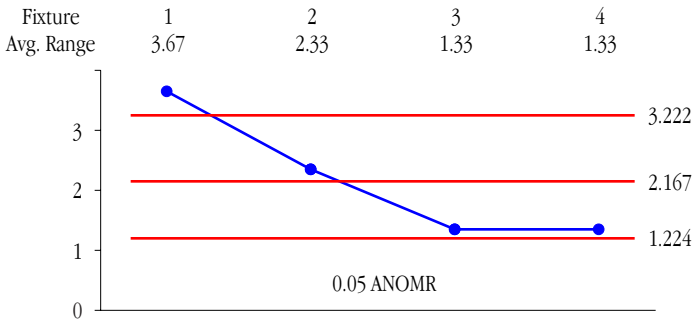

Figure 3: ANOMR for Comparing Average Ranges from Figure 1

Here we see that the average range for Fixture 1 is detectably greater than the overall average range. Thus, we can say that we may have two levels of measurement error present in these data. Now the question becomes one of whether this difference is of any practical importance.

THE PROBABLE ERROR OF THE MEASUREMENTS

One of the most important outputs of any measurement system evaluation is the Probable Error. The Probable Error (PE) is the 50th percentile for the distribution of measurement error and represents the effective *actual* resolution of the measurement system.

$$PE = 0.675 \frac{\bar{R}}{d_2}$$

so

$$PE \text{ for Fixture } 1 = 0.675 \frac{3.67}{2.326} = 1.06$$

This means that half the time a measurement with Fixture 1 will err by 1 unit or more, and half the time it will err by one unit or less. Since these measurements are recorded to the nearest whole number of units, they are recorded to the appropriate number of digits, and they are basically good to the last digit.

Since Figure 3 shows the average ranges for Fixtures 2, 3, and 4 to be indistinguishable from each other, we combine them when estimating their probable error. The average of these three average ranges is 1.67. Thus we estimate the probable error for Fixtures 2, 3, and 4 to be PE = 0.48 units. Since we want the measurement increment to fall somewhere in the range of 0.22 PE to 2.2 PE, and since these fixtures use a measurement increment of 1.0 unit, they also produce values that are good to the last recorded digit. So while Fixture 1 displayed more variation than the other fixtures, all of the fixtures produce values that are good to the last recorded digit, and there is no practical difference in measurement resolution between the fixtures.

However, the ANOMR in Figure 3 still tells us that Fixture 1 has detectably more measurement error than Fixtures 2, 3, and 4, and an effort to understand and remove the causes of excessive variation on Fixture 1 will likely be necessary.

Using the overall average range of 2.167 we might characterize these four fixtures as producing values with an overall probable error of 0.63 units. This is of importance since the measurement error will help us determine how to react to the suspected bias effects we check for below.

CHECKING INSTRUMENTS FOR BIAS EFFECTS

We use the Analysis of Main Effects (ANOME) to check for bias effects between the $m = 4$ fixtures. The original study consisted of $k = 12$ subgroups of size $n = 5$. The Overall Grand Average is 62.90. If there is no detectable bias between the four fixtures their grand averages should all fall within the ANOME limits. The overall average range is 2.167. From Table 1 at the end of this article the ANOME scaling factor with a five percent overall alpha-level is 0.244. Thus the ANOME detection limits are:

$$\begin{aligned} \textit{0.05 ANOME limits} &= \textit{Grand Average} \pm \textit{.244 * Grand Average Range} \\ &= \textit{62.90} \pm \textit{0.244 * 2.167} = \textit{62.37 and 63.43} \end{aligned}$$

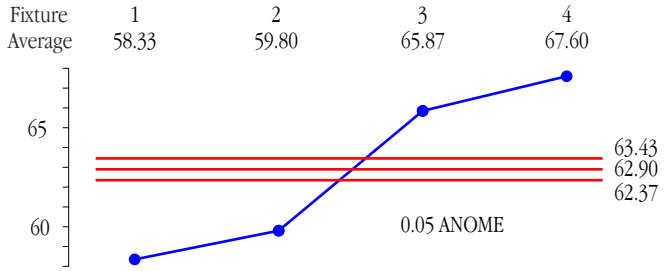

Figure 4: ANOME for Comparing Grand Averages from Figure 1

Figure 4 shows detectable biases between these four fixtures. By comparing the grand average for each fixture with the overall grand average of 62.90 we estimate these biases as –4.57, –3.10, 2.97, and 4.70 respectively.

