STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:1487–1497
Published online 26 February 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3561

The Wilcoxon–Mann–Whitney test under scrutiny[‡]

Morten W. Fagerland[*,†] and Leiv Sandvik

Ullevål Department of Research Administration, Oslo University Hospital, Norway

SUMMARY

The Wilcoxon–Mann–Whitney (WMW) test is often used to compare the means or medians of two independent, possibly nonnormal distributions. For this problem, the true significance level of the large sample approximate version of the WMW test is known to be sensitive to differences in the shapes of the distributions. Based on a wide ranging simulation study, our paper shows that the problem of lack of robustness of this test is more serious than is thought to be the case. In particular, small differences in variances and moderate degrees of skewness can produce large deviations from the nominal type I error rate. This is further exacerbated when the two distributions have different degrees of skewness. Other rank-based methods like the Fligner–Policello (FP) test and the Brunner–Munzel (BM) test perform similarly, although the BM test is generally better. By considering the WMW test as a two-sample T test on ranks, we explain the results by noting some undesirable properties of the rank transformation. In practice, the ranked samples should be examined and found to sufficiently satisfy reasonable symmetry and variance homogeneity before the test results are interpreted. Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS: Wilcoxon–Mann–Whitney test; Fligner–Policello test; Brunner–Munzel test; rank transformation; robustness; heteroscedasticity

1. INTRODUCTION

The Wilcoxon–Mann–Whitney (WMW) test, also known as the rank sum test, is routinely used to compare the means or medians of two independent populations. In medical research, a common application is to compare treatment outcomes, or patient characteristics in nonrandomized groups, when the data are continuous and distributions are skewed. This usage of the WMW test is not in accordance with the original intentions, which is to test the null hypothesis that $P(X<Y)=0.5$, where X and Y are random samples from the two populations at interest. Nonetheless, interpretation

[*]Correspondence to: Morten W. Fagerland, Ullevål Department of Research Administration, Oslo University Hospital, Norway.
[†]E-mail: morten.fagerland@medisin.uio.no
[‡]Supporting information may be found in the online version of this article.

Received 5 June 2008
Accepted 17 January 2009

of p-values from the WMW test as evidence for or against a difference in means or medians is widespread. Although not correct in general, there are situations where $P(X<Y)=0.5$ translates to equality of medians, for example, when the two populations have equal shapes and equal scales. This is often referred to as the pure shift model. Under this model, the WMW test is type I error robust against nonnormality, and has high power compared with other tests, especially when distributions are highly skewed [1–4].

For practical applications, the pure shift model is not very realistic. If distributions are skewed and means are different, variances are most likely different too. This is exemplified by considering that the normal distribution is the only standard distribution for which the mean and variance are independent. The gamma distribution with parameters b and c has mean bc and variance b^2c [5]. Likewise, the lognormal distribution with parameters μ and σ has mean $e^{\mu+\sigma^2/2}$ and variance $e^{\sigma^2+2\mu}(e^{\sigma^2}-1)$ [5]. In both these examples, there is a relationship between the mean and the variance. Similar expressions can be found for other common distributions. Thus, if data follow some nonnormal distribution, a difference in means is likely to be accompanied by a difference in variances. The same is true if a difference in medians is observed [6]. Another situation in which the pure shift model is violated is when the two distributions have different degrees of skewness, which affects both type I error rates and power [7, 8].

Several studies have shown, or suggested, that the WMW test is sensitive to deviations from the pure shift model [2, 4, 6, 9–11]. True significance levels sometimes below the nominal level, sometimes far above the nominal level, have been observed. Still, recommendations vary. Some authors recommend that the WMW test should be used only when the variances are known to be equal [2, 12]. A variance ratio limit of 1.5 has been suggested [13]. Others are less specific [3, 9].

