Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 16, Number 5, February 2011 ISSN 1531-7714

The operating characteristics of the nonparametric Levene test for equal variances with assessment and evaluation data

David W. Nordstokke, *University of Calgary*
Bruno D. Zumbo, *University of British Columbia*
Sharon L. Cairns, *University of Calgary*
Donald H. Saklofske, *University of Calgary*

Many assessment and evaluation studies use statistical hypothesis tests, such as the independent samples t test or analysis of variance, to test the equality of two or more means for gender, age groups, cultures or language group comparisons. In addition, some, but far fewer, studies compare variability across these same groups or research conditions. Tests of the equality of variances can therefore be used on their own for this purpose but they are most often used alongside other methods to support assumptions made about variances. This is often done so that variances can be pooled across groups to yield an estimate of variance that is used in the standard error of the statistic in question. The purposes of this paper are twofold. The first purpose is to describe a new nonparametric Levene test for equal variances that can be used with widely available statistical software such as SPSS or SAS, and the second purpose is to investigate this test's operating characteristics, Type I error and statistical power, with real assessment and evaluation data. To date, the operating characteristics of the nonparametric Levene test have been studied with mathematical distributions in computer experiments and, although that information is valuable, this study will be an important next step in documenting both the level of non-normality (skewness and kurtosis) of real assessment and evaluation data, and how this new statistical test operates in these conditions.

When conducting assessments or evaluations in the social, psychological or educational context it is often required that groups be compared on some construct or variable such as math achievement or emotional intelligence. Nordstokke & Zumbo (2007, 2010) remind us that when conducting these comparisons, typically using means or medians, we must be cognizant of the assumptions that are required for validly making comparisons between groups. It was highlighted by those authors that the assumption of homogeneity of variances is of key importance and must be considered prior to conducting these tests.

The assumption of equality of variances is based on the premise that the population variances on the variable being analyzed for each group are equal. The assumption of homogeneity of variances is essential when comparing two groups, because if variances are unequal, the validity of the results are jeopardized (i.e., increased Type I error rates leading to invalid inferences) (Glass et al., 1972). There are at least three possible occasions where testing for equality of variances are a concern. The first is when one wants to make inferences about population variances because they are of scientific interest on their own. For example, a health researcher may be interested in studying the effects of a new drug that helps prevent mood swings on some members of a mood management program. The researcher hypothesizes that the drug will decrease the severity of mood swings in patients. In this case, the researcher is interested in the overall increase or decrease in the

severity of mood swings (operationalized as the change in the range of mood scores from high extremes to low extremes to more moderate shifts in mood) in which case a test for equal variances would be conducted to test for differences. This is needed because those in the group that received the program would be hypothesized to have less severity in the range of their scores. The second is when there is suspected heterogeneity of variances in a t-test or an analysis of variance (ANOVA). A researcher is interested in spatial ability and uses a categorical variable, such as gender, as a grouping variable in a t-test. It cannot be assumed that males and females vary equally on spatial ability so, prior to the t-test; a test for equal variances must be carried out. A third occasion when one might be concerned about heterogeneity of variances is in a t-test or ANOVA in which the numbers of observations in the groups are widely disparate (Glass, 1966; Glass, Peckham, & Sanders, 1972). When there is reasonable evidence suggesting that the variances of two or more groups are unequal, a preliminary test of equal variances is conducted prior to conducting the t-test or ANOVA.

It cannot necessarily be assumed that groups of participants are homogeneous or exchangeable, and so there is no basis to assume equality of variances when testing the null hypothesis of no difference between two or more groups. Furthermore, if this assumption is ignored, the results of the statistical test (i.e., t-test and ANOVA) are greatly distorted leading to incorrect inferences based on the results. Of note is that nonparametric tests are also susceptible to issues of unequal variances when testing for equal medians (Harwell, Rubinstein, Hayes, & Olds, 1992; Zimmerman & Zumbo, 1993a; 1993b), thus switching to a nonparametric statistical approach does not alleviate the problem of unequal variances.

When testing for equal variances between groups, a problem arises when samples are collected from populations that result in skewed data. Data can become skewed because there are extreme scores in one end of the distribution resulting in an asymmetrically shaped distribution. In fact, it can be argued that, in many cases, data commonly collected in educational, behavioral and health research do not meet the assumption of normality or symmetry (Bradley, 1977; Micceri, 1989).

