Power Comparisons of the Mann-Whitney *U* and Permutation Tests

Abstract:
Though the Mann-Whitney *U*-test and permutation tests are often used in cases where distribution assumptions for the two-sample *t*-test for equal means are not met, it is not widely understood how the powers of the two tests compare. Our goal was to discover under what circumstances the Mann-Whitney test has greater power than the permutation test. The tests' powers were compared under various conditions simulated from the Weibull distribution. Under most conditions, the permutation test provided greater power, especially with equal sample sizes and with unequal standard deviations. However, the Mann-Whitney test performed better with highly skewed data.

Background and Significance:

In many psychological, biological, and clinical trial settings, distributional differences among testing groups render parametric tests requiring normality, such as the z test and t test, unreliable. In these situations, nonparametric tests become necessary. Blair and Higgins (1980) illustrate the empirical invalidity of claims made in the mid-20th century that t and F tests used to detect differences in population means are highly insensitive to violations of distributional assumptions, and that non-parametric alternatives possess lower power. Through power testing, Blair and Higgins demonstrate that the Mann-Whitney test has much higher power relative to the t-test, particularly under small sample conditions. This seems to be true even when Welch's approximation and pooled variances are used to "account" for violated t-test assumptions (Glass et al. 1972).

With the proliferation of powerful computers, computationally intensive alternatives to the Mann-Whitney test have become possible. The computationally intensive permutation test offers a non-parametric solution to non-normality like the rank-based Mann-Whitney test.[1] At the price of computing power, permutation methods retain information on the magnitude and variability within data. This improved information preservation, coupled with the precision we might expect from a more computationally intensive method, suggests the permutation test may be a more powerful alternative to the Mann-Whitney test. In this paper, we present the results of power testing the permutation and Mann-Whitney methods under various sample and distribution conditions.

Power:

For our analysis, power is used to assess the performance of the Mann-Whitney and permutation methods. In statistical hypothesis testing, power is the probability that we will correctly reject the null hypothesis. In other words, when the null hypothesis is incorrect and should be rejected, power is the likelihood that we detect significant evidence and do indeed reject the null. Tests with high power are particularly important in the sort of early-stage research we are trying to simulate. Poor power can be the difference between advancing an important idea to the forefront of research and mistakenly filing the idea away to collect dust. It is essential that the methods that offer high power be identified and employed in scientific research.

Methods:

We compared the powers of the permutation test and Mann-Whitney test under various condition combinations using simulations.[2] For each comparison, a random sample was taken from two distributions under the specified conditions. The samples were compared using the permutation test with 10,000 iterations and the Mann-Whitney test.[3] This sampling and comparison procedure was then repeated 10,000 times for the combination of conditions, and the empirical power was calculated for both tests under the given combination of conditions.[4] Data was sampled from the two-parameter Weibull distribution.[5] The Weibull is particularly versatile as its shape and skewness can be modified with the Weibull shape

---

[1] Also referred to as the Mann-Whitney-Wilcoxon test and Wilcoxon rank-sum test, this test compares the rankings of the elements of both groups when sorted. For more information on this test, see S. Kuiper and J. Sklar's *Practicing Statistics* (2013).

[2] Refer to Appendix 1 for the R simulation code.

[3] Due to the time needed to run permutation tests, we limited each test to 10,000 iterations.

[4] To better understand consistency of each test, we ran every test three times and took the mean of the resulting powers.

[5] For information on the Weibull distribution, see R. Pruim's *Foundations and Applications of Statistics* (2011).

parameter (k), and its mean and variance can be modified with the Weibull scale parameter (γ) (Papulous and Pillai 2002). This property allows the Weibull to approximate other probability distributions.[6]

Conditions:

The following factor levels were chosen to simulate varying types of distributions that could arise in research situations:

- Shape parameters of 1, 2, and 3 were used to approximate highly-skewed, moderately-skewed, and symmetric distributions respectively.
- Sample size pairs of ($n_1 = 10$, $n_2 = 10$), ($n_1 = 10$, $n_2 = 30$), and ($n_1 = 30$, $n_2 = 30$) allow us to consider respectively the small, unbalanced, and large samples often found in early-stage research.
- Mean differences[7] of 1, 1.5, and 2 provide cases where power varies greatly with other conditions, allowing us to detect possible power differences between the two tests.
- Standard deviations of 1, 2, and 4 in the second distribution allow us to assess power in homoscedastic, borderline heteroscedastic, and highly heteroscedastic instances.[8]

We manipulated the Weibull scale parameter to produce the desired standard deviations in our Weibull distributions. The standard deviation of the underlying distribution of the first sample is always 1. The above factors motivate 3^4, or 81, possible conditions under which we compare the Mann-Whitney and permutation methods. Fig. 1 shows an example of two Weibull distributions we used in our simulation analysis.