The WMW test is sometimes referred to as a test of medians, or a nonparametric version of the two-sample T test. Both these appellations are misleading. Conover and Iman [14] have shown that the WMW test is equivalent—in terms of reject versus not reject at a given significance level—to a T test performed after the scores have been replaced by ranks. It follows that the performance of the WMW test depends on the performance of the T test. More specifically, the shortcomings of the T test are inherited by the WMW test whenever the ranked data are subject to violations of the assumptions of normality and equal variances. Consequently, the properties of the rank transformation, as applied to samples from two possibly nonequal distributions, are the key to grasp the performance of the WMW test.

So far, studies have examined only a small number of different situations in which the pure shift model has been violated. The range of skewness values have been limited and sample sizes have usually been confined to small values (<50). The effects of variance heterogeneity have seldom been studied in conjunction with skewness, and seldom for small and realistic differences, for example, when the ratio of the standard deviations is in the range 1.1–1.5. Concerns about the consequences of using the WMW test under violations of the pure shift model have been raised, but as yet no comprehensive investigation to reveal the extent of this problem has been carried out.

The purpose of this paper is to study the performance of the WMW test for a wide range of skewed distributions that include small and common deviations from the pure shift model (Sections 2 and 3). For comparison, several other tests are included in this study. We aim to explain these results in light of the equivalence of the WMW test and the two-sample T test on ranks (Section 4). Finally, guidelines for proper use of the WMW test, and some concluding remarks, are presented (Section 5).

DOI: 10.1002/sim

2. NOTATION AND SIMULATION SETUP

Consider two populations, A and B, from which we have two samples, X of size m and Y of size n. Let R_X denote the sum of the ranks in sample X. For large samples, the test statistic

$$U = mn + m(m+1)/2 - R_X$$

has a normal distribution. Under the null hypothesis that $P(X<Y)=0.5$, the mean and variance of U are

$$E(U) = mn/2, \quad \mathrm{Var}(U) = mn(m+n+1)/12$$

In general, $P(X<Y)=0.5$ cannot be translated to equality of means or medians, but for some situations, for example, when the only difference in the distributions of A and B is a shift in location, it can. As discussed in the Introduction, departures from the pure shift model are common, and for many such situations, $P(X<Y)=0.5$ may not be fulfilled when means or medians are equal. In practice, the WMW test is often used to make inference on a difference in means or medians, and it is in this capacity we want to study the WMW test.

There is another version of the WMW test that uses the exact permutation distribution of ranks to compute p-values. For large samples, this test is computationally expensive, and software commonly uses this test only when samples are small. SPSS [15], for example, uses the exact WMW test when $mn \leqslant 400$ and $mn/2 + \min(m,n) \leqslant 220$. We will consider the exact test only briefly, and compare it with the approximate test for some small sample sizes. In the following, references to the WMW test are to the approximate version of the test.

Several modifications of the WMW test have been proposed. Most prominent of these are the Fligner–Policello (FP) test [16] and the Brunner–Munzel (BM) test [11]. The usual parametric alternatives are the two-sample T test and the modified two-sample T test for unequal variances (the Welch U test [17]). It has also been proposed to use the Welch U test on the ranks of the samples [18]. These five tests were included in the study.

We examined the true significance level of the six tests by computer simulations. The true significance level was estimated by observing the rejection rate for data sampled under two different null hypotheses:

(i) H_0: equal population means *versus* H_1: unequal population means.
(ii) H_0: equal population medians *versus* H_1: unequal population medians.

For symmetric distributions, (i) and (ii) are equivalent, but for asymmetric distributions, different results from using (i) and (ii) were expected.

The gamma distribution was used to produce skewed samples. This distribution is sufficiently flexible to allow distributions with a wide range of skewness values (β) to be generated. Also, it approximates distributions commonly encountered in medical research quite well [4].

The effects of unequal variances were studied by specifying the ratio (θ) of the standard deviation of population A to the standard deviation of population B. We used $\theta = 1.0, 1.1, 1.25, 1.5, 2.0$.