Many of the current tests of equality of variances that are widely recommended such as Levene's test for equality of variances based on means are founded on the assumption of symmetric distributions (e.g., normality). It has been demonstrated using computer simulation that violations of symmetry increase the Type I error rate of the Levene test (e.g., Shoemaker, 2003; Zimmerman, 2004). Further, statistical researchers have investigated other approaches to testing for equality of variances. Conover, Johnson, and Johnson (1984) reviewed these approaches and provided simulation results investigating their performance under various conditions of violating their assumptions. They investigated the robustness of 56 tests for equal variances, and demonstrated that the median based Levene test (Brown and Forsythe, 1974) is the most valid in terms of maintenance of its nominal Type I error rate and average power values[1]. However, to this point, there is no consensus amongst methodologists and researchers regarding what the "gold standard" or test of choice is when the assumption of symmetry of distribution and particularly normality has been violated.

A newly developed test for equality of variances, the nonparametric Levene test, which utilizes the *method of ranks* (Friedman, 1937) has demonstrated its robustness of validity through maintenance of its nominal Type I error and its statistical power via a series of simulations (Nordstokke & Zumbo, 2010). In their study, the newly developed nonparametric Levene test was compared to the median based Levene test across a large number of conditions that varied in terms of its degrees of distributional symmetry, unequal sample sizes, and overall sample size. The nonparametric Levene test outperformed the median test consistently when the population distributions that were being sampled were asymmetric to varying degrees. As Nordstokke and Zumbo (2010) describe it, the nonparametric Levene test involves pooling the data from both groups, in the two group situation, ranking the scores, placing the rank values back into their original groups, and conducting the Levene test on the ranks. This test can be defined as,

$$\text{ANOVA}\left(\left|R_{ij} - \overline{X}_j\right|\right),$$

wherein R_{ij} is calculated by pooling the values from each of the (j) groups and ranking the scores. An analysis of variance is conducted on the absolute value of the mean of the ranks for each group ($\overline{X}_j$) subtracted from each

[1] Average power is defined as the power of the test averaged across a number of simulation conditions. See Conover, Johnson, and Johnson (1984).

individual's rank (R_{ij}). From a computational point of view, this nonparametric Levene test uses Conover and Iman's (1981) notion of the rank transformation as a bridge between parametric and nonparametric statistics and simply involves (i) pooling the data and replacing the original scores by their ranks and then (ii) separating the data back into their groups and (iii) applying the conventional mean-based Levene test to the ranks. This can be easily accomplished using widely available software such as SPSS or SAS. When the data are extremely non-normal, perhaps caused by several outliers or some other intervening variables, the transformation changes the distribution and makes it uniform. Conover and Iman (1981) suggested conducting parametric analyses such as the analysis of variance on rank transformed data. Rank transformations are appropriate for simple tests of equal variances because, if the rankings between the two groups are widely disparate, it will be reflected by a significant result. For example, if the ranks of one of the groups tend to have values whose ranks are clustered near the top and bottom of the distribution and the other group has values whose ranks cluster near the middle of the distribution, the result of the nonparametric Levene test would lead one to conclude that the variances are not homogeneous. Thus the nonparametric Levene test is, essentially, a parametric analysis of variance conducted on rank transformed data.

The next logical step for the development of the nonparametric Levene test is to investigate its validity on "real-world" assessment and evaluation data; therefore, the purpose of this paper is to investigate and demonstrate the performance of the nonparametric Levene test using assessment and evaluation data.

Methods

Data

Data for this simulation study were gathered from two sources. All simulations were conducted using SPSS. The first data source (data set #1) (n = 4,600) came from an evaluation study that was conducted at the University of Calgary Counselling Center, Calgary, AB, Canada. The variable from the evaluation study was age (i.e., number of years old). The skew of the population distribution was 2.051, the kurtosis was 6.27. This data sources was continuous in nature. The second data source (n = 9,200) was from the Canadian Broadcasting Corporation "Test the Nation", a nationwide televised program that measured the cognitive functioning of the participants. Three subscale scores, calculated for this paper, consisted of a Language scale score (data set #2, 6 items) (skew = .13, kurtosis = .61), a Math scale score (data set #3, 5 items) (skew = -.47, kurtosis = 1.29), and a Memory scale score (data set #4, 5 items) (skew = -1.17, kurtosis = 5.31). These three subscale scores were calculated by combining the responses to several Likert scaled items to yield the scale score. A fourth demographic variable (data set #5) was selected from this data set asking participants to report the number of pairs of shoes they owned (skew = 1.35, kurtosis = .59). The shapes of the population distributions are illustrated in Figure 1.

