General Article

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Advances in Methods and Practices in Psychological Science
April-June 2021, Vol. 4, No. 2, pp. 1–14

DOI: 10.1177/2515245921999602
www.psychologicalscience.org/AMPPS
SAGE

Psychologists Should Use Brunner-Munzel's Instead of Mann-Whitney's *U* Test as the Default Nonparametric Procedure

Julian D. Karch
Methodology and Statistics Department, Institute of Psychology, Leiden University

Abstract

To investigate whether a variable tends to be larger in one population than in another, the *t* test is the standard procedure. In some situations, the parametric *t* test is inappropriate, and a nonparametric procedure should be used instead. The default nonparametric procedure is Mann-Whitney's *U* test. Despite being a nonparametric test, Mann-Whitney's test is associated with a strong assumption, known as *exchangeability*. I demonstrate that if exchangeability is violated, Mann-Whitney's test can lead to wrong statistical inferences even for large samples. In addition, I argue that in psychology, exchangeability is typically not met. As a remedy, I introduce Brunner-Munzel's test and demonstrate that it provides good Type I error rate control even if exchangeability is not met and that it has similar power as Mann-Whitney's test. Consequently, I recommend using Brunner-Munzel's test by default. To facilitate this, I provide advice on how to perform and report on Brunner-Munzel's test.

Keywords

hypothesis testing, assumptions, nonparametric, Type I error, power, Mann-Whitney *U* test, Wilcoxon rank-sum test, open materials

Received 4/17/20; Revision accepted 1/29/21

In psychology, a common task is to investigate whether a variable tends to be larger in one population than in another. The most used statistical technique for this research question is the parametric *t* test. Although the recommended Welch version of the *t* test (Delacre et al., 2017) assumes normality, it is asymptotically robust[1] to violations of this assumption, that is, in large samples, it typically has the correct Type I error rate and high power even if normality is violated.

However, in some situations, applying the *t* test is inappropriate even in large samples. In particular, the *t* test is often inappropriate for ordinal data because it is, for example, sensitive to the choice of the coding scheme. Consider the following example. A research group is interested in comparing their new depression therapy approach with a control group. As outcome variable, the depression score after 12 months of (mock) therapy is measured as an ordinal variable with the levels *no improvement*, *slight improvement*, *substantial improvement*, and *symptom-free*, which are coded as 1, 2, 3, and 4, respectively. For the control and the treatment groups, 10,000 participants[2] each are available, and the relative frequencies are as displayed in Table 1. Using the recommended Welch version of the *t* test (Delacre et al., 2017) leads to the following results: $t(19{,}778.84) = 0.00$, $p > .999$. Consequently, one would act as if the new therapy method does not work. However, when coding no improvement as 0, such that the new coding scheme is 0, 2, 3, and 4, the results of Welch's *t* test are $t(19{,}722.70) = -2.78$, $p = .005$, with a higher average level of improvement for the treatment ($M = 2.35$) compared with the control group ($M = 2.30$). Consequently, one would act as if the new therapy approach works. Thus, the results of the statistical analysis, and

Corresponding Author:
Julian D. Karch, Methodology and Statistics Department, Institute of Psychology, Leiden University
E-mail: j.d.karch@fsw.leidenuniv.nl

Table 1. Relative Frequencies of the Level of Improvement for the Control Group and the Treatment Group

	No improvement	Slight improvement	Substantial improvement	Symptom-free
Control	20%	30%	30%	20%
Treatment	15%	35%	35%	15%

thus the fate of the new therapy approach, depend on the more or less arbitrary coding scheme.

A test that is insensitive to the coding scheme and is thus more appropriate for ordinal data is the nonparametric Mann-Whitney *U* test. For the example, the result for the Mann-Whitney *U* test is $U = 50{,}000{,}000,\ p > .999$. for both coding schemes.

Unfortunately, the means by which the Mann-Whitney test is applied in psychology is often flawed. To reveal this, I need to clarify which hypothesis the Mann-Whitney test assesses. The Mann-Whitney test can be used to test many different hypotheses (Fay & Proschan, 2010). In psychology, it is typically presented as either a test of equality of the two populations regarding all aspects, which is often called *equality of distributions* (Howell, 2012) or *equality of population medians* (Divine et al., 2018). In the statistical literature, it has been argued that the Mann-Whitney test actually tests stochastic equality (Chung & Romano, 2016; Divine et al., 2018). If outcome scores are denoted by *X* in one population and by *Y* in the other, then stochastic equality is defined as $P(X<Y) = P(X>Y)$. The notation $P(X < Y)$ denotes the probability of a random observation from Population 1 (represented by *X*) to be bigger than a random observation from Population 2 (represented by *Y*). Conversely, $P(X < Y)$ describes the probability for the opposite. Thus, stochastic equality tests whether the probability of an observation from one population being bigger than an observation from the other deviates from random expectation.

The first problem is that the Mann-Whitney test is associated with strong assumptions for all three hypotheses (Chung & Romano, 2016; Fay & Proschan, 2010), which are rarely met in psychology. For example, for the equality of medians hypothesis, the assumption of equal variances across populations must be met, which is rarely the case in psychology (see Delacre et al., 2017).

The second problem is that if the assumptions are not met, this can completely invalidate the Mann-Whitney test (Chung & Romano, 2016). More specifically, the Type I error rates of the Mann-Whitney test can be seriously inflated even for large sample sizes. Thus, the Mann-Whitney test is generally not even asymptotically robust to violations of its assumptions. This is in contrast with the parametric counterpart of the Mann-Whitney test, Welch's *t* test, which is known to be asymptotically robust regarding the violations of its core assumption, normality (Delacre et al., 2017). Thus, somewhat counterintuitively, the nonparametric Mann-Whitney test makes rather strong assumptions and is, in some aspects, more vulnerable to violations of its assumptions than its parametric counterpart.

This observation calls for alternatives to the Mann-Whitney test. Because the Mann-Whitney test can be used for testing different hypotheses, a substitute for each hypothesis is needed. In this article, I concentrate on the stochastic equality hypothesis because first, it is the most appropriate nonparametric operationalization of the intuitive hypothesis that a variable tends to be of equal size across two populations (see Cliff, 1993; Neuhäuser & Ruxton, 2009; or the explanation presented in the next section), and second, it has repeatedly been argued to be the most appropriate hypothesis for the Mann-Whitney test (Chung & Romano, 2016; Divine et al., 2018).

Fortunately, multiple alternatives to the Mann-Whitney test exist that are asymptotically robust tests of stochastic equality. These can be categorized into "classical" procedures and "resampling procedures." Whereas the classical procedures employ theoretical sampling distributions, the resampling procedures estimate the sampling distribution empirically. The classical procedures include the Fligner-Policello (Fligner & Policello, 1981), Cliff's (1993), and the Bunner-Munzel (Brunner & Munzel, 2000) tests. The resampling procedures include the permutation version of the Brunner-Munzel test (Neubert & Brunner, 2007), the Reiczigel approach (Reiczigel et al., 2005), and the Ruscio approach (Ruscio & Mullen, 2012).