In part one we suggested that biases that are smaller than 1.128 SD(E) or 1.67 PE will not have any impact in practice. Here our overall estimate of PE is 0.63 units. Thus, biases that are within 1.05 units of each other are unlikely to be important in practice. Since all four of the bias effects estimated above differ by more than 1.05 units, we have to say that all four fixtures are operating at different levels. (Remember, all of these data were obtained by the same engineer, testing the same parts, in the same way, using each of the four fixtures.)

If careful evaluation of the fixtures reveals no clear way to eliminate these biases, we may decide to adjust the measurements from each fixture to compensate for these biases. This approach was used in this case. The readings from Fixture 1 were adjusted by adding 5 units to each reading. The measurements from Fixture 2 had 3 units added. The measurements from Fixture 3 had 3 units subtracted, and those from Fixture 4 had 5 units subtracted. In this way all of the measurements are adjusted to be equivalent to each other. With this protocol in place a new study was conducted by the same engineer, using the same three production parts. The adjusted data are shown with the average and range chart in Figure 5.

Here our "eye test" reveals reasonable parallelism between the four average charts, comparable grand averages for each Fixture, much greater similarity for the average ranges for the four fixtures and internal consistency for each fixture. We confirm these initial impressions using ANOME and ANOMR charts.

	Fixture 1			Fixture 2			Fixture 3			Fixture 4		
Part	A	B	C	A	B	C	A	B	C	A	B	C
	73	52	62	72	53	63	73	52	65	74	53	65
	74	52	64	73	54	64	73	53	66	72	53	65
	74	52	65	72	53	64	73	53	65	71	52	64
	74	53	66	74	53	65	74	54	65	73	53	65
	75	54	65	75	54	65	74	53	66	74	54	65
Average	74.0	52.6	64.4	73.2	53.4	64.2	73.4	53.0	65.4	72.8	53.0	64.8
Range	2	2	4	3	1	2	1	2	1	3	2	1

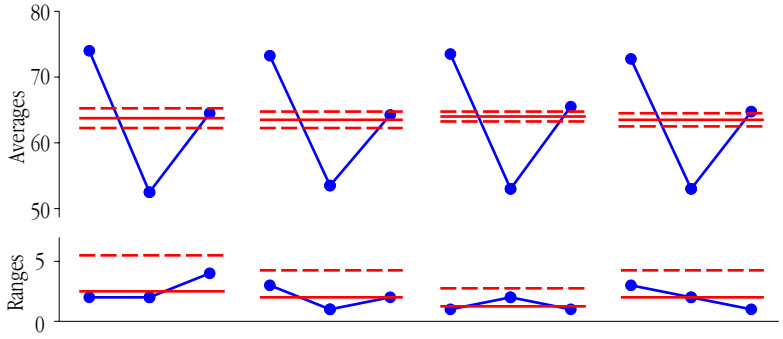

Figure 5: Average and Range Charts for New Data Following Adjustment to Remove Biases

	Grand Average	Average Range	Upper Average Limit	Lower Average Limit	Upper Range Limit
Fixture 1	63.67	2.67	65.21	62.13	5.6
Fixture 2	63.60	2.00	64.75	62.45	4.2
Fixture 3	63.93	1.33	64.70	63.16	2.8
Fixture 4	63.53	2.00	64.69	62.38	4.2

Figure 6: Chart Limits for Figure 5

The grand averages for the four fixtures are compared with an ANOME chart. With $k = 12$, $n = 5$, and $m = 4$ our five percent ANOME scaling factor from Table 1 is the same as before, 0.244, and our detection limits are:

*Post Adjust Confirmation Run 0.05 ANOME Limits = 63.68 ± 0 .244 * 2.00 = 63.19 to 64.17.*

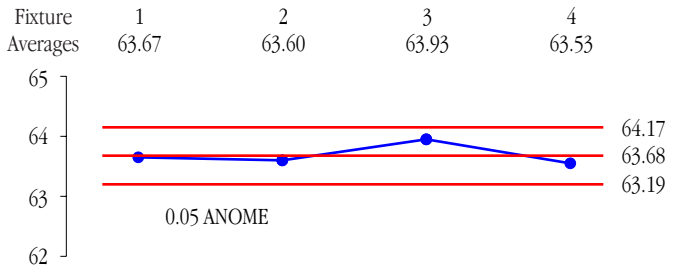

Figure 7: ANOME for Comparing Grand Average from Figure 5

With these adjustments to the readings we now have four fixtures with no detectable bias between them. But do the four fixtures have the same amount of measurement error? This

question is addressed by the ANOMR chart which compares average ranges.