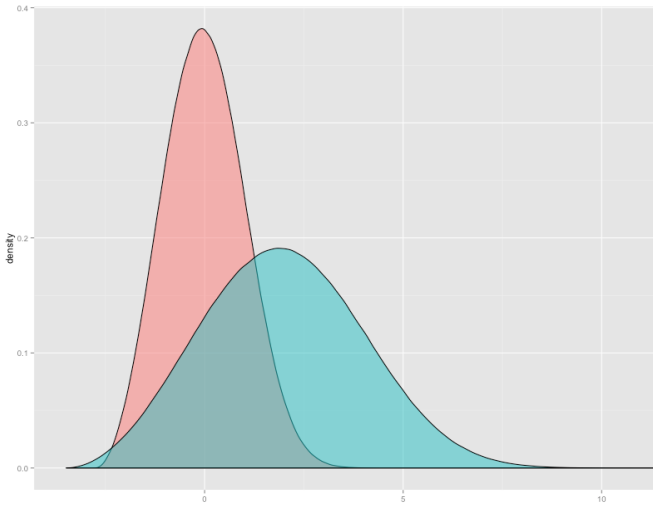


Figure 1. Probability density plots for two Weibull distributions being compared. Both distributions have shape parameter 3. The pink distribution has a standard deviation of 1 and the blue distribution has a standard deviation of 2. The difference in means between the two distributions is 2.

Results:

We calculated the empirical powers for each of three repetitions to ensure that any observed differences were consistent. As the maximum observed standard deviation between the three powers from repeated trials of the same condition was .0097, it is reasonable to believe that mean power differences greater than .02 are not due to chance.

For the distributions shown in Fig. 1, we see a sizable difference in the powers of the permutation test (0.7658) and Mann-Whitney test (0.6812).[9] We see even larger power differences whenever the two

[6] Particularly of interest was the Weibull distribution's ability to approximate the exponential and normal distributions. For more information on these distributions, see R. Pruim's *Foundations and Applications of Statistics* (2011).

[7] Here, defined as the difference between the mean of the two Weibull distributions. In these simulations, the distribution with the smaller standard deviation always had the lower mean.

[8] Homoscedasticity is often informally assumed when the maximum sample standard deviation is less than two times as large as minimum sample standard deviation.

[9] Refer to Appendix 3 for results from all 81 conditions.

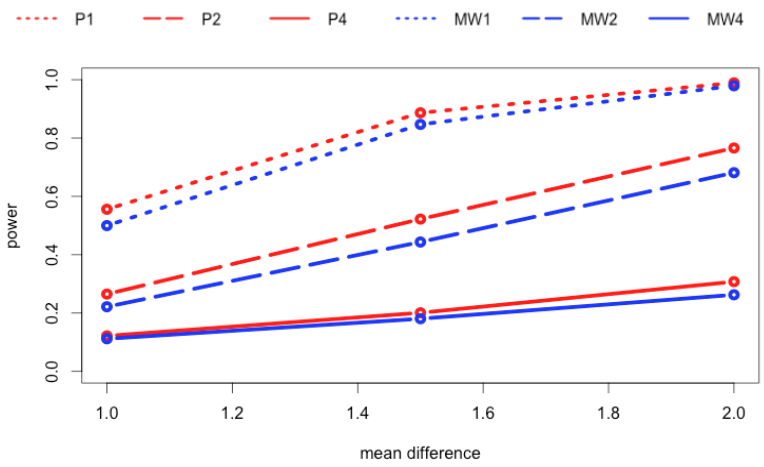


Figure 2. Power curves for Weibull distributions with shape parameter 3, sample sizes $n_1 = 10$ and $n_2 = 10$. Red curves are for the permutation test and blue curves are for the Mann-Whitney test. The numbers in the legend refer to the standard deviation of the second distribution (s.d. of the first distribution is always 1).

distributions are skewed, have a large mean difference, and exhibit extreme heteroscedasticity.