The performance of the classical procedures is similar, and no classical approach is generally superior (Delaney & Vargha, 2002). The resampling procedures have not yet been compared extensively with each other. A first small comparison (Neubert & Brunner, 2007) suggests that the Reiczigel test and the permutation Brunner-Munzel test behave similarly to each other as well as to the classical Brunner-Munzel test. However, the Reiczigel test was too conservative, and the classical Brunner-Munzel test was too liberal, problems that did not occur for the permutation Brunner-Munzel test. Although the results of Ruscio and Mullen (2012) suggest that the Ruscio approach outperforms the classical approaches for confidence interval generation, its utility as a hypothesis test is questioned by its inventors (Ruscio & Mullen, 2012). In summary, the current evidence seems to favor the permutation Brunner-Munzel test weakly. In addition, the permutation Brunner-Munzel test has the advantage that it is close to the Mann-Whitney test, which is also a permutation test. Therefore, I focus on the permutation Brunner-Munzel test in this article.

It has not been investigated how the Brunner-Munzel test and the Mann-Whitney test compare in terms of power, particularly when the assumptions are met. This is important because typically a modification of a test to make it more robust reduces its power. To address this question, I perform a power comparison. Unexpectedly, the Brunner-Munzel test was almost always equally as powerful as the Mann-Whitney test or even more powerful. The only exception was skewed data together with unequal sample sizes, for which the Mann-Whitney test had a small power advantage.

Despite the advantages of the Brunner-Munzel test, the Mann-Whitney test is still widespread in psychology, whereas the Brunner-Munzel test is mostly unknown. A Google Scholar search for articles published between 2015 and 2020 in journals that contain *psychology* in their name and *Brunner-Munzel* in their text led to two results, whereas 3,350 results contained *Mann-Whitney*.[3] One reason for the continuing popularity of the Mann-Whitney test in psychology might be that whereas the flaws of the Mann-Whitney test and the advantages of the Brunner-Munzel test are well known in the statistical literature, those articles tend to be too technical to be accessible to applied psychologists or even teachers of psychological methods. Furthermore, the arguments that I summarized so far are spread throughout multiple statistical articles. Moreover, the articles describing the Brunner-Munzel test (Brunner & Munzel, 2000; Neubert & Brunner, 2007) introduce the Brunner-Munzel test in a rather technical manner, disconnected from the Mann-Whitney test. Finally, practical advice on how to apply and report on the Brunner-Munzel test is missing.

I thus begin this article by reviewing and demonstrating the flaws of the Mann-Whitney test in more detail. Then, I introduce the Brunner-Munzel test in a nontechnical manner, as a straightforward modification of the Mann-Whitney test. I continue with a simulation study, comparing the power of the Mann-Whitney and Brunner-Munzel tests. After that, I provide practical advice for applied researchers, in particular, which software to use for the Brunner-Munzel test and how to report on and interpret its results.

Flaws of the Mann-Whitney Test

Notation

A variable has been observed across two groups. For Group 1, the observations are denoted as $x_1,\ldots, x_{n_1}$ and for Group 2 as $y_1,\ldots, y_{n_2}$. Thus, the sample sizes of Groups 1 and 2 are n_1 and n_2, respectively. For Group 1, the observations are assumed to be a random sample of Population 1 and for Group 2 of Population 2. The distributions of Populations 1 and 2 are P and Q, respectively.

Multiple perspectives on the Mann-Whitney test

For analyzing ordinal data, the Mann-Whitney test is a valid test under at least three different perspectives (Fay & Proschan, 2010). A perspective is defined by a combination of null hypothesis, alternative hypothesis, and assumptions. A test is called valid for a certain perspective if its Type I error rate is always below the desired significance level α (typically set to 5%) when the assumptions are met. The three valid perspectives are equality of distributions, equality of medians, and stochastic equality.

Below I list the null hypothesis, alternative hypothesis, and assumptions for all three perspectives. Note that all perspectives additionally assume independence, that is, all observations are assumed to be independent of each other.

- Equality of distributions
  - Null hypothesis: Population distributions are equal in all aspects; $P = Q$.
  - Alternative hypothesis: Population distributions differ in any aspect; $P \neq Q$.
  - Assumptions: None.
- Equality of medians
  - Null hypothesis: Population medians are equal; $Mdn(P) = Mdn(Q)$.
  - Alternative hypothesis: Population medians are unequal; $Mdn(P) \neq Mdn(Q)$.
  - Assumptions: If the null hypothesis is true (no differences in medians), the population distributions are identical ($P = Q$).
- Stochastic equality
  - Null hypothesis: $P(X > Y) = P(X < Y)$.
  - Alternative hypothesis: $P(X > Y) \neq P(X < Y)$.
  - Assumptions: If the null hypothesis is true, the population distributions are identical ($P = Q$).

The assumptions for the different perspectives are all a special case of the Mann-Whitney test's core assumption, exchangeability. In the Mann-Whitney test setting, exchangeability reduces to if the null hypothesis is true, the two population distributions must be identical.

Note that the equality of distributions null hypothesis is wrong if the population distributions differ in any aspect. It encompasses the other two null hypotheses, in the sense that if any of those is wrong, then the equality of distributions null hypothesis is wrong.

Applying the three perspectives to the depression therapy example from the introduction leads to answers to the following different and, in principle, equally valid research questions. I assume now that the numbers displayed in Table 1 represent the population values. For

equality of distributions, the relative frequencies across the two groups are different, so the distributions are not equal. For equality of medians, the median improvement level in both groups is slight improvement. Thus, regarding the median improvement level, the therapy and the control groups do not differ. For stochastic equality, the probability that a random person from the treatment population has a larger increase than a random person from the control population [$P(X > Y)$] is 0.36, which is the same as the reverse [$P(X < Y) = 0.36$], and consequently, the two distributions are stochastically equal.

Requirements for a reasonable test

To demonstrate that the Mann-Whitney test is not a reasonable test for any of those perspectives, I concentrate on two requirements. Essentially, they formalize asymptotical robustness of a test to its assumptions and ensure that a test still works reasonably if its assumptions are not met.

The first requirement is *asymptotic validity* under realistic assumptions. Again, a test is valid if its Type I error rate is always lower than the desired significance level α. Asymptotic validity weakens validity because it requires the Type I error rate to only approach the significance level α with increasing sample size. However, validity is required only if the assumptions are met, whereas here, asymptotic validity is also required if the assumptions are not met.

The second requirement is *consistency* under realistic assumptions and relates to power. Asymptotic validity alone is easy to obtain. For example, a test that randomly rejects the null hypothesis with probability α without considering the data is asymptotically valid but useless. Thus, power also needs to be considered. Classically, one aims at identifying the test with uniformly highest power among all valid tests. Such a test is called a *uniformly most powerful test*. However, uniformly most powerful tests can be identified only under rigorous assumptions. For example, even if all assumptions are met, Student's t test is not uniformly most powerful. A weaker and obtainable criterion is consistency. A test is consistent if its power always approaches 1 as the sample size increases.

Problems with asymptotic validity

First, I argue that the assumptions associated with the equality of medians and stochastic equality perspectives are not realistic in psychology. The assumptions for both perspectives are similar. If there is no difference between the two populations concerning the tested aspect (equality of medians or stochastic equality), there may not be a difference at all. This situation is unrealistic in psychology. For example, there are often variance differences between two populations despite there not being mean differences (Delacre et al., 2017).