Five sample size combinations were selected: $(m,n) = (25,25)$, $(50,50)$, $(25,100)$, $(100,25)$, and $(100,100)$. For $m=100$, $n=25$, the largest sample was associated with the largest standard deviation. For $m=25$, $n=100$, the largest sample was associated with the smallest standard deviation.

DOI: 10.1002/sim

Table I. Summary of the simulation setup.

Tests	WMW: Wilcoxon–Mann–Whitney FP: Fligner–Policello BM: Brunner–Munzel T: two-sample T test U: Welch U (modified T test) RU: Welch U on ranked samples
Null hypotheses	equal means; equal medians
Nominal significance level	$\alpha = 0.05$
Sampling distribution	gamma*
Sample sizes (m, n)	(25, 25), (50, 50), (25, 100), (100, 25), (100, 100)
Standard deviation ratios	$\theta = 1.0, 1.1, 1.25, 1.5, 2.0$
Equal skewness values†	$\beta_A = \beta_B = 0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0$
Skewness distribution A‡	$\beta_A = 1.0, 1.5, 2.0, 2.5, 3.0$
Skewness distribution B‡	$\beta_B = 0.5, 1.0, 1.5, 2.0, 2.5$
Replications	100 000
Programming language	Matlab [19] with the Statistics Toolbox

*Normal distribution for $\beta = 0$.
†Distribution setup (i).
‡Distribution setup (ii).

For every sample size combination and every standard deviation ratio, two distribution setups were used:

(i) skewness of population A = skewness of population B.
(ii) skewness of population A = skewness of population $B + 0.5$.

In the latter case, the distribution with the greatest skewness had the largest standard deviation. The range of skewness values for both (i) and (ii), and a summary of the simulation setup, is presented in Table I.

We also carried out some additional simulations outside the full scope of the above setup. First, on the WMW test with one very large sample size combination ($m = n = 1000$). Second, the exact WMW test and the approximate WMW test were compared for three small sample size combinations: $(m, n) = (10, 10)$, $(10, 25)$, and $(25, 10)$.

3. RESULTS

We present the results of the simulation study in Web Tables 1–20 (Supplementary materials, available for downloading at the journal web site). In this section, we introduce the criteria for robustness, and illustrate and describe the results by giving excerpts from the Web tables. Not every detail of the results will be discussed here, so to see the full picture, inspection of the Web tables is necessary. The results from the additional simulations ($m + n < 50$ and $m = n = 1000$) are not presented in the Web tables, but will be discussed briefly below.

3.1. Robustness criteria

All the simulated significance levels were classified as either 10 per cent robust, 20 per cent robust, or nonrobust according to whether they were within 10 per cent, within 20 per cent, or

DOI: 10.1002/sim

beyond 20 per cent of the nominal significance level of $\alpha = 0.05$. In the Web tables, cells are colored green (10 per cent robust), yellow (20 per cent robust), or red (nonrobust). For black and white viewing, the 10 per cent robust cells have white text on a dark background, the 20 per cent robust cells have black text on a light gray background, and the nonrobust cells have black text on a dark gray background. A summary is given in Table II.

3.2. Distributions with equal skewness

3.2.1. The Wilcoxon–Mann–Whitney test. We illustrate the general results by considering the case of $m = 25$, $n = 100$ (Table III). For this sample size combination, the smallest sample has the largest standard deviation. The effect of increasing the ratio of the standard deviations is an increase in true significance levels. This effect is exacerbated by increasing skewness. There is no qualitative difference between the two null hypotheses, but in most cases, the null hypothesis of equal means produced larger significance levels that were further from the nominal level than the null hypothesis of equal medians. These general observations are valid for all the studied sample size combinations, except for situations where the largest sample had the largest standard deviation ($m = 100$, $n = 25$; Web Tables 7 and 8). In these latter cases, increasing the standard deviation ratio tend to decrease significance levels, while increasing skewness increased significance levels. For a small number of settings, these competing effects canceled each other out, resulting in true significance levels that were close to the nominal level. Occasionally, the two null hypotheses generated widely different results. For example, when $m = 100$, $n = 25$, $\theta = 1.25$, and $\beta_A = \beta_B = 2.5$, the true significance level was 13.4 per cent under the null hypothesis of equal means and 3.5 per cent under the null hypothesis of equal medians (Web Tables 7 and 8).

