Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.



Volume 23 Number 13, September 2018 ISSN 1531-7714

# A Note on Using the Nonparametric Levene Test When Population Means Are Unequal

Benjamin R. Shear, *University of Colorado Boulder*
David W. Nordstokke, *University of Calgary*
Bruno D. Zumbo, *University of British Columbia*

This computer simulation study evaluates the robustness of the nonparametric Levene test of equal variances (Nordstokke & Zumbo, 2010) when sampling from populations with unequal (and unknown) means. Testing for population mean differences when population variances are unknown and possibly unequal is often referred to as the Behrens-Fisher problem when the populations are normally distributed, and the generalized Behrens-Fisher problem when the populations are non-normal. The nonparametric Levene test was developed to overcome reductions in power of the original Levene test of equal variances in the case of the generalized Behrens-Fisher problem. We use a Monte Carlo computer simulation to demonstrate that sampling from populations with unequal and unknown means can lead to incorrect (either inflated or decreased) Type I error rates of the nonparametric Levene test. Centering samples using either sample means or medians does not correct the Type I error rates. This note is intended to make applied researchers aware of this problem when testing for the equality of population variances with the NPL test and in general.

The main objective of this simulation study is to demonstrate how differences in population means can impact the accuracy of the nonparametric Levene test of equal variances[1]. Nordstokke and Zumbo (2010) introduced the nonparametric Levene (NPL) test for equality of population variances (or scale) that can be used when samples exhibit non-normality, for example when sampling from skewed distributions. The NPL test involves ranking observed scores and then conducting the original, mean-based Levene test (Levene, 1960) for equal variances on the ranked data. The test has been shown to have good Type I error and power properties when sampling from skewed populations, with both simulated and real data (Nordstokke & Zumbo, 2010; Nordstokke, Zumbo, Cairns, & Saklofske, 2011).

This paper uses a computer simulation to demonstrate an important aspect of the NPL test that was implicitly assumed in the studies by Nordstokke and his colleagues; that is, it is explicit in the earlier studies but not stated in the description of the computational steps of the nonparametric Levene test. The assumption states that samples have been drawn from populations with equal means, although not necessarily equal variances. In the case of comparing variances across two

[1] In the statistical literature, statistics such as the mean or median are referred to as measures of "location" of a random variable, while statistics such as the variance (and standard deviation) or interquartile range are measures of "scale" used to describe the variability or spread of a random variable. As such, one will encounter either terms like "tests of equal variance" or "tests of scale" in the statistical research literature. While the former is often used to describe parametric and the latter nonparametric statistical tests, tests of equal variance are a subset of the more general category of tests of equal scale. Because the statistical literature is not standard or consistent in its usage, we will use the two terms interchangeably.

groups, Nordstokke and Zumbo (2010) describe the computational steps for the NPL test as:

1. Pool observed data together and rank all scores, giving the lowest score the rank of 1.
2. Separate the ranked data back into their original groups.
3. Conduct the original mean-based Levene test of equal variances (described below) on the ranked data.

The original mean-based Levene test for equal variances proceeds by conducting an analysis of variance (ANOVA) on the absolute deviations of the observed scores from their respective group-specific means. That is, the original Levene's test is computed as an ANOVA on the absolute values of the deviations $e_{ij} = X_{ij} - \bar{X}_{\bullet j}$, where $X_{ij}$ is the score for observation *i* in group *j* and $\bar{X}_{\bullet j}$ is the mean of group *j*. In the original Levene test, the $X_{ij}$ are the observed scores, whereas in the NPL test the $X_{ij}$ are the pooled ranks described above. The resulting *F*-statistic from this test is used to evaluate the null hypothesis that the samples are drawn from populations with equal variances.

The NPL test assumes that the samples are drawn from populations with equal means, so that the sample means are also equal in expectation. In Nordstokke and Zumbo (2010), for example, this assumption was satisfied because the simulated data came from populations with equal means. In Nordstokke et al. (2011) mean-centering was described as part of the simulation methodology in order to manipulate group variance ratios, but was not described as part of the steps in computing the NPL test. In this paper, we demonstrate how unequal (and unknown) population means can adversely affect the Type I error rates of the NPL test, and discuss the implications for use of the test. As we discuss in more detail below, this problem arises only when population means are both unequal and unknown; if population means differ by a known amount, the difference can be adequately adjusted and the NPL test used as if population means were equal. As a reminder, a Type I error in this context would be one in which the researcher falsely concludes there is a difference in the population variances when in fact there is not.

## Background

In education and social science-based research contexts it is often important to determine whether two or more groups being studied have statistically equal variances. Nordstokke et al. (2011) discuss two reasons why testing for differences in scale might be important. First, the nominal Type I error rate (and other characteristics) of many statistical procedures for comparing group means, such as the t-test or ANOVA, may be biased if population variances are unequal, and this problem can be especially relevant when group sample sizes are unequal. Many nonparametric statistical tests of location (i.e., means or medians) are subject to these same problems (Nordstokke & Colp, 2018; Nordstokke et al., 2011). Therefore, a test of equal variances may be used as a preliminary test before comparing population means or medians. As a second case, one may be interested in comparing variances directly as an outcome of a study. One may be interested in studying how a particular treatment affects variability rather than the average outcome, for example, or may be studying a treatment with heterogeneous effects that impacts both the variance and mean of an outcome. In these latter cases, the variability of scores may actually be the outcome of interest.