This observation calls for investigating the performance of the Mann-Whitney test under more realistic assumptions. I use the assumption of the two populations' variances being finite because the parametric Welch t test is asymptotically valid and consistent under this assumption (Fay & Proschan, 2010). For both perspectives, I generated data such that the corresponding null hypothesis is true.

For the equal medians perspective, the distribution P of the first population was the uniform distribution on the interval 0–1. The distribution Q of the second distribution was a mixture of two uniform distributions. With probability 0.5, data were generated from a uniform distribution on the interval 0–0.5 or from a uniform distribution on the interval 0.5–1.5. Note that the median of both populations was thus 0.5.

For the stochastic equality perspective, both populations were normally distributed with a zero mean. The only difference was that for Population 1, the variance was 1, and for Population 2, it was 4. Note that the hypothesis of stochastic equality is met.

I estimated the Type I error rates for increasing sample sizes. The sample sizes for Group 1 were 10, 50, 100, 250, 500, 750, and 1,000. For the equal medians perspective, the sample size of Group 2 was equal to the sample size of Group 1. For the stochastic equality perspective, the sample size of Group 2 was always 2 times bigger than the sample size of Group 1.

In Figure 1, I display the estimated Type I error rates of the Mann-Whitney test under the equal median and stochastic equality perspectives. Clearly, the Mann-Whitney test is not asymptotically valid. For the equal median perspective, the Type I error rate increased with sample size and even converged to 1. The Type I error rate for the stochastic equality perspective seems to be stable at around 0.09.

Problems with consistency

For the difference in distributions perspective, the Mann-Whitney test is asymptotically valid under the finite variances assumptions. The only assumption needed for the validity of the Mann-Whitney test for the difference in distribution perspective is independence (Fay & Proschan, 2010). Consequently, it is also asymptotically valid. However, the Mann-Whitney test is also not reasonable under this perspective because it is not consistent.

The simulation study I performed for showing that the Mann-Whitney test is not asymptotically valid for the stochastic equality perspective showcases this. Note that in this scenario, the population distribution differed by their variances. Thus, the null hypothesis of equal distributions was false. However, as is visible in Figure 1b,

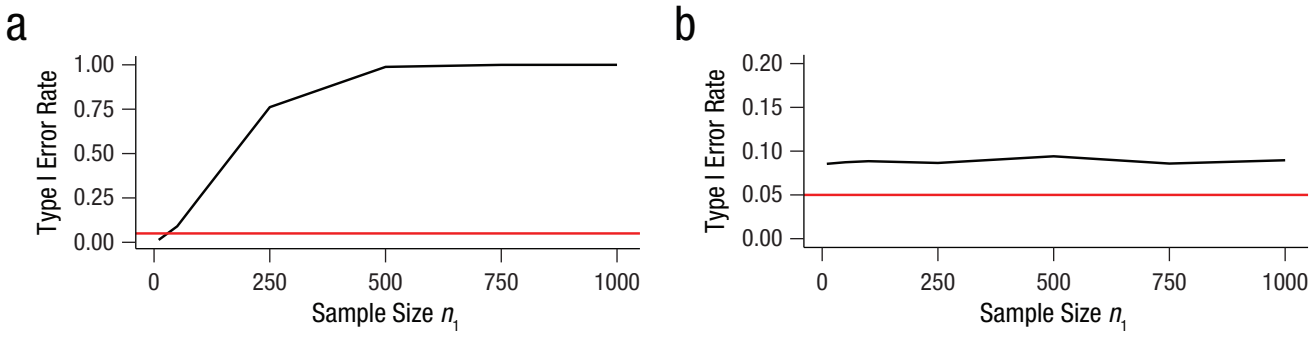

Fig. 1. Estimated Type I error rates of the Mann-Whitney test. The horizontal line displays 0.05, the desired Type I error rate.

the power of the Mann-Whitney test remained relatively constant around 0.09 with increasing sample size instead of approaching 1, as would be the case for a consistent test.

Brunner-Munzel Test

Importance of the stochastic equality perspective

The observations from the previous section call for alternatives to the Mann-Whitney test for each perspective. In this article, I focus on the stochastic equality perspective because more often than not, this is the perspective psychologists should take when applying the Mann-Whitney test. To unpack this statement, I discuss the different perspectives and when they should be chosen.

The equality of distributions perspective stands out because it does not answer a directional question (Schlag, 2015), in the sense that it does not investigate whether a variable tends to be larger in one population than in another. Note that the equality of distributions null hypothesis is wrong if the population distributions differ in any aspect.

However, psychologists typically are interested in making a directional statement (Schlag, 2015). This is especially the case when the Mann-Whitney test is applied because it is understood to be the nonparametric equivalent of the *t* test (Rayner, 2018, Section 1.5), which clearly is a directional test. For example, clinical psychologists typically are interested in establishing that patients who underwent therapy are better off than the control participants, not just that the distributions differ. For these directional questions, the equality of distributions perspective is inappropriate. In contrast, the median and the stochastic equality perspective are both directional.

Which directional perspective is the most appropriate to take should be determined for each research project individually. However, the stochastic equality perspective is closer to the individual person, and thus often to the substantive research question, than the equality of medians perspective. For the equality of medians perspectives, all individuals are summarized by one value, the median. The medians are then compared across the groups. This strategy, summarizing first and then comparing, leads to a lot of information loss. The stochastic equality perspective, in contrast, compares all individuals with each other directly, and thus all available information is considered. I illustrate with the therapy example. Consider that the population frequencies are as displayed in Table 2. Thus, no person in the treatment group had worse improvement than any person in the control group, and many persons in the treatment group had substantially better improvements than all persons in the control group. Thus, clearly, the therapy works. However, under the equality of medians perspective, the therapy and the control treatment are equally effective because for both, the median is slight improvement. In contrast, the stochastic equality perspective correctly captures the difference between the two groups because it compares the individual improvement levels directly: $P(\text{improvement treatment} > \text{improvement control}) = 64\% > 0 = P(\text{improvement treatment} < \text{improvement control})$, that is, the probability that the therapy leads to a bigger improvement than the control treatment is 64%, whereas the probability that the control treatment leads to the bigger improvement than the therapy is 0%.