Fig. 2 shows the power curves for all distributions we tested with shape parameter 3 and sample sizes $n_1 = 10$ and $n_2 = 10$.[10] The points on the P2 and MW2 curves with mean difference 2 correspond to the scenario shown in Fig. 1. Notice that as heteroscedasticity increases, we see a greater difference between the powers of the two tests.

Discussion:

Under most conditions, the permutation test provided higher power. This trend holds in all relevant subsets of the data except when Weibull shape is 1, as we will discuss shortly. The permutation test had higher power when sample sizes were equal, and performed particularly well when both samples had size 30. Even when sample sizes were unequal, permutation tests outperformed Mann-Whitney tests, though by a smaller margin. The permutation test was also fairly robust to unequal standard deviations. Generally we find that when relative sample sizes differ or heteroscedasticity is present, the permutation method offers the better alternative.

The only instances in which the Mann-Whitney test performed better were under highly skewed distributions, i.e. when Weibull shape was 1. Blair and Higgins (1980) find a similar result when comparing the t-test and Mann-Whitney test. We attribute this outcome to the way we specified our permutation method to evaluate differences in means. Means are highly sensitive to skewed distributions as a few extreme values can have great weight on the average. A ranking system, as used in the Mann-Whitney test, does not account for the magnitude of data and is robust to such pull from extreme observations. Perhaps using permutation tests to evaluate medians instead of means would have avoided such bias. In further research, using a median-specified permutation test may allow the permutation test to perform better in all possible instances. With a mean specification, however, the Mann-Whitney should be used when data is highly skewed.

This paper provides a guideline for early stage research and other instances in which testing assumptions are frequently violated. As explicated in past work, when normality cannot be assumed, parametric methods fall drastically short of non-parametric tests. We show that in almost all instances, the mean-specified permutation test offers a much more powerful alternative to the Mann-Whitney test. Only in instances where, even after transformations, distributions are very highly skewed, should the Mann-Whitney test be used instead. Choosing the optimal test can maximize power, improving the likelihood that researchers correctly detect significant differences.

[10] Refer to Appendix 2 for the full set of power curves for each of the comparisons we conducted.

References:

Blair, C.R. and J.J. Higgins. 1980. A Comparison of the Power of Wilcoxon's Rank-Sum Statistic to that of Student's *t* Statistic Under Various Nonnormal Distributions. *Journal of Educational and Behavioral Statistics* 5(4): 309-335.

Glass, G.V., P.D. Peckham, and J.R. Sanders. 1972. Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance. *Review of Educational Research* 42(3): 237-288.

Papulous, A. and S.U. Pillai. 2002. Probability, Random Variables and Stochastic Processes 4th Edition. McGraw-Hill, Boston.

Appendices:

(1) R Code for Simulations:

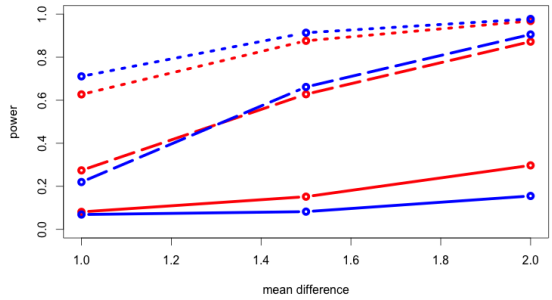
```
# Run permutation test
randtest <- function(x,y,fun=mean,reps) {
#n = sample size; m = no. in x
 n <- length(x)
 m <- length(y)
 data <- c(x,y)
# Store results in a numeric vector
 results <- numeric(reps)
#Run the permutation reps times
 for (i in 1:reps) {
   simtemp <- sample(data)
   results[i] <- fun(simtemp[1:n])-fun(simtemp[(n+1):(n+m)])
 }
#Find differences
 test.stat <- abs(fun(x)-fun(y))
 two.sided.p <- sum(abs(results)>=test.stat)/reps
 return(list(results=results,two.sided.p=two.sided.p,test.stat=test.st
at))
}

############
# Run simulations
simulation <- function(n1,n2, shape, scale, mean,
                       simreps, randreps) {
#Create vectors to store results
 resultsr <- numeric(simreps)
 resultsmw <- numeric(simreps)
#Run simpreps simulations
 for (i in 1:simreps) {
#Generate data from Weibull distributions
   x <- rweibull(n1,shape,scale)
   y <- rweibull(n2,shape,scale)+
     mean*weibullparinv(shape,scale)$sigma/sqrt(n2)
#Find p-values for both Mann-whitney and permutation tests
   resultsr[i] <- randtest(x,y,reps=randreps)$two.sided.p
   resultsmw[i] <- wilcox.test(x,y,alternative=
                                c("two.sided"))$p.value }
#Calculate Type II error
 rerror.p <- sum(resultsr >= 0.05)/simreps
 mwerror.p <- sum(resultsmw >= 0.05)/simreps
#Calculate power
rpower <- 1-rerror.p
 mwpower <- 1-mwerror.p
 return(list(rerror.p=rerror.p,mwerror.p=mwerror.p,resultsr=resultsr,r
esultsmw=resultsmw)) }
```

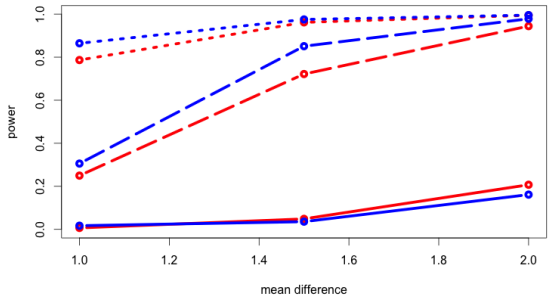
```
###############
# Simulation run 3 times
#Create vectors to store results
rpowers <- numeric(3)
mwpowers <- numeric(3)
#Run tests three times
for (i in 1:3) {
 # sim1 <- simulation(n1,n2, shape, variance, mean, simreps, randreps)
 sim1 <- simulation(10,10, 1, 1, 1.5, 10000, randreps=10000)
#Store powers from 3 tests
 rpowers[i] <- sim1$rpower
 mwpowers[i] <- sim1$mwpower
}
```

(2) Power Curves:

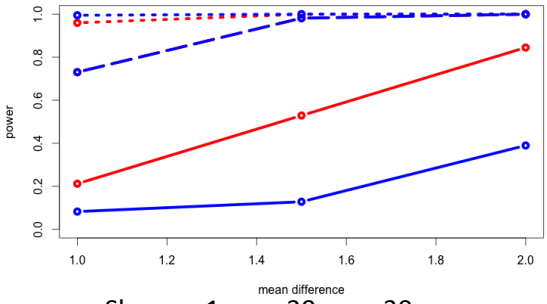
Figure 3. Power curves for Weibull distributions. For each combination of conditions, the simulation conducted 10,000 repetitions of the Mann-Whitney test and the permutation test for means, which had 10,000 iterations. We conducted three repetitions of this simulation. The points on the plots indicate the mean powers from these three repetitions. In the legend below, "P" refers to permutation test, "MW" refers to Mann-Whitney test, and the numbers refer to the standard deviation of the second distribution (the standard deviation of the first distribution is always 1).