The use of effect sizes is a third reason that researchers may be concerned about testing for equality of variances across groups. The interpretation of many standardized effect sizes (e.g. Cohen's *d*) can be impacted when variances are unequal because the standardizer for these effect sizes is based on the variance (Grissom & Kim, 2005). Given recommendations to report and interpret standardized effect sizes (e.g., Wilkinson & Task Force on Statistical Inference, 1999), this implies that researchers need to address issues of variance (in)equality across groups to meaningfully report and interpret effect sizes. The next section describes how sampling from populations with unequal and unknown means can adversely affect nonparametric tests of scale such as the NPL test when used in these contexts. The subsequent section discusses cases where the population means differ by a known amount.

## Nonparametric Levene Test and Unequal Means

Many early tests for equality of scale were dependent upon the assumption that populations were normally distributed. In order to overcome problems caused by non-normality, Levene (1960) proposed a new

test in which an ANOVA is conducted on the absolute value of the residuals within each group, as described above. Subsequent studies have shown that an alternative version of the Levene test proposed by Brown and Forsythe (1974), using the sample medians instead of the sample means to calculate the deviations, maintains correct Type I error rates under a wider range of conditions, including when sampling from non-normal populations with potentially unequal sample sizes and unequal population means (Boos & Brownie, 2004; Conover, Johnson, & Johnson, 1981; Lim & Loh, 1996). We refer to this version of the Levene test as the LevMED test. Unfortunately, the LevMED test has also been shown to have low power under certain conditions, particularly when sampling from asymmetric non-normal populations (Lim & Loh, 1996; Nordstokke & Zumbo, 2010).

To address this problem, Nordstokke and Zumbo (2010) developed a nonparametric version of the Levene test based upon ranks rather than raw scores. As described above, this test conducts the original mean-based Levene test, but using ranks instead of raw scores. Simulation studies of the small (i.e., finite) sample properties of the NPL test have shown that it has correct Type I error rates under a variety of different sample-sizes and sample size ratios when sampling from both symmetric and skewed populations. In addition, it has been shown to have higher power values than the LevMED under many of these conditions, hence addressing an important limitation of the LevMED test (Nordstokke & Colp, 2014; Nordstokke & Zumbo, 2010).

To date, all simulation studies evaluating the NPL test have assumed population means are equal. That is, they have generally been situated in the (generalized) Behrens-Fisher case described below, where it makes sense to assume equal population means. However, prior studies have shown that sampling from populations with unequal means can cause problems for other nonparametric rank-based tests of scale. Olejnik and Algina (1987), for example, included two nonparametric tests based on ranks that required an assumption of equal population means in their simulation study. They found that when sampling from symmetric distributions with unequal means, these tests tended to become conservative (i.e., had lower than anticipated Type I error rates), while they were liberal (i.e., had inflated Type I error rates) when sampling from asymmetric populations with unequal means.

To gain intuition about how unequal population means could adversely affect nonparametric tests of scale such as the NPL test, consider the following examples. First, imagine one is sampling from two populations with extremely unequal variances, as could be the case in many non-experimental settings. Further, assume that the means of these two populations are also very unequal, and that the two population distributions do not overlap. If we draw two equally-sized samples $n_1 = n_2$, with total sample size $n_1 + n_2 = N$, and conduct the NPL test, the group with the smaller mean will take on the first $N/2$ ranks, while the group with larger mean will take on the remaining $N/2$ ranks. These sets of ranks will now have exactly equal variances, the NPL test conducted using the ranks will yield a statistically non-significant result, and the researcher will incorrectly fail to reject the null hypothesis. Hence the power of the NPL test (and other rank-based tests) can be drastically reduced when population means are unequal.

Conversely, imagine a scenario in which one is sampling from populations with equal variances but unequal means. If the samples are of very different sizes (i.e., $n_1$ is much larger than $n_2$) and the samples again do not overlap (or overlap very little), the larger sample will have ranks spanning a larger range. This could lead to a statistically significant NPL test result, leading to a false rejection of the null hypothesis and, in general, inflated Type I error rates. These scenarios illustrate that population mean differences may interact with other factors (such as the relative sample sizes and degree of asymmetry in the population) in complex ways.

## Nuisance Parameters

A key aspect of these problems is that the population means are both unequal and unknown. If the population means were unequal but known (or if the difference in population means were known), one could simply adjust for the difference in means by subtracting an appropriate value from each sample. When the population means are unknown, however, relying on sample means or medians to make adjustments may not be an adequate solution. Olejnik and Aligna (1987), for example, included two alternative versions of their nonparametric tests of scale, using scores that had been centered to the sample means or medians, in an attempt to correct for unknown population mean differences. These alternative versions generally maintained the correct Type I error rates when sampling from

symmetric populations with unequal means, but tended to have inflated Type I error rates when sampling from asymmetric populations with either equal or unequal means.

This problem is referred to more generally in the statistical literature as one of *nuisance parameters*. *Nuisance parameters* are parameters of a distribution that are not of direct interest, but that must be accounted for when comparing the parameters of direct interest. In general, the variances are *nuisance parameters* when testing for differences in means, while the means are *nuisance parameters* when testing for differences in variances. Accurately testing for mean differences when the population variances (or their ratios) are unknown is referred to as the Behrens-Fisher problem when populations are normally distributed (Scheffé, 1970), and the generalized Behrens-Fisher problem when populations are non-normally distributed (Zumbo & Coulombe, 1997). One must either make an assumption about the variances (e.g., that they are equal) or appropriately adjust test procedures to account for unequal variances. Alternatively, using the NPL test to compare population variances when population means are unknown results in the opposite problem; in this case, the unknown population means are *nuisance parameters* that either must be assumed equal or accounted for.