Mann-Whitney U *statistic*

Because the Brunner-Munzel test modifies the Mann-Whitney test, I first introduce the computational details

Table 2. Relative Frequencies of the Level of Improvement for the Control Group and the Treatment Group

	No improvement	Slight improvement	Substantial improvement	Symptom-free
Control	40%	60%	0%	0%
Treatment	0%	60%	0%	40%

of the latter. For the definition of the Mann-Whitney U test statistic, I introduce the following function, which quantifies whether an observation x_i from Group 1 is smaller, bigger, or equal than an observation y_i from Group 2:

$$S(x_i, y_j) = \begin{cases} 1 & \text{if } x_i < y_j \\ \frac{1}{2} & \text{if } x_i = y_j \\ 0 & \text{otherwise} \end{cases}$$

The Mann-Whitney U statistic is then the sum of this function over all possible Group 1, Group 2 pairings

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} S(x_i, y_j).$$

Stochastic superiority statistic $\hat{\mathbf{p}}''$

The Brunner-Munzel test builds on an equivalent version of the Mann-Whitney test, which uses an easier to interpret modification of the U statistic. Instead of using the sum over all possible pairings, the average value is calculated. The resulting test statistic is

$$\hat{p}'' = \frac{U}{n_1 n_2}.$$

The test statistic $\hat{p}''$ is inherently linked to the stochastic equality perspective. The stochastic equality null hypothesis $P(X > Y) = P(X < Y)$ can be equivalently expressed as $p'' = P(X < Y) + 0.5\, P(X = Y) = 0.5$. Thus, if the Populations 1 and 2 are stochastistically equal, $p'' = 0.5$. If $P(X < Y) > P(X > Y)$, then $p'' > 0.5$, and if $P(X < Y) < P(X > Y)$, then $p'' < 0.5$. The term $0.5\, P(X = Y)$ is needed to take into account ties, which are almost guaranteed in ordinal data. I call the concept p'' *stochastic superiority*. The stochastic superiority p'' is also used as an alternative effect size obtained from Cohen's d and is known as the *probability of superiority* (Ruscio & Mullen, 2012). An important distinction is that stochastic superiority, as defined here, is estimated without distributional assumptions. In contrast, normality is assumed if it is obtained from Cohen's d.

The stochastic superiority estimate $\hat{p}''$ is what statisticians call a consistent estimator of the stochastic superiority p. Thus, with larger sample sizes, $\hat{p}''$ gets closer and closer to the population value p''. In contrast to Mann-Whitney U, the stochastic superiority statistic $\hat{p}''$ can thus be interpreted easily. If $\hat{p}''$ is around 0.5, this is evidence for stochastic equality. If $\hat{p}''$ is significantly bigger (smaller) then 0.5, this is evidence for Population 2 (Population 1) being stochastically superior, that is, tending to bigger values.

From stochastic superiority statistic $\hat{\mathbf{p}}''$ *to* P *value*

For the Mann-Whitney test, multiple techniques exist to translate the stochastic superiority statistic $\hat{p}''$ into a p value. The most accurate technique employs the permutation approach, which is described in detail in Good (2005).

Essentially, the permutation approach is a generally applicable technique to translate a test statistic into a p value. It comes with one core assumption, exchangeability, and thus is the reasion why exchangeability is also the core assumptions of the Mann-Whitney test. In addition, it is well known that the permutation test is not asymptotically valid when exchangeability is violated (Chung & Romano, 2013). This explains why the Mann-Whitney test is not asymptotically valid under the equality of medians and stochastic equality perspectives if the corresponding assumptions are violated.

However, statisticians have invented a general technique, called *studentization*, which makes a permutation test asymptotically valid even if exchangeability is violated (Chung & Romano, 2013; Janssen, 1997). Neubert and Brunner (2007) applied this technique to the Mann-Whitney test to make it asymptotically valid. The resulting test is the Brunner-Munzel test.

Studentizing the stochastic superiority statistic $\hat{\mathbf{p}}''$

The core idea behind studentization is surprisingly simple. The test statistic is transformed such that it has approximately a standard normal distribution under the null hypothesis. However, the required derivations are quite technical. The interested reader is referred to Neubert and Brunner (2007) and Brunner et al. (2018). Here, I will present only the result.

The modified test statistic is the studentized stochastic superiority statistic

$$\hat{p}^* = \frac{\hat{p}'' - 0.5}{\hat{\sigma}_p}$$

with the pooled standard deviation $\hat{\sigma}_p$ defined as follows

$$\hat{\sigma}_p = \sqrt{\frac{1}{n_1 n_2}\left(\frac{1}{n_2}\hat{\sigma}_1 + \frac{1}{n_1}\hat{\sigma}_2\right)}$$

with $\hat{\sigma}_i$ as defined in Neubert and Brunner (2007, p. 5194).

As for the Mann-Whitney test, optimally, a permutation approach is employed to translate the studentized stochastic superiority statistic $\hat{p}^*$ into a p value. The resulting test is known as the *permutation* Brunner-Munzel test. Note that the permutation versions of the

Mann-Whitney test and the Brunner-Munzel test differ only by the fact that the former uses the normal stochastic superiority $\hat{p}''$ and the latter uses the studentized stochastic superiority estimate $\hat{p}^*$ as test statistic.

However, this rather small difference leads to substantially different properties. The Mann-Whitney test is not asymptotically valid for the stochastic equality perspective. Thus, if the assumption of exchangeability is not met, the Type I error rate can be substantially higher than the significance level, even for large samples. I demonstrated this in the preceding section (see Fig. 1). In contrast, Brunner and Munzel (2000) proved that the Brunner-Munzel test is asymptotically valid under the rather general and reasonable assumption that the variances of both population distributions are finite. Note that this is exactly the same condition under which the parametric Welch *t* test is asymptotically valid.

Simulation Study

The fact that the Brunner-Munzel test is asymptotically valid does not guarantee that the Type I error rates are satisfactory for sample sizes as they occur in psychology. In addition, the theoretical results do not quantify how the Brunner-Munzel test compares with the Mann-Whitney test in terms of power. In particular, if the assumptions of the Mann-Whitney test are met, it could be that the Mann-Whitney test has higher power and thus should be preferred in this case. The appropriate tool to investigate these questions is a simulation study.

Multiple simulation studies have already investigated under which conditions the Type I error rate of the Brunner-Munzel test is sufficiently close to the significance level α (Brunner & Munzel, 2000; Delaney & Vargha, 2002; Neubert & Brunner, 2007; Neuhäuser, 2010; Neuhäuser & Ruxton, 2009). In general, the results are reassuring. Even for small samples, the Type I error rate was sufficiently close to the significance level α. As one example, in the simulation study performed by Brunner and Munzel (2000), the Type I error rates were between 4.6% and 5.7% for a significance level of $\alpha = 5\%$ under all investigated conditions with samples sizes $n_1, n_2 > 10$.

Rather surprisingly, to my knowledge, no previous study compared the power of the Brunner-Munzel and Mann-Whitney tests. Thus, this is the focus of this simulation study. I also report Type I error rates because no previous study used a design specifically tailored to mirror the factors as they occur in psychology and to keep this article self-contained.

The full details of the simulation study can be found in the Supplemental Material available online. Here, I focus on summarizing the results.

Concerning Type I error rates, the results are in line with the results from previous simulation studies. The estimated Type I error rate of the Brunner-Munzel test was never higher than 6%. For sample size $n_1 \geq 50$, it was even always in the range of 4.90% to 5.16%. This is in contrast to the Mann-Whitney test for which the Type I error rate was outside of the range 4% to 6% in 15.66% of the conditions. Even for the largest sample size considered ($n_1 = 150$), the Type I error rate of the Mann-Whitney test ranged from 3.21% to 10.23%.

For power, I consider only the conditions in which exchangeability was met. Only in these conditions is it guaranteed that both tests have the same Type I error rate, which is required for a meaningful power comparison. Overall, the power of both tests was similar. The mean of the absolute power differences was 1.38%, and the median was 0.50% (see also Table S2 in the Supplemental Material).