As before we use the overall average range to construct the ANOMR chart. The overall average range is 2.00, and the five percent ANOMR scaling factors from Table 2 are 0.565 and 1.487, resulting in limits of 1.13 and 2.97.

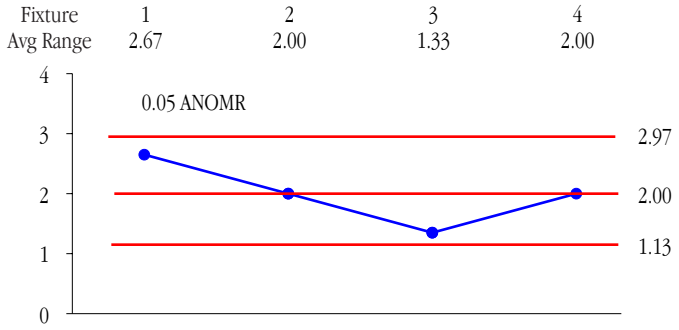

Figure 8: ANOMR for Comparing Average Ranges from Figure 5

With all of the average ranges within the limits this ANOMR chart shows no detectable difference in measurement error between the four fixtures.

But what happened to the excessive variation for Fixture 1 shown in the Initial Run? Thankfully, it was easy to keep a record of production data (computer controlled operation with lot based data files collected automatically) and it was a simple step to figure out that process variation on Fixture 1 appeared to diminish as each "inspection day" progressed. A quick study showed that there was a much longer "warm up" time needed for the laser micrometers used in Fixture 1 in comparison to the other fixtures (the method sheets for all 4 fixtures were revised to reflect this new information).

MANUFACTURING SPECIFICATIONS

Up to this point, we have avoided any reference to specification limits; but now it is finally necessary. Figure 9 shows an R Chart with all Confirmation Run data factored into the limits.

$$\textit{Upper Range Limit} \;=\; D_4 * \bar{R} \;=\; 2.114 * 2.00 \;=\; 4.23$$

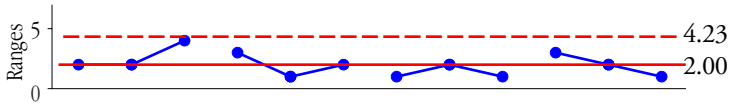

Figure 9: Range Chart from Figure 5 with Common Limits

With the ranges all falling within the common limits we can now calculate the overall probable error. From Figure 9:

$$SD(E) \;=\; \frac{\bar{R}}{d_2} \;=\; \frac{2.00}{2.326} \;=\; 0.86$$

and

$$PE \;=\; 0.675\ SD(E) \;=\; 0.58$$

We can use this value to establish "Manufacturing Specifications" (See "Where Do Manufacturing Specifications Come From", Wheeler, Quality Digest July 2010). 96% Manufacturing Specifications are found by tightening the watershed specifications by two Probable Errors. Here the specifications of 70 ± 10 result in watershed specifications of 59.5 and 80.5. Thus we find:

Lower Manufacturing Limit: 59.5 + 2*0.58 = 60.66 which becomes 61

Upper Manufacturing Limit: 80.5 – 2* 0.58 = 79.34 which becomes 79

When our observed measurement is either 61 or 79 there is at least a 96% chance that the product will actually be within the specifications of 60 to 80. (Values between 62 and 78 will be even more likely to be conforming.) Although these limits appear to be what would be a "good guess" under normal circumstances when setting manufacturing tolerance limits, this process leaves nothing to emotion or subjectivity. Instead, it sets the manufacturing limits based on a statistical foundation derived from measurement capability, the *actual* behavior of the data, risk assessment, and a balance of tolerance for marginal product combined with an understanding of how that product affects the performance of the system.

SUMMARY

One of the axioms of data analysis is that we have to detect a difference before we can estimate that difference, and only then can we assess the practical importance of that difference. Since we can detect very small differences when we use enough data, it is crucial that we know how to assess practical importance.