Table II. Robustness criteria and table colors.

Category	Significance level bounds	Color viewing	Black/white viewing
10 per cent robust	$4.5 \leqslant p \leqslant 5.5$	green	White text on dark background
20 per cent robust	$4.0 \leqslant p \leqslant 6.0$	yellow	Black text on light gray background
nonrobust	$p < 4.0$ or $p > 6.0$	red	Black text on dark gray background

p is the true significance level.

Table III. True significance levels (p) of the WMW test for a nominal significance level of 5 per cent.

H_0: equal means				$m = 25$, $n = 100$				H_0: equal medians						
							Std. ratio							
10.7	12.1	16.4	24.2	35.4	48.9	62.6	2.00	10.7	10.9	11.8	13.2	15.0	18.4	22.4
8.1	8.9	11.0	15.7	23.8	34.9	48.9	1.50	8.1	8.5	9.1	10.2	12.1	15.4	19.6
6.7	6.9	7.8	9.9	13.6	21.5	32.6	1.25	6.7	7.0	7.1	7.6	9.1	11.8	16.4
5.6	5.7	5.8	6.4	7.7	10.8	16.7	1.10	5.6	5.7	5.8	6.2	6.8	8.7	12.5
4.8	4.8	4.7	4.8	4.9	4.9	5.0	1.00	4.8	4.8	5.1	5.0	5.2	5.9	8.7
0.0	0.5	1.0	1.5	2.0	2.5	3.0	Skewness	0.0	0.5	1.0	1.5	2.0	2.5	3.0

Data from normal distributions (skewness = 0) and gamma distributions (skewness > 0).

DOI: 10.1002/sim

An interesting result was that the ability of the WMW test to maintain the nominal significance level decreased when the total sample size increased. This effect was quite large, and illustrates that the WMW test does not share the asymptotic robustness of the two-sample T test.

If the results are subjected to the robustness criteria defined at the start of this section, 36 per cent of the simulated significance levels are 10 per cent robust, 11 per cent are 20 per cent robust, and 53 per cent are nonrobust. Alarmingly, only a 10 per cent difference in standard deviations combined with moderate to large degrees of skewness led to nonrobust significance levels for most settings. As noted above, this nonrobustness is especially severe for large sample sizes. Additional simulations (not shown) with a sample size of $m=n=1000$ produced true significance levels of 99 per cent under the null hypothesis of equal means and 40 per cent under the null hypothesis of equal medians when $\theta=1.1$, and $\beta_A=\beta_B=3.0$.

If the purpose is to compare medians, the significance level of the WMW test may be unacceptable even when standard deviations are equal. This is exemplified in Table III where the true significance level is inflated for large degrees of skewness.

3.2.2. Comparison with the other tests. For equal sample sizes, the performance of the four rank-based tests are qualitatively equal (Web Tables 1–4 and 9–10). The WMW test and the Welch U test on ranks (RU) have about the same true significance levels. The FP test is slightly better than the WMW test, and the BM test is slightly better than the FP test. Under the null hypothesis of equal means, the two-sample T test (T) and the Welch U test (U) are superior to the rank-based tests with significance levels close to the nominal level for almost all the simulated settings. Under the null hypothesis of equal medians, the parametric tests are slightly better when $m=n=25$, but slightly worse when $m=n=100$.