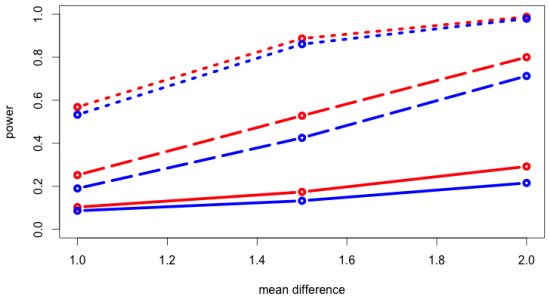
Shape = 1, n_1 = 10, n_2 = 10



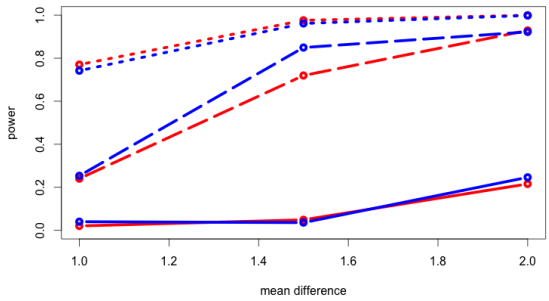
Shape = 1, n_1 = 10, n_2 = 30



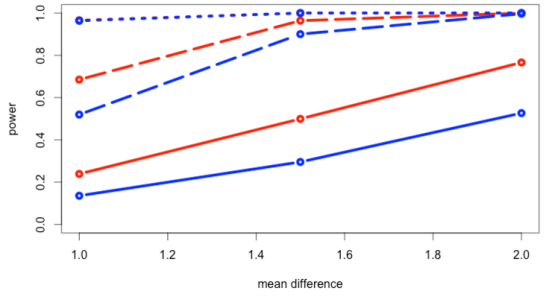
Shape = 1, n_1 = 30, n_2 = 30



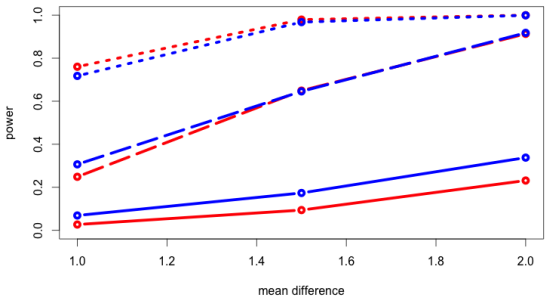
Shape = 2, n_1 = 10, n_2 = 10



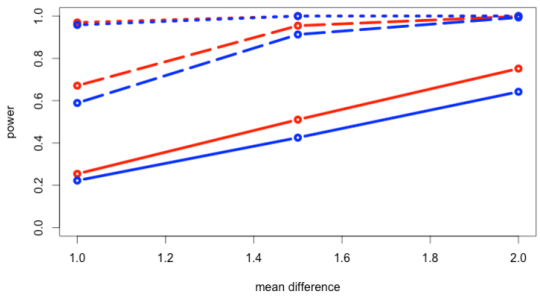
Shape = 2, n_1 = 10, n_2 = 30



Shape = 2, n_1 = 30, n_2 = 30



Shape = 3, n_1 = 10, n_2 = 30



Shape = 3, n_1 = 30, n_2 = 30

(3) Results:

Table 1. Permutation and Mann-Whitney test powers. For each combination of conditions, the simulation conducted 10,000 repetitions of the Mann-Whitney test and the permutation test for means, which had 10,000 iterations. We conducted three repetitions of this simulation. If the Mann-Whitney test outperformed the permutation test, the row is highlighted in grey.