In Part One we demonstrated how measurement error dominates bias effects between instruments when those bias effects are smaller than 1.128 times the standard deviation of measurement error. Here the instruments begin to become equivalent. Depending upon the need for precision in a given application, we may choose to compensate for detectable biases by adjusting the values obtained from each instrument.

In Part Two we introduced the Analysis of Mean Moving Ranges (ANOMmR) as a way to test for differences in measurement error between instruments. Here we also further illustrated the adjustment for bias effects between instruments.

In Part Three we have looked at using multiple parts to compare instruments. When we have detectable bias effects, we should always estimate and evaluate the size of these effects relative to measurement error. Based on these three parts we propose the following guidelines for action:

1. When an instrument fails to display internal consistency it cannot be called a measurement device. When the same thing is measured repeatedly and ranges for those repeated measurements fall above the upper range limit we have evidence that the measurement operation is being carried out inconsistently. When this happens the reason for the inconsistency needs to be found and corrected.
2. If two instruments have detectably different levels of measurement error, and if either of the Probable Errors of these instruments is two or more times the size of the measurement increment, then the two instruments will need to be evaluated separately. As the Probable Error and Measurement Increment become similar in size the measurements begin to be good to the last recorded digit and measurement error

becomes less important in the measurement process.

3. When instruments display equivalent amounts of measurement error, they may be compared for detectable bias effects. When detectable bias effects are smaller than 1.67 Probable Errors measurement error will dominate the bias effects. Here the decision regarding the need to make adjustments for the bias effects will depend upon the precision needed in the application where the measurements are being used.

Tables for ANOME and ANOMR

The following tables are excerpts from more extensive tables given in ***Analyzing Experimental Data*** by Donald J. Wheeler and are used with permission.

Table 1: Analysis of Main Effects (ANOME)

Use the following scaling factors to test for
Bias Effects between m Instruments or m Operators
with an overall alpha level of five percent.

Detection Limits = *Grand Average* ± $ANOME_{.05}$ [*Average Range*]

$ANOME_{.05}$		$n = 2$	$n = 3$	$n = 4$	$n = 5$
$k = 4$	$m = 2$	0.832	0.385	0.261	0.203
$k = 6$	$m = 2$	0.609	0.299	0.207	0.161
	$m = 3$	1.083	0.519	0.355	0.275
$k = 8$	$m = 2$	0.502	0.254	0.175	0.137
	$m = 4$	1.159	0.568	0.391	0.305
$k = 9$	$m = 3$	0.813	0.408	0.283	0.220
$k = 10$	$m = 2$	0.436	0.223	0.156	0.122
	$m = 5$	1.204	0.599	0.415	0.324
$k = 12$	$m = 2$	0.388	0.202	0.141	0.111
	$m = 3$	0.675	0.346	0.241	0.189
	$m = 4$	0.884	0.450	0.313	0.244
	$m = 6$	1.230	0.621	0.431	0.338
$k = 14$	$m = 2$	0.356	0.185	0.130	0.102
	$m = 7$	1.254	0.638	0.445	0.348
$k = 15$	$m = 3$	0.589	0.306	0.214	0.168
	$m = 5$	0.926	0.477	0.332	0.260
$k = 16$	$m = 2$	0.329	0.172	0.121	0.096
	$m = 4$	0.739	0.382	0.269	0.209
	$m = 8$	1.273	0.653	0.455	0.357
$k = 18$	$m = 2$	0.307	0.162	0.115	0.090
	$m = 3$	0.530	0.276	0.194	0.153
	$m = 6$	0.957	0.495	0.346	0.271
	$m = 9$	1.290	0.663	0.463	0.363
$k = 20$	$m = 2$	0.290	0.153	0.109	0.085
	$m = 4$	0.648	0.339	0.237	0.187
	$m = 5$	0.779	0.405	0.285	0.224
	$m = 10$	1.303	0.672	0.471	0.368
$k = 21$	$m = 3$	0.486	0.255	0.179	0.141
	$m = 7$	0.982	0.509	0.358	0.280
$k = 24$	$m = 2$	0.261	0.139	0.099	0.076
	$m = 3$	0.449	0.237	0.167	0.130
	$m = 4$	0.584	0.307	0.217	0.170
	$m = 6$	0.810	0.423	0.298	0.233
	$m = 8$	0.999	0.521	0.367	0.287
	$m = 12$	1.326	0.688	0.484	0.379

Table 2: Analysis of Mean Ranges (ANOMR)
Use the following scaling factors to test for differences between the Average Ranges for *m* Instruments or *m* Operators with an overall alpha level of five percent.