Variance ratios

Five levels of variance ratios (var1/var2) are utilized in this design. The first level (1/1) represents the case where variances are equal and the Type I error rates for the nonparametric Levene are investigated. The other levels (3/1, 2/1, 1/2 and 1/3) represent the instances in the design where the variances are unequal and the statistical power of the nonparametric Levene test is investigated. The design was created so that there were direct pairing and inverse pairing in relation to unbalanced groups and direction of variance imbalance. Direct pairing occurs when the larger sample sizes are paired with the larger variance and inverse pairing occurs when the smaller sample size is paired with the larger variance (Tomarken & Serlin, 1986). This was done to investigate a more complete range of data possibilities. In addition, Keyes and Levy (1997) drew our attention to concern with unequal sample sizes, particularly in the case of factorial designs – see also O'Brien (1978, 1979) for discussion of (the original versions of) Levene's test in additive models for variances. Findings suggest that the validity and efficiency of a statistical test is somewhat dependent on the direction of the pairing of sample sizes with the ratio of variance (Tomarken & Serlin, 1986).

Simulation

Each of the five data sets selected from the two data sources was treated as a population. For each of the data sources, a 3x4 completely crossed design was utilized with three levels of sample size ratios (n1/n2: 1/1, 2/1, and 3/1) and five levels of variance ratios (var1/var2: 1/3, 1/2, 1/1, 2/1, and 3/1). Each population was exhaustively randomly sampled into sets of 40,