When $m=25$, $n=100$ (Web Tables 5 and 6), the FP test was better than the WMW test for $\theta\geqslant 1.25$, but worse otherwise. The BM and RU tests were markedly better than the WMW and the FP tests, and maintained the nominal significance level for small amounts of skewness. The T test behaved similarly to the WMW and FP tests. The U test was slightly better than the BM and RU tests when $\beta\leqslant 1.0$ for both null hypotheses, and for some combinations of θ and $\beta>1.0$ values under the null hypothesis of equal medians.

For the sample size combination $m=100$, $n=25$, the BM and RU tests, and sometimes the U test, were far better than the WMW test (Web Tables 7 and 8). The T test maintained the significance level only when the standard deviations were equal. For $\theta>1.0$, the T test was conservative with significance levels below the nominal level.

3.2.3. The exact WMW test versus the approximate WMW test. The exact and the approximate WMW tests were compared for the sample size combinations $(m,n)=(10,10)$, $(10,25)$, and $(25,10)$. Only small differences between the two tests were detected. For both tests and all three sample sizes, the patterns of results were very similar to what was observed for the approximate WMW test for the sample sizes $(m,n)=(25,25)$, $(25,100)$, and $(100,25)$. Two differences are worth mentioning: (i) both tests were conservative with true significance levels in the range 4.1–4.5 per cent when $m=n=10$ and $\theta=1.0$; (ii) the true significance levels for combinations of large standard deviation ratios and great amounts of skewness were slightly closer to the nominal level than for the larger sample size combinations. This last observation is in agreement with our previous finding that the ability of the WMW test to maintain the nominal significance level seems to decrease with increasing sample size.

DOI: 10.1002/sim

3.3. Distributions with different degrees of skewness

3.3.1. The WMW test. As could be expected, when sampling from distributions in which skewness differed, true significance levels deviated severely from the nominal level. Table IV presents the case of $m=100$, $n=25$. From this example, we can see that the range of true significance levels is large, and that the WMW test can be conservative and liberal under quite similar settings. For both null hypotheses, the true significance level increases with increasing skewness. This is true for most of the sample size combinations, but for small standard deviation ratios ($\theta \leqslant 1.5$), the opposite effect is sometimes seen, for example, when $m=25$, $n=100$ (Web Tables 15 and 16). Table IV illustrates the varying effects of increasing the standard deviation ratio. Under the null hypothesis of equal medians, the true significance level decreases, but under the null hypothesis of equal means, the true significance level sometimes increases, sometimes decreases. It is typical for the unequal skewness case that the true significance level is often closer to the nominal level when the standard deviations are slightly different ($\theta=1.1$ or 1.25) than when the standard deviations are equal ($\theta=1.0$).

The decreasing robustness of the WMW test with increasing sample size is also observed when the degrees of skewness are different. When $m=n=1000$, true significance levels range from 5.2 to 100.0 per cent, and the number of true significance levels that are within 20 per cent of the nominal level is three out of 50 (results not shown). Considering the sample size combinations defined in Table I, 26 per cent of the simulated significance levels are 10 per cent robust, 13 per cent are 20 per cent robust, and 61 per cent are nonrobust.

In general, it is difficult to predict the performance of the WMW test in a given situation. There are a number of competing factors that influence the direction and magnitude of the true significance level: the degrees of skewness, the standard deviation ratio, the total sample size, and the sample size ratio.

3.3.2. Comparison with the other tests. When the sample sizes are equal (Web Tables 11–14 and 19–20), the T and U tests have true significance levels very close to the nominal level under the null hypothesis of equal means. In this situation, the two parametric tests are clearly superior to the rank-based tests. When the distributions are aligned by their medians, the rank-based tests are

Table IV. True significance levels (p) of the WMW test for a nominal significance level of 5 per cent.