When using the NPL test as a preliminary step towards testing hypotheses about population means in the (generalized) Behrens-Fisher case, the concern is often with maintaining correct Type I error rates. In other words, the concern is with ensuring an accurate test when the null hypothesis about means is true and the population means are equal. Hence it may be reasonable to resolve the *nuisance parameter* problem by assuming that the population means are equal for the purposes of using the NPL test as a preliminary step prior to the formal test of equal population means.

The second and third contexts for testing variances noted above, however, move beyond the traditional case in which a test of variances is used as a preliminary step towards testing hypotheses about locations, to one in which the variances are of direct interest. In these cases, assuming population means are equal may not be a satisfactory approach. Even in the Behrens-Fisher context, using tests of scale in experimental versus non-experimental research contexts is also a relevant concern. While it may be plausible to assume equal population means in an experimental context where participants are sampled from a single population and randomly assigned to experimental conditions, it may not be a plausible assumption in non-experimental research contexts with pre-existing groups. The *nuisance parameter* problem re-emerges in these contexts. In this study, we investigate how the *nuisance parameter* of mean differences can affect the NPL test for equality of variances.

## Purpose

There appear to be adequate solutions to test for equality of variances when populations are normally distributed (or symmetric), in both the Behrens-Fisher case and more generally when the population means differ. In the generalized Behrens-Fisher case, when population means are unknown but can be assumed equal, tests such as the NPL may provide a good approach. However, when the population means cannot be assumed equal and are unknown, as is the case in many non-experimental settings, there are less clear solutions. Widely recommended tests, such as the LevMED test, which is robust across many conditions, may suffer from such low power that the possibility of Type II errors becomes a concern.

The NPL test was developed to help address this problem by providing a robust test of scale when sampling from non-normal, asymmetric populations, but its operating characteristics (i.e., Type I error rate and power) have not been studied when moving beyond the generalized Behrens-Fisher case to conditions in which population means are both unknown and potentially unequal. Because it is necessary to ensure correct Type I error rates before evaluating the power of a test, this study explores whether the NPL test maintains correct Type I error rates under a variety of conditions when the population means differ. Specifically, this study addresses two research questions: a) is the NPL test robust when population means are unequal? and b) are versions of the NPL test that center scores using sample means or medians prior to ranking robust when population means are unequal?

## Methods

A Monte Carlo computer simulation was used to explore how the Type I error rates of the NPL test are affected by unequal population means. A computer simulation allows us to evaluate how well a statistical test works under known conditions, and provide insight about whether the test is likely to work well in applied contexts.

The simulation compares the performance of four tests of scale: the parametric LevMED test, the original (uncentered) NPL test, a mean-centered version of the NPL test (NPL-MEAN) and a median-centered version of the NPL test (NPL-MED). To conduct the NPL-MEAN and NPL-MED tests, we added an initial step prior to those described above for the NPL test. For the NPL-MEAN test, all scores were centered to their respective group sample means prior to pooling and ranking. For the NPL-MED test, scores were centered using their respective group sample medians prior to ranking.

To investigate performance of the various tests, four factors were varied: (1) total sample size ($N$=96, 48, 24), (2) sample size ratio ($n_1/n_2$ =1/3, 1/2, 1/1, 2/1, 3/1), (3) standardized mean difference between populations ($D$=0, 1, 2), and (4) population skewness (skewness=0, 1, 2, 3). In all conditions the true population variances were equal, so that simulation results document the Type I error rates of the tests. The mean differences were standardized by the (common) population standard deviation. A fully crossed computer simulation design with 3x5x3x4=180 conditions was utilized. Conditions were chosen to cover a range of contexts and to match those considered in prior studies (e.g., Nordstokke & Zumbo, 2010), with the only difference being that mean differences were also included as an experimental factor. For each condition we generated 5,000 pairs of random samples, where each sample was drawn from a population with the indicated parameters. We then conducted the four tests for the equality of variances on each pair of samples and recorded whether the result was statistically significant at the 0.05 level. Although the true population means are known in this simulation study, we applied the tests as if the population means were unknown, as would be done in practice. Following Bradley's (1978) liberal criterion, we considered any test that maintained a Type I error rate below 0.075 (for a nominal rate of 0.05) to be robust. Bradley also suggests that a test should maintain a Type I error rate of at least 0.025 (again for a nominal rate of 0.05) to be considered accurate. Although a hypothesis test is considered conservative and could have reduced power when Type I error rates fall below the nominal level, the test is not invalid per se. Hence, we focus only on whether each test has inflated Type I error rates, and would be considered invalid. All simulations and data analyses were conducted using the software package R (R Core Team, 2017). Custom functions were written to carry out the various NPL tests.