However, for the symmetric distributions, the Brunner-Munzel test had more power than the Mann-Whitney test.[4] The power advantage ranged from 0% to 4.05%.

For the skewed distribution, the opposite pattern emerged. Although the Mann-Whitney test was not more powerful in all conditions, it was more powerful in the majority of the conditions. The Mann-Whitney test's maximum power advantage was 10.33%.

In summary, the Type I error rate of the Brunner-Munzel test converged very quickly to the significance level of $\alpha = 5\%$. Both tests performed similarly in terms of power; the Brunner-Munzel test had a slight advantage for the symmetric distributions, and the Mann-Whitney had a slight advantage for the skewed distribution.

Brunner-Munzel Test in Practice

Example data set

In this section, I discuss practical considerations by applying the Brunner-Munzel test to an example data set. The example data come from the Eurobarometer 73.2 (European Commission, 2012). The data are open and available at http://doi.org/10.4232/1.11429. As part of the Europarameter 73.2, participants were asked the question, "How often during the past 4 weeks have you felt downhearted and depressed?," which had the ordinal answer options *all the time*, *most of the time*, *sometimes*, *rarely*, and *never* and the noninformative *don't know*. The research question is whether either men or women feel depressed more often. Because the data are ordinal and the research question is directional, the appropriate formalization of this question is to test for stochastic equality.

Computational considerations

The first practical challenge is computational. After removing all missing values, including the don't know responses, the data set consisted of responses from 14,430 women and 12,199 men. Calculating the test statistic for all possible permutations of the data, as required by the permutation Brunner-Munzel test, is computationally infeasible. The number of permutations needed would be $> 10^{7973}$. The computational infeasibility is not limited to huge samples. For example, if both groups have a size of 100, the number of permutations required is $> 9 \times 10^{58}$, which is still computationally infeasible.

This computational infeasibility of the Brunner-Munzel test for moderately sized samples is shared by all permutation tests, thus also by the Mann-Whitney test. In general, there are two approaches to solve this problem. The first uses a random number of permutations instead of all permutations (this is sometimes called *approximate* or *Monte Carlo approach*). The second uses a parametric distribution to approximate the permutation distribution (this is called the *asymptotic approach*).

For the Brunner-Munzel test, there seems to be no consensus yet which approach is better. However, only for the asymptotic approach, detailed evidence that it is reasonably accurate for moderately large samples is available (Brunner & Munzel, 2000). In addition, in this article's simulation study, I used the asymptotic approach, confirming those results. Thus, for now, I recommend using the asymptotic approach if the exact approach is computationally infeasible.

The difference between the exact permutation approach and the asymptotic approach is most pronounced for very small samples, in which the exact permutation approach is computationally feasible. In particular, the permutation approach has better Type I error control in very small samples ($n_1, n_2 < 10$) compared with the asymptotic approach (Neubert & Brunner, 2007). However, starting at small samples ($n_1, n_2 \geq 10$), the Type I error control of the asymptotic approach is reasonable (Brunner & Munzel, 2000).

Reporting results

Before reporting the results, I introduce guidelines for reporting on the Brunner-Munzel test. I follow the American Psychological Association principles for reporting on statistical tests as closely as possible. Note that my recommendations are different from the standard recommendations for the Mann-Whitney test (Field, 2017, p. 296). This is not because the Brunner-Munzel test is conceptually different from the Mann-Whitney test but, rather, is because of the misconception in the literature that the Mann-Whitney test is a reasonable test for median differences.

The appropriate test statistic to report is the studentized stochastic superiority estimate $\hat{p}^*$. If the exact or the approximate permutation versions of the test are used, it does not have associated degrees of freedom. For the asymptotic version, the studentized stochastic superiority is assumed to be distributed according to a t distribution with certain degrees of freedom,[5] equivalently to the t test. Consequently, the degrees of freedom should be reported.

As a measure of effect size, the raw stochastic superiority estimate $\hat{p}''$ can be recommended because it directly estimates the stochastic superiority.

Applying those guidelines to the example data set leads to the following results: Women and men were not stochastically equal in their reporting of how often they felt depressed, $\hat{p}^*(26100.38) = -16.05$, $p < .001$, $\hat{p}'' = 0.4462$. The probability that a random woman reported less frequent depressive feelings than a random man was 0.4462, splitting ties evenly. Consequently, women tended to feel depressed more often than men.

Interpretation of stochastic superiority effect size $\hat{p}''$

To guide the interpretation of the stochastic superiority effect size, multiple approaches have been proposed (Cliff, 1993; Divine et al., 2018). I discuss the most prominent here to help readers understand the stochastic superiority effect size better and give readers a tool set to explain the results of applying the Brunner-Munzel test to their readers.

It helps to interpret the stochastic superiority effect size $\hat{p}''$ as summarizing the outcome of a contest. In this contest, the observations from the two groups compete. In the example, the two groups are women and men, and the observations are how often a participant felt depressed. Each observation from a group competes with all observations from the other group. A match between two observations is decided as follows. If one observation is bigger, then it wins, and one point is awarded to its group. If both observations are equal, then half a point is awarded to both groups. The stochastic superiority effect size $\hat{p}''$ is then the proportion of points won by the second group. Consequently, the men won $\hat{p}'' = 44.62\%$ of the points, and the women won $1 - \hat{p}'' = 55.38\%$ of the point. Thus, because the women won significantly more points, they tended to report more frequent depressive feelings.

The stochastic equality effect size can be visualized as a bubble plot (see Fig. 2). In this plot, points above the diagonal represent Group 1, Group 2 pairs in which the Group 2 observation was bigger. For points below the diagonal, the Group 1 observation was bigger, and for points on the diagonal, both observations were equal.

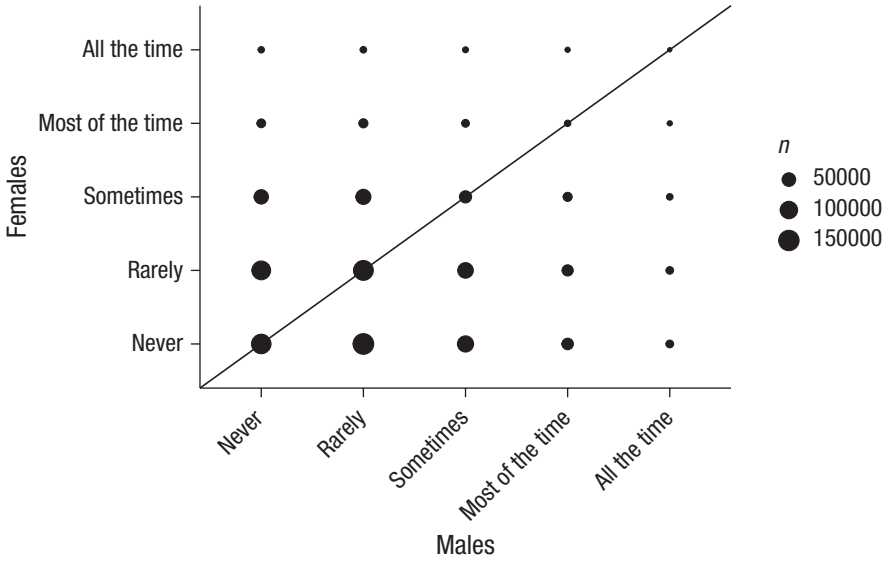

Fig. 2. Bubble plot for the example data set.