	Distribution Shape	Difference in Means	Standard Deviation of 2nd Distribution	n1	n2	r.power1	r.power2	r.power3	mw.power1	mw.power2	mw.power3	Difference In Mean Powers
1	1	1	1	10	10	0.6244	0.6235	0.6331	0.7117	0.7080	0.7130	-0.0839
2	2	1	1	10	10	0.5665	0.5691	0.5696	0.5335	0.5320	0.5320	0.0359
3	3	1	1	10	10	0.5571	0.5578	0.5524	0.5005	0.5012	0.4976	0.0560
4	1	1	2	10	10	0.2750	0.2733	0.2738	0.2216	0.2191	0.2187	0.0542
5	2	1	2	10	10	0.2507	0.2525	0.2535	0.1885	0.1920	0.1900	0.0621
6	3	1	2	10	10	0.2662	0.2649	0.2641	0.2190	0.2237	0.2219	0.0435
7	1	1	4	10	10	0.0803	0.0799	0.0831	0.0703	0.0678	0.0685	0.0122
8	2	1	4	10	10	0.1031	0.1026	0.1041	0.0887	0.0843	0.0861	0.0169
9	3	1	4	10	10	0.1205	0.1221	0.1228	0.1082	0.1153	0.1120	0.0100
10	1	1.5	1	10	10	0.8716	0.8760	0.8810	0.9102	0.9154	0.9151	-0.0374
11	2	1.5	1	10	10	0.8866	0.8860	0.8875	0.8627	0.8572	0.8612	0.0263
12	3	1.5	1	10	10	0.8851	0.8862	0.8881	0.8463	0.8494	0.8443	0.0398
13	1	1.5	2	10	10	0.6226	0.6374	0.6236	0.6566	0.6644	0.6639	-0.0338
14	2	1.5	2	10	10	0.5264	0.5365	0.5215	0.4241	0.4337	0.4180	0.1029
15	3	1.5	2	10	10	0.5231	0.5243	0.5189	0.4435	0.4438	0.4432	0.0786
16	1	1.5	4	10	10	0.1529	0.1490	0.1535	0.0828	0.0774	0.0867	0.0695
17	2	1.5	4	10	10	0.1740	0.1682	0.1793	0.1318	0.1317	0.1348	0.0411
18	3	1.5	4	10	10	0.2086	0.1949	0.1991	0.1866	0.1788	0.1765	0.0202
19	1	2	1	10	10	0.9687	0.9712	0.9650	0.9777	0.9788	0.9762	-0.0093
20	2	2	1	10	10	0.9878	0.9879	0.9870	0.9788	0.9788	0.9791	0.0087
21	3	2	1	10	10	0.9891	0.9883	0.9900	0.9785	0.9800	0.9788	0.0100
22	1	2	2	10	10	0.8708	0.8759	0.8697	0.9035	0.9059	0.9073	-0.0334
23	2	2	2	10	10	0.8002	0.7975	0.8022	0.7153	0.7091	0.7136	0.0873
24	3	2	2	10	10	0.7663	0.7650	0.7661	0.6830	0.6796	0.6810	0.0846
25	1	2	4	10	10	0.2978	0.2975	0.2965	0.1562	0.1557	0.1535	0.1421
26	2	2	4	10	10	0.2891	0.2928	0.2948	0.2139	0.2152	0.2176	0.0767
27	3	2	4	10	10	0.3143	0.3081	0.3004	0.2690	0.2604	0.2586	0.0449
28	1	1	1	10	30	0.7874	0.7854	0.7872	0.8632	0.8627	0.8677	-0.0779
29	2	1	1	10	30	0.7648	0.7698	0.7759	0.7375	0.7431	0.7452	0.0282
30	3	1	1	10	30	0.7646	0.7543	0.7620	0.7209	0.7112	0.7207	0.0427
31	1	1	2	10	30	0.2429	0.2457	0.2604	0.3000	0.3002	0.3167	-0.0560
32	2	1	2	10	30	0.2419	0.2358	0.2430	0.2556	0.2469	0.2581	-0.0133
33	3	1	2	10	30	0.2548	0.2485	0.2431	0.3089	0.3040	0.3072	-0.0579
34	1	1	4	10	30	0.0069	0.0070	0.0060	0.0184	0.0166	0.0158	-0.0103
35	2	1	4	10	30	0.0222	0.0194	0.0199	0.0415	0.0386	0.0404	-0.0197
36	3	1	4	10	30	0.0255	0.0288	0.0268	0.0650	0.0734	0.0687	-0.0420
37	1	1.5	1	10	30	0.9623	0.9630	0.9603	0.9768	0.9749	0.9741	-0.0134
38	2	1.5	1	10	30	0.9764	0.9764	0.9754	0.9623	0.9605	0.9613	0.0147
39	3	1.5	1	10	30	0.9786	0.9811	0.9778	0.9675	0.9676	0.9677	0.0116