H_0: equal means					$m=100$, $n=25$	H_0: equal medians				
					Std. ratio					
5.9	12.6	25.0	45.1	67.7	2.00	1.7	1.7	1.9	2.0	2.3
5.0	7.8	12.9	22.7	39.5	1.50	2.8	3.0	3.2	4.1	7.2
4.9	5.8	7.2	9.3	14.8	1.25	4.0	4.5	5.4	8.2	13.6
5.1	5.4	5.7	5.7	6.1	1.10	5.2	5.9	7.7	11.1	16.8
5.6	5.5	6.1	7.5	10.4	1.00	6.0	7.1	9.1	13.3	18.8
1.0	1.5	2.0	2.5	3.0	Skewness population *A*	1.0	1.5	2.0	2.5	3.0
0.5	1.0	1.5	2.0	2.5	Skewness population *B*	0.5	1.0	1.5	2.0	2.5

Data from normal distributions (skewness$=0$) and gamma distributions (skewness>0).

DOI: 10.1002/sim

better, except for some small sample size settings ($m=n=25$). As in the equal skewness case, the four rank-based tests perform very similarly, with the BM test somewhat better.

For a sample size of $m=25$, $n=100$, none of the six tests perform well under the null hypothesis of equal means (Web Table 15). The situation is better under the null hypothesis of equal medians (Web Table 16) where the BM, U, and RU tests improve upon the WMW test in some settings, but not in all.

When the largest sample has the largest standard deviation ($m=100$, $n=25$; Web Tables 17 and 18), the results are difficult to categorize and describe. All tests perform well in at least a few number of settings and very poorly in several others. Under the null hypothesis of equal means, the U test is clearly superior, while under the null hypothesis of equal medians, the BM and T tests performed best, although under different combinations of skewnesses and standard deviation ratios.

4. RANK TRANSFORMATIONS

In this section, we explore the implications of Conover and Iman's result that the WMW test is equivalent to the two-sample T test on ranks [14]. The authors show that the two-sample T statistic performed after the original scores have been replaced by ranks is a monotonically increasing function of the standardized WMW test statistic. The WMW test and the T test on ranks are therefore α-equivalent, meaning that for a given significance level α, the tests will always agree on whether to reject or not reject the null hypothesis.

The T test has been studied extensively, and is known to be nonrobust under violations of normality and variance homogeneity [1, 4, 9, 20, 21]. Consequently, the severe nonrobustness of the WMW test can be explained by considering the properties of the rank transformation.

There are several ways to apply a rank transformation to two-sample data, but the most common method, and the one that is used to derive the WMW test statistic, is to replace the lowest score by one, the next lowest score by two, and so on until the highest score is replaced by $m+n$. In theory, there are several benefits of the rank transformation. Outliers are pulled toward the center of the samples, skewness and variance are reduced, and the ordering of the original samples is maintained. However, unless the two samples are identically distributed, the moments of the two samples can be effected quite differently, and the exact properties of the ranked samples are difficult to predict. Differences in means, variances, or skewnesses may be amplified or even introduced where no differences existed in the original samples.

An example will illustrate this effect. Consider two samples, X and Y, and their first three moments:

Sample	Mean	Std	Skewness
$X=1, 10, 100$	37	54.7	0.69
$Y=20, 24, 67$	37	26.1	0.69

The means and skewnesses are equal, but the standard deviation of the X's is twice the standard deviation of the Y's. We denote the rank transformed samples by RX and RY, and calculate

DOI: 10.1002/sim

their moments:

Sample	Mean	Std	Skewness
$RX = 1, 2, 6$	3	2.65	0.60
$RY = 3, 4, 5$	4	1.00	0.00

Clearly, the rank transformation has not succeeded in reducing variance hetereogeneity. A difference in means has been introduced, and skewness has been eliminated in one sample, but only slightly reduced in the other. The two ranked samples differ in all three moments, whereas the original samples differed only in standard deviations. This undesirable effect of the rank transformation is not likely to occur for every data set, or even for most data sets, but our simulation study suggests that it is quite common, and can be difficult to predict.