The following procedure was used to generate samples from populations with the desired characteristics. For each condition, we generated two independent samples, of size $n_1$ and $n_2$. Each sample was generated as a random draw from an approximate chi-squared distribution using the "rchisq()" function in R, with the degrees of freedom ($df$) adjusted to achieve the desired skewness. The $df$ values used were 1000, 7.4, 2.2, and 0.83, to simulate population skewness levels of approximately 0, 1, 2, and 3, respectively, and for consistency with prior simulation studies using the NPL test (e.g., Nordstokke & Zumbo, 2010)[2]. Next, we subtracted $df$ from each sampled value and divided all

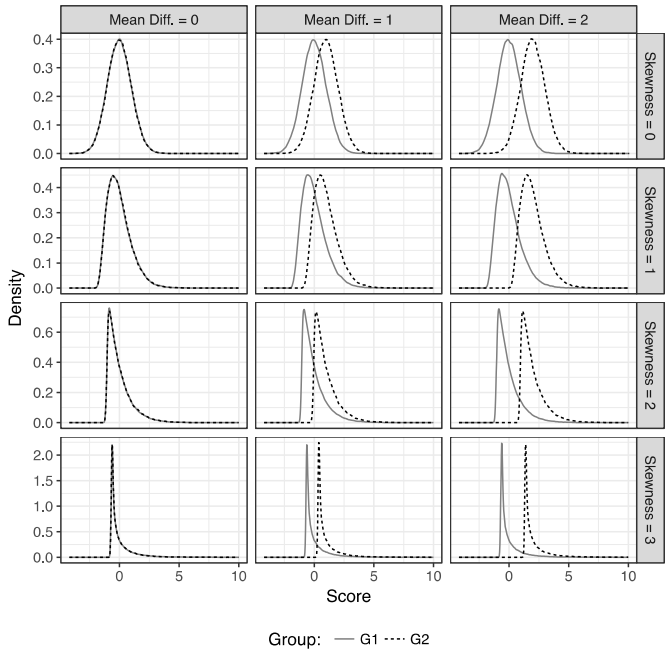


**Figure 1.** Density plots of parent population distributions for different levels of skewness and mean differences. Densities are estimated based on random samples of size N=100,000 from the data generation procedure described in the text.

[2] A chi-square distribution with $k$ degrees of freedom has a mean of $k$, a variance of $2k$ and skewness of $\sqrt{8/k}$. These properties are used as described below in order to sample from standardized populations that maintain the shape and skewness of a chi-squared distribution. In a true chi-squared distribution $k$ is a positive integer greater than 0, and hence the distributions used here are approximate chi-squared distributions.

sampled values by $\sqrt{2 * df}$. This resulted in two samples from populations with the desired level of skewness, each with a population mean of zero and variance of 1. A constant (either 0, 1, or 2) was then added to all observations from the second sample to achieve the desired difference in population means. The skewness of both populations being compared was equal within each condition.

To better illustrate the nature of the populations being sampled, Figure 1 shows estimated density plots for the conditions included. Each row of plots represents a different level of skewness, while each column shows a different magnitude of population mean differences. For example, the top right panel shows the populations sampled from in the condition for which skewness = 0 and the mean difference = 2. These density plots were estimated by generating samples of size N=100,000 from the relevant populations following the procedure described above.

## Results

The simulation results are presented in Tables 1 through 4. Each table presents observed Type I error rates for one of the four tests described above. Observed Type I error rates exceeding the 0.075 criterion are indicated in bold. Table 1, for example, presents the observed Type I error rates for the LevMED test. The results indicated that, as anticipated, the LevMED test-maintained Type I error rates at or near the 0.05 level for all conditions, and never exceeded the 0.075 criterion. Hence the LevMED test would be considered robust across all conditions studied, consistent with prior research.

Table 2 presents Type I error rates for the original NPL test with no centering. The first three columns, for conditions in which the population means are equal, show that the NPL test is robust to varying sample size ratios and levels of skewness in the population distributions. Type I error rates when there were no

**Table 1.** Type I Error Rates for the Median-based Levene Test (LevMED)

| | | Mean Diff. = 0 | | | Mean Diff. = 1 | | | Mean Diff. = 2 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Skew** | **n1/n2** | **N=24** | **N=48** | **N=96** | **N=24** | **N=48** | **N=96** | **N=24** | **N=48** | **N=96** |
| 0 | 0.33 | 0.038 | 0.040 | 0.044 | 0.040 | 0.041 | 0.042 | 0.034 | 0.042 | 0.044 |
| | 0.5 | 0.041 | 0.045 | 0.049 | 0.038 | 0.042 | 0.050 | 0.037 | 0.040 | 0.047 |
| | 1 | 0.035 | 0.039 | 0.042 | 0.033 | 0.039 | 0.046 | 0.039 | 0.046 | 0.046 |
| | 2 | 0.041 | 0.041 | 0.045 | 0.035 | 0.041 | 0.048 | 0.037 | 0.048 | 0.051 |
| | 3 | 0.041 | 0.044 | 0.043 | 0.040 | 0.045 | 0.045 | 0.037 | 0.042 | 0.044 |
| 1 | 0.33 | 0.041 | 0.047 | 0.044 | 0.044 | 0.043 | 0.053 | 0.041 | 0.042 | 0.045 |
| | 0.5 | 0.044 | 0.042 | 0.041 | 0.044 | 0.047 | 0.045 | 0.038 | 0.042 | 0.049 |
| | 1 | 0.041 | 0.045 | 0.047 | 0.050 | 0.048 | 0.050 | 0.036 | 0.045 | 0.049 |
| | 2 | 0.042 | 0.046 | 0.049 | 0.042 | 0.044 | 0.048 | 0.043 | 0.051 | 0.047 |
| | 3 | 0.045 | 0.042 | 0.038 | 0.040 | 0.042 | 0.043 | 0.040 | 0.045 | 0.048 |
| 2 | 0.33 | 0.048 | 0.044 | 0.046 | 0.044 | 0.052 | 0.045 | 0.050 | 0.046 | 0.048 |
| | 0.5 | 0.047 | 0.041 | 0.050 | 0.047 | 0.050 | 0.046 | 0.052 | 0.047 | 0.052 |
| | 1 | 0.046 | 0.046 | 0.047 | 0.048 | 0.046 | 0.051 | 0.045 | 0.046 | 0.050 |
| | 2 | 0.049 | 0.047 | 0.046 | 0.048 | 0.047 | 0.043 | 0.045 | 0.050 | 0.047 |
| | 3 | 0.046 | 0.047 | 0.040 | 0.046 | 0.046 | 0.046 | 0.045 | 0.050 | 0.051 |
| 3 | 0.33 | 0.049 | 0.036 | 0.043 | 0.048 | 0.049 | 0.045 | 0.054 | 0.044 | 0.042 |
| | 0.5 | 0.047 | 0.048 | 0.042 | 0.046 | 0.042 | 0.050 | 0.048 | 0.042 | 0.044 |
| | 1 | 0.047 | 0.047 | 0.052 | 0.048 | 0.050 | 0.056 | 0.047 | 0.043 | 0.054 |
| | 2 | 0.051 | 0.044 | 0.050 | 0.045 | 0.044 | 0.045 | 0.050 | 0.047 | 0.050 |
| | 3 | 0.051 | 0.042 | 0.043 | 0.045 | 0.046 | 0.043 | 0.054 | 0.041 | 0.041 |