The stochastic equality effect size summarizes this plot by computing the proportion of points above the diagonal, including half of the points on the diagonal.

O'Brien and Castelloe (2006) suggested to transform the stochastic superiority effect size $\hat{p}''$ to the so-called Wilcoxin-Mann-Whitney (WMW) odds via $WMW_{odds} = \hat{p}''/(1 - \hat{p}'')$. Thus, the WMW_{odds} quantify how many points Group 2 won in relation to Group 1. For the example, this amounts to $0.4462/(1 - 0.4462) = 80.58\%$. Thus, the men won only 80.58% of the points that the women won. The WMW_{odds} can also be interpreted as odds. In the example, the odds are roughly 8:10 for a random women reporting less depressive symptoms than a random man, splitting ties evenly. If and only if there is stochastic equality, the WMW_{odds} are 1 in the population.

Another transformation is Cliff's δ (Cliff, 1993). Cliff's δ is estimated as follows: $\hat{\delta} = (1 - \hat{p}'') - \hat{p}''$. It is thus similar to the WMW_{odds} because it also compares the points won by the two groups. However, instead of comparing the points using division, it uses the difference. In particular, Cliff's $\hat{\delta}$ compares how many more points are won by Group 1 compared with Group 2. For the example, Cliff's $\hat{\delta} = 10.76\%$. Thus, women won 10.76% more points than men. If and only if there is stochastic equality, Cliff's $\hat{\delta}$ is 0 in the population.

For this example, one can reveal the causes of the significant stochastic superiority effect size by directly comparing the distributions. In Table 3, the relative frequencies are compared across the two groups. Note that 38.74% of the men but only 31.03% of the women reported that they never felt depressive symptoms. The option rarely was chosen essentially equally often by both genders. In contrast, for sometimes, most of the time, and all of the time, the frequencies are all higher for the women. Together with the lower frequency of women in the never category, this explains the significant stochastic superiority effect size.

Confidence intervals

It is often recommended to report confidence intervals additionally to or instead of the outcome of a hypothesis test. For the example data set, accurate confidence intervals for the stochastic superiority can be obtained from the asymptotic Brunner-Munzel test and confidence intervals for the other effect sizes by transforming the obtained confidence interval for the stochastic superiority (see Appendix). The resulting confidence intervals are [0.44, 0.45] for stochastic superiority, [0.78, 0.83] for the WMW_{odds}, and [0.09, 0.12] for Cliff's δ.

For small samples or if the population stochastic superiority is close to 0 or 1, the confidence interval obtained in this fashion should be treated with caution. There is still no consensus on which approach to use in this case. Potential remedies include a bootstrap approach (Ruscio & Mullen, 2012), inverting the permutation Brunner-Munzel test (Pauly et al., 2016), and multiple other approaches (see Brunner et al., 2018, Section 3.7.2). Implementations of some of these approaches can be found in the R packages *rankFD*, *nparcomp*, and *RProbSup*.

Table 3. Relative Frequencies of the Frequency of Depressive Symptoms for Men and Women

	Never	Rarely	Sometimes	Most of the time	All the time
Men	38.74%	33.63%	21.73%	4.92%	0.98%
Women	31.03%	33.46%	26.65%	7.35%	1.51%

Code

In the Appendix, I provide the R code used for the example analysis discussed here.

Discussion

Summary

I have demonstrated that the Mann-Whitney test is not a reasonable test of equality of medians, distributions, or stochastic equality. This observation motivated identifying a reasonable test for each perspective. Here, I focused on the stochastic equality perspective because it is the most appropriate formalization of the intuitive hypothesis that one population tends to have larger values than another. To this end, I introduced the Brunner-Munzel test. I demonstrated that it is a reasonable test for the stochastic equality perspective with often equal or even higher power than the Mann-Whitney test. Only for skewed data and unequal sample sizes was the Mann-Whitney test more powerful. As a consequence, I recommend that psychologists should use the Brunner-Munzel test as the default nonparametric test instead of the Mann-Whitney test. They should use the Mann-Whitney test only if they are confident that their data meet the assumptions of exchangeability and are skewed, which will rarely be the case. Note that contrary to common practice, preliminary assumption checks, whether based on formal hypothesis tests or visualization techniques, cannot be recommended to establish exchangeability or skewness because preliminary assumption checks are generally flawed. Among other problems, they distort the Type I and II error rates of the actual test of interest (Wells & Hintze, 2007). Valid alternative approaches to establishing assumptions are theory and reason (Wells & Hintze, 2007). To enable researchers to apply the Brunner-Munzel test, I provided practical guidance.

Nonparametric or parametric testing

I did not yet address the question regarding when researchers should adopt the parametric Welch *t* test over the nonparametric Brunner-Munzel test. This question has already received a lot of attention (Delaney & Vargha, 2002; Rietveld & van Hout, 2015; Ruxton, 2006; Ruxton & Neuhäuser, 2019), but there still seems to be no consensus. For now, my recommendation is as follows.

The Welch *t* test is the recommended procedure for testing equality of means (Delacre et al., 2017), whereas I recommend the Brunner-Munzel test for testing stochastic equality. Thus, the choice between the two tests should first and foremost be guided by which hypothesis a researcher intends to test. Because each hypothesis is associated with a different research question, the choice is essentially determined by which research question a researcher aims to answer. Note that this is contrary to the common practice of choosing between tests on the basis of pretesting of assumptions, which, again, should be avoided (Wells & Hintze, 2007).

For ratio and interval data, I illustrate using the running depression therapy example. However, now I assume that the improvement was measured using an interval variable indicating the extent of the improvement, with high values indicating a large improvement. In this example, the equality of means hypothesis is equivalent to the research question of whether the therapy, on average, leads to more improvement than the control treatment. In contrast, the stochastic equality hypothesis is equivalent to whether for a random patient, the therapy has a higher chance to lead to a bigger improvement than the control treatment. I illustrate that those are profoundly different questions using a concrete example. In the therapy group, 90% of the patients did not improve, whereas 10% improved by 20, which is considered a substantial improvement. In the control group, all patients improved by 1, which is a slight improvement. Using a comparison of means, one would conclude that the therapy works given that the mean improvement was bigger ($M = 2$) in the therapy group than in the control group ($M = 1$). In contrast, using stochastic equality, one would conclude that the therapy does not work given that for two random patients from both groups, there is a 90% chance that the patient who was in the control group had a bigger improvement.

Again, which research question to ask should be decided individually by each researcher. However, psychological researchers should consider testing the stochastic equality hypothesis more often than is currently done. For example, for an individual patient, identifying the treatment with a higher chance of leading to a bigger improvement seems more relevant than identifying the therapy with the larger average improvement. At the very least, testing for stochastic equality and reporting the associated effect sizes provides useful additional information.