With standard software, the WMW test is performed without giving any consideration to the effects of the rank transformation. The transformation is done as an intermediate step, and the properties of the ranked samples are hidden from the user. We recommend that a rank transformation is performed independently of the WMW test, and that the properties of the ranked samples are carefully considered. Particular attention should be paid to the standard deviations and skewnesses of the ranked samples. If either of these two moments differ considerably, the transformation may have done more harm than good, and p-values obtained with the WMW test can be grossly inaccurate.

5. DISCUSSION

As a test of means or medians, the Wilcoxon–Mann–Whitney (WMW) test can be severely nonrobust for deviations from the pure shift model. Our simulation study demonstrates that this problem is more serious than previously thought. We show that a variety of minor deviations from this model can lead to true significance levels that are alarmingly far from the nominal level. We have argued that when distributions are skewed, differences in means are likely accompanied by differences in variances. The assumption of the pure shift model is thus often unrealistic. In medical research, skewed data are common [3], and it is precisely in these situations the WMW test is most frequently used.

It has been suggested that comparing medians results in slightly better robustness than comparing means [4]. This is indeed the case for many situations, but there are also situations where the opposite is true, for example, when variances are equal, skewness is severe, and sample sizes are unequal. Our study shows that the problem of lack of robustness of the WMW test is not resolved by considering the WMW test as a test of medians. This description of the test is, at best, inaccurate.

The WMW test was designed to test the null hypothesis that $P(X<Y) = 0.5$. Sometimes, this can be interpreted as a test of means or medians—for example, under the pure shift model—but this is not true in general. Unless the shapes and scales of the population distributions are well known, it is difficult to determine whether such an interpretation is appropriate. As our simulation study shows, there are many situations in which the WMW test has a very high probability of rejection when either the means or medians are equal. It is not straightforward to interpret a significant p-value in such situations, apart from concluding that the two populations differ in some way.

DOI: 10.1002/sim

One might argue that the WMW test can be used to test the null hypothesis that two populations are identical, and that a medical researcher would be equally interested in other differences between treatments than a shift in location. We agree that the WMW test is better suited to this type of inquiry, but our study shows that there are several situations in which the test has very low power to detect variance or skewness heterogeneity, for example, in unbalanced designs where the largest sample has the largest variance.

Our critique applies to the large sample approximate version of the WMW test. However, in a comparison of the exact with the approximate WMW test for some sample size combinations in the range $20 \leqslant m+n \leqslant 35$, we found little to distinguish the two tests. The patterns of results were largely the same as for the approximate test on the larger sample size combinations ($m+n \geqslant 50$).

Zimmerman [22] has pointed out that the rank transformation reduces variance heterogeneity, but does not eliminate it. This observation is then used to explain the sensitivity of the WMW test to unequal variances. In our study, this line of approach has been further explored. We have shown that the rank transformation can alter means, standard deviations, and skewnesses of the two samples differently. The only situation in which the rank transformation is guaranteed to achieve a beneficial effect is when distributions are identical and sample sizes are equal. For deviations from these rather strict assumptions, the effects of the rank transformation on sample moments are unpredictable.

In our simulation study, the WMW test was compared with the Fligner–Policello test (FP), the Brunner–Munzel test (BM), the two-sample T test (T), the Welch U test (U), and the Welch U test on ranks (RU). The four rank-based tests (WMW, FP, BM, and RU) performed similarly, although the BM test was frequently a little better than the others. When the sample sizes were equal, the parametric tests (T and U) were superior to the rank-based tests under the null hypothesis of equal means, but not under the null hypothesis of equal medians. When the sample sizes were unequal, the BM, RU, and U tests performed best. For several settings, small changes in population properties led to large alterations in the performance of the tests. We recommend that Web Tables 1–20 are consulted for more detailed recommendations.