Note: skew denotes the skewness of the population distributions; n1/n2 indicates the ratio of sample sizes; N denotes the total sample size of both samples combined; Mean Diff.=standardized mean difference between population distributions.

**Table 2.** Type I Error Rates for the Nonparametric Levene Test (NPL)

| | | Mean Diff. = 0 | | | Mean Diff. = 1 | | | Mean Diff. = 2 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Skew** | **n1/n2** | **N=24** | **N=48** | **N=96** | **N=24** | **N=48** | **N=96** | **N=24** | **N=48** | **N=96** |
| 0 | 0.33 | 0.049 | 0.051 | 0.054 | 0.038 | 0.072 | **0.122** | **0.302** | **0.622** | **0.876** |
| | 0.5 | 0.049 | 0.053 | 0.052 | 0.014 | 0.025 | 0.038 | 0.070 | **0.295** | **0.668** |
| | 1 | 0.047 | 0.044 | 0.043 | 0.013 | 0.012 | 0.008 | 0.000 | 0.000 | 0.000 |
| | 2 | 0.049 | 0.047 | 0.047 | 0.015 | 0.032 | 0.070 | 0.064 | **0.308** | **0.767** |
| | 3 | 0.047 | 0.052 | 0.046 | 0.037 | **0.092** | **0.162** | **0.290** | **0.654** | **0.921** |
| 1 | 0.33 | 0.047 | 0.053 | 0.050 | 0.068 | 0.071 | **0.079** | **0.430** | **0.499** | **0.669** |
| | 0.5 | 0.048 | 0.045 | 0.048 | 0.037 | 0.044 | 0.066 | **0.157** | **0.300** | **0.470** |
| | 1 | 0.047 | 0.047 | 0.049 | 0.043 | 0.074 | **0.183** | 0.003 | 0.003 | 0.002 |
| | 2 | 0.048 | 0.048 | 0.048 | 0.048 | **0.199** | **0.595** | 0.066 | **0.643** | **1.000** |
| | 3 | 0.047 | 0.054 | 0.043 | **0.083** | **0.313** | **0.717** | **0.347** | **0.953** | **1.000** |
| 2 | 0.33 | 0.053 | 0.055 | 0.055 | **0.178** | **0.183** | **0.208** | **0.522** | **0.497** | **0.650** |
| | 0.5 | 0.049 | 0.050 | 0.054 | **0.109** | **0.152** | **0.251** | **0.337** | **0.323** | **0.438** |
| | 1 | 0.045 | 0.047 | 0.046 | **0.098** | **0.253** | **0.597** | 0.011 | 0.010 | 0.012 |
| | 2 | 0.053 | 0.052 | 0.053 | **0.132** | **0.627** | **0.997** | **0.130** | **0.905** | **1.000** |
| | 3 | 0.060 | 0.055 | 0.054 | **0.182** | **0.807** | **1.000** | **0.511** | **0.997** | **1.000** |
| 3 | 0.33 | 0.048 | 0.053 | 0.049 | **0.372** | **0.301** | **0.310** | **0.630** | **0.496** | **0.673** |
| | 0.5 | 0.044 | 0.045 | 0.044 | **0.267** | **0.233** | **0.266** | **0.512** | **0.354** | **0.451** |
| | 1 | 0.043 | 0.046 | 0.048 | **0.132** | **0.238** | **0.500** | 0.014 | 0.013 | 0.029 |
| | 2 | 0.053 | 0.042 | 0.045 | **0.173** | **0.908** | **1.000** | **0.261** | **0.989** | **1.000** |
| | 3 | 0.054 | 0.053 | 0.054 | **0.392** | **0.985** | **1.000** | **0.746** | **1.000** | **1.000** |

Note: skew denotes the skewness of the population distributions; n1/n2 indicates the ratio of sample sizes; N denotes the total sample size of both samples combined; Mean Diff.=standardized mean difference between population distributions. Observed Type I error rates greater than 0.075 are indicated in bold.

population mean differences remained at or near the 0.05 level and never were above the 0.075 criterion for these conditions. As the population mean differences increased, Type I error rates became inflated above the nominal 0.05 level as the difference in population means increased, the level of skewness increased, and sample size increased. In some cases, the false positive rate (i.e., Type I error rate) became 100% in the conditions with the largest total sample size (*N*=96), most unequal sample size ratios, largest mean differences, and largest skewness. Even in conditions with more modest mean differences (mean difference of 1) and skewness (skewness of 1), the Type I error rate was 0.717 when the total sample size was N=96 and the sample size ratio was $n_1/n_2 = 3$ (i.e., $n_1 = 24$ and $n_2 = 72$). These results suggest the NPL test is not robust to differences in population means, and the inflation in Type I error rates can increase as the population distributions become less symmetric.