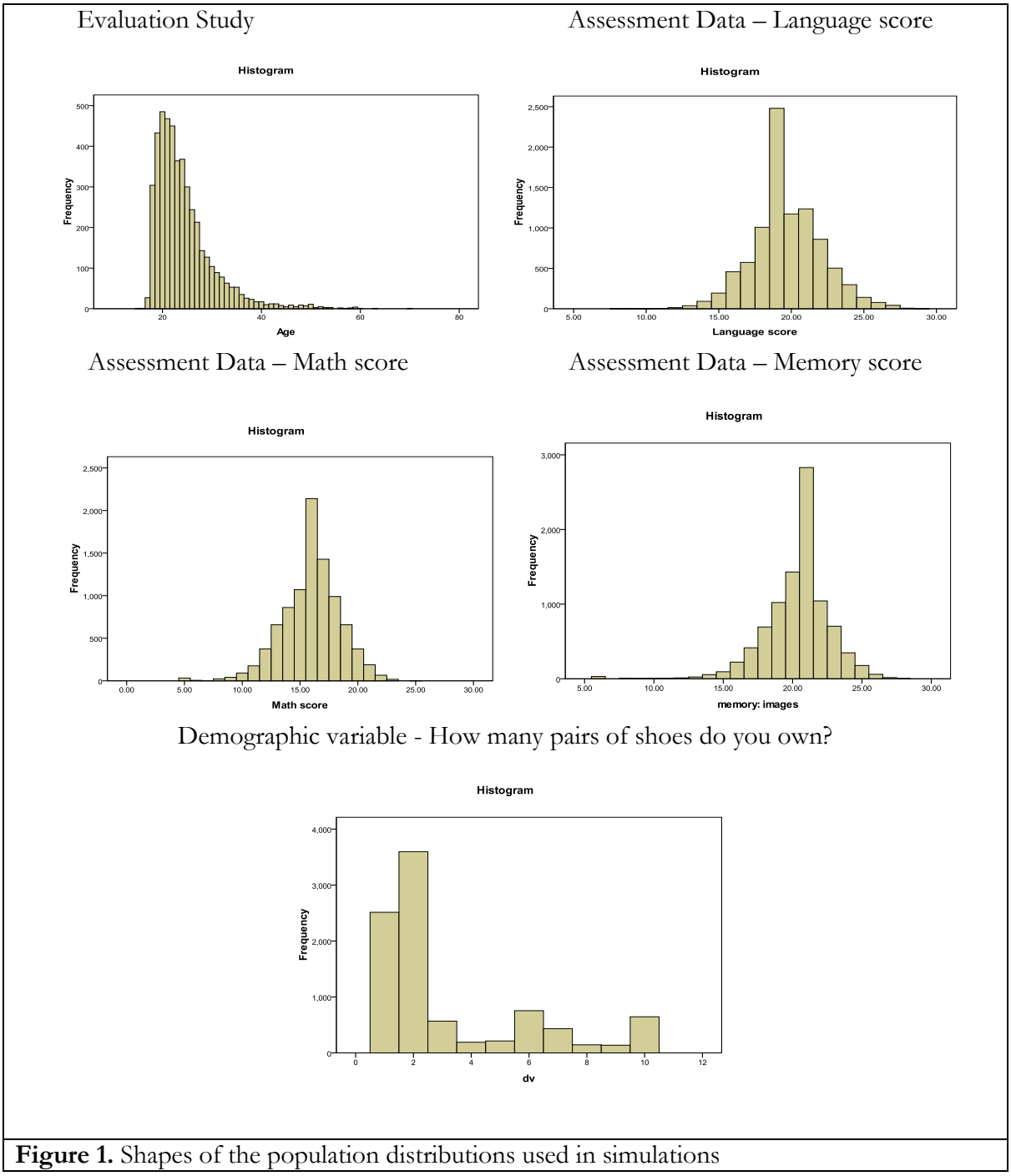

Figure 1. Shapes of the population distributions used in simulations

and members of each of the sets were further randomly assigned into two groups. This resulted in 115 sets of grouped data, which will be henceforth called replications, for the first data source and 230 replications for the second data source. Each replication (involving two groups of 20) was entered into the simulation. What follows are the steps involved in conducting the simulation in one cell of the design where the sample size ratio (n1/n2) is 1/1 and the variance ratio (var1/var2) is 1/1. Once the population has been exhaustively sampled to create the grouped data, the mean of each of the groups are centered and the variance for each group was manipulated to one of the ratios outlined in the design. In this case the ratio is 1/1, so the variance of both of the groups was unchanged. Next, for each of the pairs of groups in the set, the scores for the two groups are pooled, ranked in ascending order, split back into their original groups, and then an

independent samples t-test is then performed on the ranked data of the two groups. A Levene's test for equality of variances is reported in this procedure as a default test to determine if the variances are statistically significantly different at the nominal alpha value of .05. The frequency of Type I errors was tabulated for each cell in the design.

The criteria for maintaining the Type I error rate of the nonparametric Levene test is .05 (±.025), this was considered to be liberal criteria according to Bradley, (1978); however, it should be noted that when Type I error rates are less than .05, the validity of the test is not jeopardized in the same way as they are when they are inflated. This makes a test invalid if Type I errors are inflated; but when the Type I error rate decreases the test becomes more conservative potentially reducing its power. Reducing power does not invalidate the results of a test, per se, so tests will be considered to be invalid only if the Type I error rate is inflated. In the cells where the ratio of variances was not equal, and that maintained their Type I error rates, statistical power is represented by the proportion of times that the nonparametric Levene's test correctly rejected the null hypothesis. Type I error rates and power are often represented as percentages. For example, if the nominal alpha is .05, that means that 5 percent of the time the test will reject the null hypothesis when it should not be rejected; and the power of the test may be .20, meaning that 20 percent of the time the test will be powerful enough to detect real differences between groups. In all cases in the present study, Type I error rates and statistical power values are converted from proportions to percentages.

Results

For the first data set (i.e., the age data from the evaluation study), the Type I error rates and the statistical power for the nonparametric Levene test is presented in Table 1. The rows of Table 1 represent the ratio of sample sizes (i.e., n1/n2), which are 1/1, 2/1 and 3/1. The columns represent the ratio of variances (i.e., var1/var2) for each of the cells of the design. In the column where the ratio of variances is 1/1, the Type I error rates for the nonparametric Levene are shown, and when the ratio of variances is unequal (e.g., 2/1), the statistical power is represented. For example, when the sample sizes are equal (1/1) and the variances are equal (1/1), the Type I error rate for the nonparametric Levene test is 3.5%. The nonparametric Levene test maintained its Type I error rate in each cell of Table 1. To give an example of a power value, when the sample sizes are (2/1) and the ratio of variances are directly paired with the sample size ratio (2/1) the statistical power of the nonparametric Levene test is 77.4%. Overall for the results in Table 1, the Type I error rates ranged between 3.5% and 5.2%, and the power values ranged between 44.3% and 93%.