In summary, the large sample approximate WMW test can be a poor method for comparing the means or medians of two populations, unless the two distributions have equal shapes and equal scales. This problem also seems to apply in various degrees to the exact WMW test, the FP test, the BM test, and the Welch U test on ranks. When using the WMW test, we recommend that the properties of the ranked samples are thoroughly investigated for signs of skewness and variance heterogeneity. If such effects are present, the Welch U test is preferable in many situations.

REFERENCES

1. Sawilowsky SS, Blair RC. A more realistic look at the robustness and type II error properties of the t test to departures from population normality. *Psychological Bulletin* 1992; **111**(2):352–360.
2. Penfield DA. Choosing a two-sample location test. *Journal of Experimental Education* 1994; **62**(4):343–360.
3. Bridge PD, Sawilowsky SS. Increasing physicians' awareness of the impact of statistics on research outcomes: comparative power of the t-test and Wilcoxon rank-sum test in small samples applied research. *Journal of Clinical Epidemiology* 1999; **52**(3):229–235.
4. Skovlund E, Fenstad GU. Should we always choose a nonparametric test when comparing two apparently nonnormal distributions? *Journal of Clinical Epidemiology* 2001; **54**:86–92.
5. Evans M, Hastings N, Peacock B. *Statistical Distributions* (3rd edn). Wiley Series in Probability and Statistics. Wiley: New York, 2000.
6. Hart A. Mann–Whitney test is not just a test of medians: differences in spread can be important. *British Medical Journal* 2001; **323**:391–393.

DOI: 10.1002/sim

7. Wilcox RR, Keselman HJ. Modern robust data analysis methods: measures of central tendency. *Psychological Methods* 2003; **8**(3):254–274.
8. Wilcox RR. *Introduction to Robust Estimation and Hypothesis Testing* (2nd edn). Academic Press: San Diego, CA, 2005.
9. Stonehouse JM, Forrester GJ. Robustness of the *t* and *U* tests under combined assumption violations. *Journal of Applied Statistics* 1998; **25**(1):63–74.
10. Zimmerman DW. Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education* 1998; **67**(1):55–68.
11. Brunner E, Munzel U. The nonparametric Behrens–Fisher problem: asymptotic theory and a small-sample approximation. *Biometrical Journal* 2000; **42**(1):17–25.
12. Moser BK, Stevens GR, Watts CL. The two-sample *T* test versus Satterthwaite's approximate *F* test. *Communications in Statistics—Theory and Methods* 1989; **18**(11):3963–3975.
13. Zimmerman DW. Failure of the Mann–Whitney test: a note on the simulation study of Gibbons and Chakraborti (1991). *Journal of Experimental Education* 1992; **60**(4):359–364.
14. Conover WJ, Iman RL. Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician* 1981; **35**(3):124–129.
15. SPSS 15.0. SPSS, Inc., Chicago, IL, 2006.
16. Fligner MA, Policello II GE. Robust rank procedures for the Behrens–Fisher problem. *Journal of the American Statistical Association* 1981; **76**(373):162–168.
17. Welch BL. The significance of the difference between two means when the population variances are unequal. *Biometrika* 1937; **29**:350–362.
18. Zimmerman DW, Zumbo BD. Rank transformations and the power of the student *t* test and Welch t' test for non-normal populations with unequal variances. *Canadian Journal of Experimental Psychology* 1993; **47**(3): 523–539.
19. Matlab 7. The MathWorks, Inc., Natick, MA, 2005.
20. Gans DJ. Use of a preliminary test in comparing two sample means. *Communications in Statistics—Simulation and Computation* 1981; **B10**(2):163–174.
21. Posten HO, Yeh HC, Owen DB. Robustness of the two-sample *t*-test under violations of the homogeneity of variance assumption. *Communications in Statistics—Theory and Methods* 1982; **11**(2):109–126.
22. Zimmerman DW. A note on homogeneity of variance of scores and ranks. *Journal of Experimental Education* 1996; **64**(4):351–362.

DOI: 10.1002/sim