There was one important exception to this pattern of increased Type I error rates for the NPL test. When the sample sizes were equal (sample size ratio of 1), larger mean differences and skewness could lead to Type I error rates either much higher or much lower than the nominal 0.05 level. We anticipated that equal sample sizes would lead to decreased Type I error rates, and when mean differences were very large (mean difference of 2) the Type I error rates for the equal sample size conditions were well below the nominal 0.05 level. However, when the population mean difference was 1 standard deviation, Type I error rates were inflated above the 0.075 criterion in some cases when skewness was 1, and in all cases when skewness was greater than 1. We hypothesize that this result is due to the

proportion of overlap between the two populations. This result highlights that although Type I error rates tend to increase as differences in population means and skewness increase, there can be a complex interplay between the magnitude of the mean differences, skewness and relative sample sizes.

Tables 3 and 4 show results for the NPL test combined with either mean (NPL-MEAN) or median (NPL-MED) centering. The results indicate that although the NPL-MEAN and NPL-MED test were robust to population mean differences when the populations were symmetric (the first six rows of each table), Type I error rates became inflated as the populations became more skewed. When there was no difference in the population means, using mean or median centering with the NPL test could lead to inflated Type I error rates if the populations were skewed. Hence, although the mean and median centering improved Type I error rates for symmetric populations, they lead to worse results when sampling from skewed populations. The NPL-MED test generally had lower Type I error rates than the NPL-MEAN test, and observed rates were only slightly above the 0.075 criterion when skewness was less than or equal to 1.

## Summary and Take-Home Messages

The primary findings and messages from this study are described in the points below. First, it is clear that population mean differences can create complications when testing for differences of scale using the NPL test. As anticipated, the simulations demonstrated that Type I error rates for the NPL test can become either too high or too low when there are population mean differences, depending upon whether populations are symmetric or asymmetric, and whether sample sizes are equal or unequal. In general, as the differences between population means and the level of skewness in the populations increased, the NPL test tended to have more inflated Type I error rates, meaning researchers would be more likely to incorrectly conclude that the

**Table 3.** Type I Error Rates for the Nonparametric Levene Test with Mean-Centering (NPL-MEAN)

| | | Mean Diff. = 0 | | | Mean Diff. = 1 | | | Mean Diff. = 2 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Skew** | **n1/n2** | **N=24** | **N=48** | **N=96** | **N=24** | **N=48** | **N=96** | **N=24** | **N=48** | **N=96** |
| 0 | 0.33 | 0.051 | 0.056 | 0.060 | 0.054 | 0.051 | 0.060 | 0.055 | 0.052 | 0.051 |
| | 0.5 | 0.059 | 0.056 | 0.060 | 0.053 | 0.055 | 0.060 | 0.054 | 0.051 | 0.059 |
| | 1 | 0.050 | 0.047 | 0.054 | 0.054 | 0.053 | 0.059 | 0.059 | 0.056 | 0.050 |
| | 2 | 0.050 | 0.051 | 0.058 | 0.054 | 0.054 | 0.061 | 0.059 | 0.056 | 0.050 |
| | 3 | 0.060 | 0.050 | 0.062 | 0.056 | 0.052 | 0.062 | 0.057 | 0.053 | 0.060 |
| 1 | 0.33 | **0.098** | **0.089** | **0.089** | **0.098** | **0.100** | **0.099** | **0.093** | **0.097** | **0.098** |
| | 0.5 | **0.088** | **0.089** | **0.096** | **0.094** | **0.089** | **0.095** | **0.092** | **0.097** | **0.088** |
| | 1 | **0.093** | **0.090** | **0.094** | **0.096** | **0.097** | **0.087** | **0.092** | **0.093** | **0.093** |
| | 2 | **0.089** | **0.091** | **0.097** | **0.092** | **0.094** | **0.095** | **0.098** | **0.094** | **0.089** |
| | 3 | **0.099** | **0.082** | **0.094** | **0.100** | **0.098** | **0.088** | **0.090** | **0.096** | **0.099** |
| 2 | 0.33 | **0.251** | **0.274** | **0.205** | **0.249** | **0.289** | **0.212** | **0.242** | **0.270** | **0.251** |
| | 0.5 | **0.248** | **0.280** | **0.212** | **0.259** | **0.274** | **0.218** | **0.261** | **0.284** | **0.248** |
| | 1 | **0.251** | **0.284** | **0.213** | **0.255** | **0.297** | **0.223** | **0.260** | **0.281** | **0.251** |
| | 2 | **0.258** | **0.281** | **0.220** | **0.247** | **0.277** | **0.209** | **0.256** | **0.285** | **0.258** |
| | 3 | **0.250** | **0.267** | **0.203** | **0.245** | **0.279** | **0.209** | **0.252** | **0.275** | **0.250** |
| 3 | 0.33 | **0.620** | **0.723** | **0.490** | **0.640** | **0.728** | **0.487** | **0.632** | **0.736** | **0.620** |
| | 0.5 | **0.656** | **0.728** | **0.504** | **0.668** | **0.736** | **0.511** | **0.656** | **0.730** | **0.656** |
| | 1 | **0.686** | **0.759** | **0.531** | **0.688** | **0.754** | **0.535** | **0.682** | **0.751** | **0.686** |
| | 2 | **0.653** | **0.751** | **0.505** | **0.652** | **0.740** | **0.527** | **0.668** | **0.747** | **0.653** |
| | 3 | **0.629** | **0.725** | **0.470** | **0.627** | **0.723** | **0.469** | **0.629** | **0.713** | **0.629** |

*Note*: skew denotes the skewness of the population distributions; n1/n2 indicates the ratio of sample sizes; N denotes the total sample size of both samples combined; Mean Diff.=standardized mean difference between population distributions. Observed Type I error rates greater than 0.075 are indicated in bold.