Upper Detection Limit = $UMR_{.05}$ [*Overall Average Range*]
Lower Detection Limit = $LMR_{.05}$ [*Overall Average Range*]

$ANOMR_{.05}$		$n = 2$		$n = 3$		$n = 4$		$n = 5$	
		LMR	***UMR***	***LMR***	***UMR***	***LMR***	***UMR***	***LMR***	***UMR***
$k = 4$	$m = 2$	0.271	1.729	0.480	1.520	0.576	1.424	0.633	1.367
$k = 6$	$m = 2$	0.396	1.604	0.575	1.425	0.656	1.344	0.702	1.298
	$m = 3$	0.136	2.132	0.333	1.775	0.445	1.623	0.511	1.536
$k = 8$	$m = 2$	0.476	1.524	0.634	1.366	0.703	1.297	0.741	1.259
	$m = 4$	0.109	2.314	0.293	1.877	0.406	1.705	0.474	1.607
$k = 9$	$m = 3$	0.247	1.914	0.442	1.626	0.539	1.504	0.597	1.434
$k = 10$	$m = 2$	0.530	1.470	0.675	1.325	0.734	1.266	0.769	1.231
	$m = 5$	0.093	2.440	0.269	1.951	0.380	1.758	0.450	1.653
$k = 12$	$m = 2$	0.571	1.429	0.702	1.298	0.757	1.243	0.789	1.211
	$m = 3$	0.330	1.786	0.512	1.536	0.598	1.432	0.647	1.376
	$m = 4$	0.211	2.047	0.403	1.706	0.503	1.566	0.565	1.487
	$m = 6$	0.083	2.523	0.252	2.006	0.363	1.799	0.433	1.686
$k = 14$	$m = 2$	0.603	1.397	0.723	1.277	0.775	1.225	0.805	1.195
	$m = 7$	0.075	2.591	0.239	2.045	0.349	1.831	0.420	1.715
$k = 15$	$m = 3$	0.394	1.699	0.561	1.480	0.638	1.387	0.683	1.334
	$m = 5$	0.189	2.143	0.379	1.763	0.479	1.608	0.544	1.524
$k = 16$	$m = 2$	0.629	1.371	0.740	1.260	0.790	1.210	0.818	1.182
	$m = 4$	0.290	1.895	0.476	1.606	0.564	1.488	0.619	1.418
	$m = 8$	0.069	2.649	0.228	2.080	0.337	1.859	0.409	1.738
$k = 18$	$m = 2$	0.650	1.350	0.755	1.245	0.802	1.198	0.829	1.171
	$m = 3$	0.437	1.634	0.597	1.436	0.670	1.352	0.711	1.305
	$m = 6$	0.174	2.211	0.360	1.805	0.462	1.641	0.527	1.552
	$m = 9$	0.065	2.695	0.220	2.109	0.328	1.880	0.401	1.757
$k = 20$	$m = 2$	0.667	1.333	0.768	1.232	0.813	1.186	0.838	1.162
	$m = 4$	0.351	1.792	0.526	1.541	0.607	1.435	0.657	1.373
	$m = 5$	0.266	1.972	0.450	1.656	0.543	1.523	0.598	1.449
	$m = 10$	0.061	2.735	0.212	2.134	0.320	1.899	0.393	1.774
$k = 21$	$m = 3$	0.473	1.586	0.624	1.403	0.692	1.326	0.732	1.280
	$m = 7$	0.162	2.259	0.346	1.835	0.449	1.668	0.515	1.572
$k = 24$	$m = 2$	0.697	1.303	0.790	1.210	0.829	1.171	0.853	1.147
	$m = 3$	0.505	1.547	0.648	1.375	0.711	1.304	0.750	1.261
	$m = 4$	0.399	1.717	0.563	1.487	0.639	1.393	0.685	1.338
	$m = 6$	0.249	2.030	0.433	1.688	0.526	1.550	0.583	1.474
	$m = 8$	0.153	2.300	0.335	1.861	0.438	1.688	0.505	1.589
	$m = 12$	0.055	2.803	0.200	2.175	0.307	1.930	0.381	1.801