40	1	1.5	2	10	30	0.7291	0.7157	0.7202	0.8538	0.8493	0.8500	-0.1294
41	2	1.5	2	10	30	0.7210	0.7224	0.7149	0.8534	0.8491	0.8457	-0.1300
42	3	1.5	2	10	30	0.6542	0.6530	0.6403	0.6485	0.6526	0.6360	0.0035
43	1	1.5	4	10	30	0.6239	0.6396	0.6417	0.6623	0.6778	0.6757	-0.0369
44	2	1.5	4	10	30	0.0496	0.0477	0.0495	0.0358	0.0363	0.0359	0.0129
45	3	1.5	4	10	30	0.0934	0.0917	0.0975	0.1772	0.1718	0.1720	-0.0795
46	1	2	1	10	30	0.9942	0.9948	0.9946	0.9961	0.9951	0.9948	-0.0008
47	2	2	1	10	30	0.9996	0.9993	0.9994	0.9983	0.9978	0.9987	0.0012
48	3	2	1	10	30	0.9999	0.9996	0.9998	0.9994	0.9989	0.9991	0.0006
49	1	2	2	10	30	0.9445	0.9431	0.9453	0.9781	0.9796	0.9775	-0.0341
50	2	2	2	10	30	0.9297	0.9299	0.9266	0.9228	0.9226	0.9210	0.0066
51	3	2	2	10	30	0.9114	0.9137	0.9138	0.9144	0.9185	0.9203	-0.0048
52	1	2	4	10	30	0.2088	0.2066	0.2062	0.1618	0.1603	0.1621	0.0458
53	2	2	4	10	30	0.2164	0.2194	0.2115	0.2505	0.2480	0.2407	-0.0306
54	3	2	4	10	30	0.2272	0.2366	0.2306	0.3389	0.3424	0.3324	-0.1064
55	1	1	1	30	30	0.9622	0.9558	0.9606	0.9950	0.9944	0.9950	-0.0353
56	2	1	1	30	30	0.9649	0.9663	0.9665	0.9630	0.9629	0.9648	0.0023
57	3	1	1	30	30	0.9691	0.9699	0.9699	0.9587	0.9584	0.9564	0.0118
58	1	1	2	30	30	0.7272	0.7288	0.7350	0.6158	0.6180	0.6197	0.1125
59	2	1	2	30	30	0.6880	0.6813	0.6862	0.5222	0.5105	0.5269	0.1653
60	3	1	2	30	30	0.6783	0.6662	0.6683	0.5903	0.5887	0.5885	0.0818
61	1	1	4	30	30	0.2101	0.2105	0.2152	0.0783	0.0824	0.0870	0.1294
62	2	1	4	30	30	0.2372	0.2440	0.2364	0.1368	0.1370	0.1344	0.1031
63	3	1	4	30	30	0.2598	0.2454	0.2574	0.2269	0.2167	0.2240	0.0317
64	1	1.5	1	30	30	0.9995	0.9993	0.9988	1.0000	0.9999	1.0000	-0.0008
65	2	1.5	1	30	30	0.9999	0.9999	1.0000	0.9999	0.9997	0.9999	0.0001
66	3	1.5	1	30	30	1.0000	1.0000	1.0000	0.9999	0.9996	1.0000	0.0002
67	1	1.5	2	30	30	0.9812	0.9806	0.9820	0.9926	0.9925	0.9937	-0.0117
68	2	1.5	2	30	30	0.9632	0.9636	0.9651	0.9007	0.8961	0.9038	0.0638
69	3	1.5	2	30	30	0.9553	0.9533	0.9542	0.9131	0.9130	0.9121	0.0415
70	1	1.5	4	30	30	0.5234	0.5279	0.5355	0.1246	0.1282	0.1318	0.4007
71	2	1.5	4	30	30	0.4951	0.5073	0.4962	0.2900	0.3001	0.2973	0.2037
72	3	1.5	4	30	30	0.5143	0.5094	0.5076	0.4290	0.4219	0.4250	0.0851
73	1	2	1	30	30	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000
74	2	2	1	30	30	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000
75	3	2	1	30	30	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000
76	1	2	2	30	30	0.9994	0.9994	0.9997	1.0000	1.0000	1.0000	-0.0005
77	2	2	2	30	30	0.9995	0.9995	0.9997	0.9964	0.9971	0.9965	0.0029
78	3	2	2	30	30	0.9985	0.9980	0.9986	0.9936	0.9931	0.9936	0.0049
79	1	2	4	30	30	0.8482	0.8406	0.8451	0.3981	0.3873	0.3836	0.4550
80	2	2	4	30	30	0.7669	0.7643	0.7677	0.5320	0.5268	0.5216	0.2395
81	3	2	4	30	30	0.7488	0.7526	0.7539	0.6339	0.6423	0.6500	0.1097