For ordinal data, testing for stochastic equality should be the default strategy because a comparison of means requires at least an interval scale (Delaney & Vargha, 2002), and the stochastic equality hypothesis and its associated effect sizes are also meaningful for ordinal data. This advice seems to conflict with the common practice of treating ordinal data as continuous, especially if the number of levels is sufficiently large, enabling a comparison of means. However, note that this practice implicitly assumes that the data are not ordinal but, rather, discrete data on an interval scale because it implies that the numerical distance between each pair of subsequent categories is equal. Consequently, whether this practice is appropriate is not related to the numbers

of categories but, rather, whether this central assumption of an interval scale is fulfilled.

In certain situations, the hypothesis of equal means and stochastic equality are equivalent, most importantly, if the distributions of the two populations are symmetric. Only in this situation should the choice between the tests be guided by their statistical properties. To make an informed choice between the two tests, a detailed comparison of the power of Welch's *t* test and the Brunner-Munzel test is needed, which has not been performed. However, the practical value of such a comparison can be questioned because one virtually never knows with certainty that the population distributions are symmetric. Consequently, by default, one should assume that they are not and select between the two tests on the basis of the different hypotheses they test.

Testing median differences and equality of distributions

I have not yet discussed how to best test for median differences or equality of distributions. Testing for equality of distributions is generally a tough problem. A test appropriate for general use must be able to detect all the numerous ways two distributions can differ. This is probably one reason why no consensus has been reached about how to best test for equality of distributions and also casts doubt on whether this will ever be the case. However, I agree with Chung and Romano (2016) that omnibus tests such as the Kolmogorov-Smirnov and the Cramér–von Mises tests, which capture the differences of the entire distributions as opposed to only testing for stochastic equality, should be preferred over the Mann-Whitney test. An overview and comparison of available classical procedures can be found in Thas (2010), whereas Wasserman (2012) provided an overview of some modern alternatives.

For testing median differences, the classically recommended procedure is Mood's median test (Brown & Mood, 1951). However, many modern alternatives and supposed improvements have been proposed (Bonett & Price, 2002; Chung & Romano, 2013; DiCiccio & Efron, 1996; Schlag, 2015; Wilcox, 2006). A detailed independent comparison of those tests is missing, and thus no recommendation can be given yet. Consequently, a comparison of available tests for median differences is recommended for future work.

Conclusion

In conclusion, when investigating directional research questions, psychologists should test for stochastic equality more often. However, instead of the Mann-Whitney test, they should use the Brunner-Munzel test by default.

Appendix

R code

In this appendix, I explain the R code used for the example analysis so that readers can adapt this code for their analysis.

First, the example data set needs to be downloaded from http://doi.org/10.4232/1.11429.

The next step is to import the data into R.

```
library(foreign)
df <- read.spss("ZA5232_v3-0-0.sav",
to.data.frame = TRUE)
```

Then, the relevant variables are selected and missing values removed.

```
df <- df[, c("v187", "v218")]
names(df) <- c("freq_dep", "sex")
df <- df[!is.na(df$freq_dep), ]
```

As the next step, the depressed factor is reverted such that a larger value represents a higher frequency of depressive symptoms and transformed into an ordered factor.

```
library(tidyverse)
df$freq_dep <- fct_rev(df$freq_dep)
df$freq_dep <- as.ordered(df$freq_dep)
males <- df$freq_dep[df$sex == "Male"]
females <- df$freq_dep[
df$sex == "Female"]
```

From the multiple implementations of the Brunner-Munzel test available in R (e.g., in the packages *lawstat*, *rankFD*, and *brunnermunzel*), I recommend the "brunnermunzel.permutation.test" function from the *brunnermunzel* package because of its straightforward interface and because it implements the recommendation to use the permutation version of the Brunner-Munzel test if computationally feasible and the asymptotic version if not computationally feasible.

The Brunner-Munzel test is performed as follows.

```
library(brunnermunzel)
res <- brunnermunzel.permutation.test(
females, males)
print(res)
```

```
##
## Brunner-Munzel Test
##
## data: x and y
## Brunner-Munzel Test Statistic = -16,
## df = 26100, p-value <2e-16
## 95 percent confidence interval:
##  0.440 0.453
## sample estimates:
## P(X<Y)+.5*P(X=Y)
##   0.446
```

Unfortunately, the degrees of freedom are rounded to a whole number. The exact degrees of freedom can be extracted as follows.

```
res$parameter
```

The Wilcoxon-Mann-Whitney odds and Cliff's δ, as well as their confidence intervals, can be extracted from the output.

```
get_wmw_odds <- function(stoc_sup) {
return(stoc_sup / (1 - stoc_sup))
}

get_cliff_delta <- function(stoc_sup) {
return(1 - 2 * stoc_sup)
}

#stochastic_superiority
stoc_sup <- unname(res$estimate)

wmw_odds <- get_wmw_odds(stoc_sup)
print(wmw_odds)
```

```
## [1] 0.806
```

```
cliff_delta <- get_cliff_delta(stoc_sup)
print(cliff_delta)
```

```
## [1] 0.108
```

```
wmw_odds_ci <- get_wmw_odds(res$conf.int)
print(wmw_odds_ci)
```

```
## lower upper
## 0.785 0.827
## attr(,"conf.level")
## [1] 0.95
```

```
cliff_delta_ci <- get_cliff_delta(
 res$conf.int)
cliff_delta_ci <- cliff_delta_ci[c(2, 1)]
print(cliff_delta_ci)
```

```
## upper lower
## 0.0944 0.1207
```

For the bubble plot, I provide a function at https://github.com/karchjd/bubble_plot. The plot presented in the article can be reproduced as follows.

```
source("https://raw.githubusercontent.
 com/karchjd/bubble_plot/master/
 bubble_plot.R")
p <- bubble_plot(females[1:1000],
 males[1:1000])
p <- the_plot+xlab("Males")+ylab("Females")
```

Note that I used only the first 1,000 women and men for computational convenience.

Transparency

Action Editor: Brent Donnellan
Editor: Daniel J. Simons
Author Contributions
J. D. Karch is the sole author of this article and is responsible for its content.
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.
Open Practices
Open Data: https://codeocean.com/capsule/6242621/tree/v4
Open Materials: https://codeocean.com/capsule/6242621/tree/v4
Preregistration: not applicable
All data and materials have been made publicly available via Code Ocean and can be accessed at https://codeocean.com/capsule/6242621/tree/v4. This article has received badges for Open Data and Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges. The code to reproduce this article is made available via Code Ocean and can be accessed at https://codeocean.com/capsule/6242621/tree/v4.

ORCID iD

Julian D. Karch https://orcid.org/0000-0002-1625-2822

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/2515245921999602

Notes

1. Asymptotical robustness is only one aspect of general robustness, as described in, for example, Wilcox (2016). Note that Welch's *t* test is not generally robust. As an example, its power can be heavily affected by violations of normality (see e.g., Wilcox, 2016).
2. I chose this unusually high sample size because it is required for the small mean difference introduced by the change in the coding scheme to be significant. However, the argument presented is relevant for all sample sizes.
3. The exact search strings used were "'mann-whitney' source: psychology" and "'brunner-munzel' source:psychology," both restricted to articles between 2015 and 2020.
4. Rarely, the Mann-Whitney test appears to be slightly more powerful. However, the power advantage was never bigger than 0.0015 and thus could not be distinguished from estimation error.
5. The exact formula is displayed in Equation 5.9 in Brunner and Munzel (2000).