**Table 4.** Type I Error Rates for the Nonparametric Levene Test with Median-Centering (NPL-MED)

| | | Mean Diff. = 0 | | | Mean Diff. = 1 | | | Mean Diff. = 2 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Skew** | **n1/n2** | **N=24** | **N=48** | **N=96** | **N=24** | **N=48** | **N=96** | **N=24** | **N=48** | **N=96** |
| 0 | 0.33 | 0.058 | 0.045 | 0.050 | 0.057 | 0.051 | 0.046 | 0.054 | 0.043 | 0.048 |
| | 0.5 | 0.057 | 0.049 | 0.050 | 0.054 | 0.046 | 0.048 | 0.050 | 0.047 | 0.048 |
| | 1 | 0.044 | 0.042 | 0.043 | 0.044 | 0.039 | 0.048 | 0.047 | 0.049 | 0.051 |
| | 2 | 0.053 | 0.044 | 0.045 | 0.049 | 0.045 | 0.052 | 0.050 | 0.050 | 0.051 |
| | 3 | 0.059 | 0.049 | 0.047 | 0.060 | 0.051 | 0.046 | 0.054 | 0.054 | 0.046 |
| 1 | 0.33 | **0.077** | 0.074 | 0.070 | 0.070 | 0.072 | **0.079** | **0.080** | 0.075 | **0.078** |
| | 0.5 | 0.072 | 0.063 | **0.078** | 0.071 | 0.071 | 0.071 | 0.075 | 0.074 | **0.083** |
| | 1 | 0.068 | 0.067 | **0.076** | 0.074 | 0.074 | **0.076** | 0.063 | 0.068 | **0.077** |
| | 2 | 0.075 | 0.065 | 0.071 | **0.076** | 0.066 | **0.077** | 0.072 | 0.073 | 0.073 |
| | 3 | **0.076** | 0.067 | 0.064 | 0.072 | 0.075 | **0.077** | 0.073 | 0.072 | **0.077** |
| 2 | 0.33 | **0.135** | **0.151** | **0.190** | **0.136** | **0.153** | **0.197** | **0.143** | **0.152** | **0.184** |
| | 0.5 | **0.136** | **0.160** | **0.199** | **0.137** | **0.166** | **0.201** | **0.134** | **0.159** | **0.202** |
| | 1 | **0.129** | **0.164** | **0.208** | **0.132** | **0.167** | **0.210** | **0.133** | **0.176** | **0.216** |
| | 2 | **0.137** | **0.166** | **0.205** | **0.134** | **0.154** | **0.196** | **0.131** | **0.164** | **0.201** |
| | 3 | **0.145** | **0.157** | **0.196** | **0.137** | **0.147** | **0.187** | **0.138** | **0.158** | **0.192** |
| 3 | 0.33 | **0.333** | **0.399** | **0.540** | **0.316** | **0.412** | **0.538** | **0.323** | **0.404** | **0.542** |
| | 0.5 | **0.333** | **0.448** | **0.563** | **0.331** | **0.466** | **0.570** | **0.332** | **0.451** | **0.569** |
| | 1 | **0.345** | **0.477** | **0.587** | **0.325** | **0.474** | **0.603** | **0.337** | **0.486** | **0.597** |
| | 2 | **0.335** | **0.455** | **0.577** | **0.326** | **0.454** | **0.582** | **0.348** | **0.458** | **0.574** |
| | 3 | **0.317** | **0.410** | **0.546** | **0.305** | **0.412** | **0.548** | **0.318** | **0.406** | **0.539** |

Note: skew denotes the skewness of the population distributions; n1/n2 indicates the ratio of sample sizes; N denotes the total sample size of both samples combined; Mean Diff.=standardized mean difference between population distributions. Observed Type I error rates greater than 0.075 are indicated in bold

population variances differ. An exception to this occurred when sample sizes were equal, in which case Type I error rates first increased and then decreased as mean differences increased.

Centering scores using sample means or sample medians improved Type I error accuracy when populations were symmetric, but lead to additional incorrect (inflated) Type I error rates when the populations were skewed; when sampling from asymmetric populations with either equal or unequal population means, both the NPL-MEAN and NPL-MED tests had inflated Type I error rates. The NPL-MED test appeared to have slightly lower (though still inflated) Type I error rates, suggesting this may be the more promising version to evaluate in future research. The LevMED test maintained correct Type I error rates for all conditions studied. These results are consistent with those found in earlier studies examining nonparametric tests of scale (e.g., Olejnik & Algina, 1987).

As with any simulation study, the results cannot necessarily be generalized to conditions beyond those included in our experimental design. However, because the conditions represent scenarios that could be encountered in practice (i.e., unknown and unequal means and skewed distributions), and because the results indicated that the tests were not robust for many of these conditions, we urge caution when using the NPL test under similar conditions. Future studies could examine whether the same problems occur with additional types of distributions, and when comparing more than two populations. Lim and Loh (1996) found that using bootstrapped critical values improved the power of the LevMED test, and future research could explore whether bootstrapped critical values or other resampling techniques could be used to improve the Type I error

rates of the NPL tests when population means are unequal and unknown. To our knowledge, a bootstrap version of the NPL test has not been described or evaluated in prior studies.