References

Bonett, D. G., & Price, R. M. (2002). Statistical inference for a linear function of medians: Confidence intervals, hypothesis testing, and sample size requirements. *Psychological Methods*, *7*(3), 370–383. https://doi.org/10.1037/1082-989X.7.3.370

Brown, G. W., & Mood, A. M. (1951). On median tests for linear hypotheses. In J. Neyman (Ed.), *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* (pp. 159–166). The Regents of the University of California.

Brunner, E., Bathke, A. C., & Konietschke, F. (2018). *Rank and pseudo-rank procedures for independent observations in factorial designs: Using R and SAS*. Springer International Publishing. https://doi.org/10.1007/978-3-030-02914-2

Brunner, E., & Munzel, U. (2000). The nonparametric Behrens-Fisher problem: Asymptotic theory and a small-sample approximation. *Biometrical Journal*, *42*(1), 17–25. https://doi.org/10.1002/(SICI)1521-4036(200001)42:1<17::AID-BIMJ17>3.0.CO;2-U

Chung, E., & Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics*, *41*(2), 484–507. https://doi.org/10.1214/13-AOS1090

Chung, E., & Romano, J. P. (2016). Asymptotically valid and exact permutation tests based on two-sample U-statistics. *Journal of Statistical Planning and Inference*, *168*, 97–105. https://doi.org/10.1016/j.jspi.2015.07.004

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114*(3), 494–509. https://doi.org/10.1037/0033-2909.114.3.494

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's *t*-test instead of Student's *t*-test. *International Review of Social Psychology*, *30*(1), 92–101. https://doi.org/10.5334/irsp.82

Delaney, H. D., & Vargha, A. (2002). Comparing several robust tests of stochastic equality with ordinally scaled variables and small to moderate sized samples. *Psychological Methods*, *7*(4), 485–503. https://doi.org/10.1037/1082-989X.7.4.485

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, *11*(3), 189–212.

Divine, G. W., Norton, H. J., Barón, A. E., & Juarez-Colunga, E. (2018). The Wilcoxon-Mann-Whitney procedure fails as a test of medians. *The American Statistician*, *72*(3), 278–286. https://doi.org/10.1080/00031305.2017.1305291

European Commission. (2012). *Eurobarometer 73.2 (Feb-Mar 2010)* (ZA5232; Version 3.0.0) [Data file]. TNS OPINION & SOCIAL. GESIS Data Archive, Cologne. https://doi.org/10.4232/1.11429

Fay, M. P., & Proschan, M. A. (2010). Wilcoxon-Mann-Whitney or *t*-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*, *4*, 1–39. https://doi.org/10.1214/09-SS051

Field, A. (2017). *Discovering statistics using IBM SPSS statistics* (5th ed.). SAGE.

Fligner, M. A., & Policello, G. E. (1981). Robust rank procedures for the Behrens-Fisher problem. *Journal of the American Statistical Association*, *76*(373), 162–168. https://doi.org/10.1080/01621459.1981.10477623

Good, P. (2005). *Permutation, parametric and bootstrap tests of hypotheses* (3rd ed.). Springer.

Howell, D. C. (2012). *Statistical methods for psychology*. Cengage Learning.

Janssen, A. (1997). Studentized permutation tests for non-i.i.d. Hypotheses and the generalized Behrens-Fisher problem. *Statistics & Probability Letters*, *36*(1), 9–21. https://doi.org/10.1016/S0167-7152(97)00043-6

Neubert, K., & Brunner, E. (2007). A studentized permutation test for the non-parametric Behrens-Fisher problem. *Computational Statistics & Data Analysis*, *51*(10), 5192–5204. https://doi.org/10.1016/j.csda.2006.05.024

Neuhäuser, M. (2010). A nonparametric two-sample comparison for skewed data with unequal variances. *Journal of Clinical Epidemiology*, *63*(6), 691–693. https://doi.org/10.1016/j.jclinepi.2009.08.026

Neuhäuser, M., & Ruxton, G. D. (2009). Distribution-free two-sample comparisons in the case of heterogeneous variances. *Behavioral Ecology and Sociobiology*, *63*(4), 617–623. https://doi.org/10.1007/s00265-008-0683-4

O'Brien, R. G., & Castelloe, J. (2006). Exploiting the link between the Wilcoxon-Mann-Whitney test and a simple odds statistic. In *Proceedings of the Thirty-First Annual SAS Users Group International Conference 2006*.

Pauly, M., Asendorf, T., & Konietschke, F. (2016). Permutation-based inference for the AUC: A unified approach for continuous and discontinuous data. *Biometrical Journal*, *58*(6), 1319–1337. https://doi.org/10.1002/bimj.201500105

Rayner, J. C. W. (2018). *Introductory nonparametrics*. bookboon.com.

Reiczigel, J., Zakariás, I., & Rózsa, L. (2005). A Bootstrap test of stochastic equality of two populations. *The American Statistician*, *59*(2), 156–161. https://doi.org/10.1198/000313005X23526

Rietveld, T., & van Hout, R. (2015). The *t* test and beyond: Recommendations for testing the central tendencies of two independent samples in research on speech, language and hearing pathology. *Journal of Communication Disorders*, *58*, 158–168. https://doi.org/10.1016/j.jcomdis.2015.08.002

Ruscio, J., & Mullen, T. (2012). Confidence intervals for the probability of superiority effect size measure and the area under a receiver operating characteristic curve. *Multivariate Behavioral Research*, *47*(2), 201–223. https://doi.org/10.1080/00273171.2012.658329

Ruxton, G. D. (2006). The unequal variance *t*-test is an underused alternative to Student's *t*-test and the Mann-Whitney U test. *Behavioral Ecology*, *17*(4), 688–690. https://doi.org/10.1093/beheco/ark016

Ruxton, G., & Neuhäuser, M. (2019). Striving for simple but effective advice for comparing the central tendency of two populations. *Journal of Modern Applied Statistical Methods*, *17*(2), Article eP2567. https://doi.org/10.22237/jmasm/1551908612

Schlag, K. H. (2015). *Who gives direction to statistical testing? Best practice meets mathematically correct tests*. SSRN. https://doi.org/10.2139/ssrn.2660977

Thas, O. (2010). *Comparing distributions*. Springer.

Wasserman, L. (2012). Modern two-sample tests. *Normal Deviate*. https://normaldeviate.wordpress.com/2012/07/14/modern-two-sample-tests/

Wells, C. S., & Hintze, J. M. (2007). Dealing with assumptions underlying statistical tests. *Psychology in the Schools*, *44*(5), 495–502. https://doi.org/10.1002/pits.20241

Wilcox, R. R. (2006). Comparing medians. *Computational Statistics & Data Analysis*, *51*(3), 1934–1943. https://doi.org/10.1016/j.csda.2005.12.008

Wilcox, R. R. (2016). *Introduction to robust estimation and hypothesis testing* (4th ed.). Elsevier.