From a statistical perspective, this study highlights the fact that the NPL test can be sensitive both to differences in population means and to differences in population variances. The original NPL test was explicitly intended to test the null hypothesis that samples are drawn from populations with equal variances, but it implicitly assumes the populations have equal means. As a result, the NPL test is also sensitive to differences in population means. A similar issue is discussed by Nordstokke and Colp (2018), who note that many nonparametric tests of location make implicit assumptions about the relative shapes of the distributions being compared, and can be sensitive to differences in distributional shape as well as location. Differences between population means are a *nuisance parameter* when testing for the equality of population variances. As a result, many discussions about tests of scale do not address mean differences, and vice versa. In the (generalized) Behrens-Fisher case, when population means can be assumed equal, the *nuisance parameter* may not be a major concern. But when population means are unknown and cannot be assumed equal, the *nuisance parameter* can cause problems for the NPL test as demonstrated in this study. As noted above, there are some contexts in which it may be reasonable to assume that population means are equal, such as experimental settings where participants are drawn from a single population. If there is the possibility of heterogeneous treatment effects in these contexts, however, this could lead to a difference in means and variances for post-treatment outcomes, and the problem may also be relevant in these experimental settings.

What are the practical implications of these results? Unfortunately, researchers will rarely know ahead of time whether the means of the populations being compared truly differ or not. This is, after all, the primary reason for using statistical tests. We recommend that researchers and data analysts use graphical and descriptive summary statistics to evaluate the plausibility of the assumption that population means are equal prior to applying the NPL test. These preliminary analyses can also be used to gain information about the relative shapes of the distributions being compared (e.g., whether they are skewed or symmetric). If the population means appear potentially unequal, then greater attention to factors such as the relative sample sizes and shape of the distributions is necessary when using the NPL test. If the distributions appear largely symmetric (i.e., not skewed), using mean or median centering versions of the NPL test (the NPL-MEAN or NPL-MED tests) may provide a more accurate comparison of the variances. If the population means cannot safely be assumed equal and there is evidence that the distributions are asymmetric, it may be more appropriate to use the LevMED test. While the LevMED test was robust in all cases studied here, it has been previously shown to have lower power than the NPL test under conditions of non-normality. In sum, although none of the tests compared here appear to be optimal under all conditions, we hope the results will help researchers understand the relevant aspects of their data when determining the best approach for testing hypotheses about the equality of population variances.

## References

Boos, D. D., & Brownie, C. (2004). Comparing variances and other measures of dispersion. *Statistical Science*, 19(4), 571–578. https://doi.org/10.1214/088342304000000503

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31(2), 144–152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x

Brown, M. B., & Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, 69(346), 364–367. https://doi.org/10.1080/01621459.1974.10482955

Conover, W. J., Johnson, M. E., & Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, 23(4), 351–361. https://doi.org/10.1080/00401706.1981.10487680

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Levene, H. (1960). Robust test for equality of variances. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling* (pp. 278–292). Stanford, CA: Stanford University Press.

Lim, T.-S., & Loh, W.-Y. (1996). A comparison of tests of equality of variances. *Computational Statistics & Data Analysis*, 22(3), 287–301. https://doi.org/10.1016/0167-9473(95)00054-2

Nordstokke, D. W., & Colp, S. M. (2014). Investigating the robustness of the nonparametric Levene test with more than two groups. *Psicológica*, 35(2), 361–383.

Nordstokke, D. W., & Colp, S. M. (2018). A note on the assumption of identical distributions for nonparametric tests of location. *Practical Assessment, Research & Evaluation*, 23(3), 1–9.

Nordstokke, D. W., & Zumbo, B. D. (2010). A new nonparametric Levene test for equal variances. *Psicológica*, 31(2), 401–430.

Nordstokke, D. W., Zumbo, B. D., Cairns, S. L., & Saklofske, D. H. (2011). The operating characteristics of the nonparametric Levene test for equal variances with assessment and evaluation data. *Practical Assessment, Research & Evaluation*, 16(5), 1–8.

Olejnik, S. F., & Algina, J. (1987). Type I error rates and power estimates of selected parametric and nonparametric tests of scale. *Journal of Educational and Behavioral Statistics*, 12(1), 45–61. https://doi.org/10.3102/10769986012001045

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Scheffé, H. (1970). Practical solutions of the Behrens-Fisher problem. *Journal of the American Statistical Association*, 65(332), 1501–1508. https://doi.org/10.1080/01621459.1970.10481179

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54(8), 594–604. https://doi.org/10.1037/0003-066X.54.8.594

Zumbo, B. D., & Coulombe, D. (1997). Investigation of the robust rank-order test for non-normal populations with unequal variances: The case of reaction time. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 51(2), 139–150. https://doi.org/10.1037/1196-1961.51.2.139

## Citation:

Shear, Benjamin R., Nordstokke, David W., & Zumbo, Bruno D. (2018) A Note on Using the Nonparametric Levene Test When Population Means Are Unequal. *Practical Assessment, Research & Evaluation*, 23(13). Available online: http://pareonline.net/getvn.asp?v=23&n=13

## Corresponding Author

Benjamin R. Shear
University of Colorado Boulder
UCB 249
Boulder, CO 80309
email: benjamin.shear [at] colorado.edu