Table 1. Type I error rates and power for nonparametric Levene test on Age data.

		Variance ratio (var1/var2)				
		Direct pairing			Inverse pairing	
		3/1	2/1	1/1	1/2	1/3
sample size ratio (n1/n2)	1/1	92.2	67.8	3.5	67.8	92.2
	2/1	94.8	77.4	5.2	59.1	82.6
	3/1	93.0	71.3	5.2	44.3	69.6

For the second data set (i.e., Language scale score for the CBC Test the Nation data), the Type I error rates and statistical power are presented in Table 2. All of the Type I error rates are within the criteria for validity ranging between 4.3% and 7%, for example, in the cell of the design where the sample sizes were equal (1/1) and the variance ratios were equal (1/1), the Type I error rate of the nonparametric Levene test was 4.3%. The statistical power values in Table 2 range from 31.7% to 90.4%.

Table 2. Type I error rates and power for nonparametric Levene test on Language score data.

		Variance ratio (var1/var2)				
		Direct pairing			Inverse pairing	
		3/1	2/1	1/1	1/2	1/3
sample size ratio (n1/n2)	1/1	90.4	58.7	4.3	58.7	90.4
	2/1	77.0	43.0	5.2	60.9	90.9
	3/1	64.8	31.7	7.0	50.9	84.8

The simulation results for the third data set (i.e., Math scale score data from the CBC Test the Nation data) are illustrated in Table 3. The Type I error rate for the nonparametric Levene was within the criteria for validity with values ranging from 5.7% to 7.4%. For example, in the cell of the design where the sample sizes were equal (1/1) and the variance ratios were equal (1/1), the Type I error rate of the nonparametric Levene test was 7.4%. The statistical power values in Table 2 range between 28.3% and 73.5%.

For the fourth data set (i.e., Memory scale score for the CBC Test the Nation data), the Type I error rates and statistical power are illustrated in Table 4. The Type I error rates for the nonparametric Levene test were valid in all cells of Table 4. For example, in the cell of the design where the sample sizes were equal (1/1) and the variance ratios were equal (1/1), the Type I error rate of the nonparametric Levene test was 7.5%. In addition, the statistical power values of the nonparametric Levene test ranged between 30.4% and 83.9%.

Table 3. Type I error rates and power for nonparametric Levene test on Math score data.

		Variance ratio (var1/var2)				
		Direct pairing			Inverse pairing	
		3/1	2/1	1/1	1/2	1/3
Sample size ratio (n1/n2)	1/1	73.5	47.4	7.4	47.4	73.5
	2/1	54.8	36.5	5.7	44.3	67.0
	3/1	51.3	28.3	6.5	34.3	69.1

Table 4. Type I error rates and power for nonparametric Levene test on Memory score data.

		Variance ratio (var1/var2)				
		Direct pairing			Inverse pairing	
		3/1	2/1	1/1	1/2	1/3
Sample size ratio (n1/n2)	1/1	83.9	56.1	7.5	56.1	83.9
	2/1	74.3	44.8	7.4	51.7	82.6
	3/1	62.6	30.4	4.8	50.9	81.7

The simulation results for the fifth data set (i.e., Demographic variable from the CBC Test the Nation data) resulted in Type I error rates of 23.9%, 15.2%, and 17.8% for the 1/1, 2/1, and 3/1 sample size ratios, respectively. Clearly, given that the Type I error rate ranged for 15.2% to 23.9% the nonparametric Levene was not within the criteria for validity for the nominal 5% Type I error rate and the statistical power was not reported.

Discussion

It is evident from the simulation results that the nonparametric Levene test for equality of variances overall performs well on real evaluation and assessment data, in terms of maintenance of its nominal Type I error rate and statistical power, when data are sampled from skewed population distributions. However, this result was not consistent across all the cells of the design in the five data sets.

The first data set (i.e., the age data from the evaluation study) provides the best results in terms of maintenance of the nominal Type I error rate and high statistical power values, in some cases reaching as high as 90% power. The reader should note that Cohen (e.g., 1988, 1992), in the absence of any other basis for setting the value for desired power, suggested that 80% be used. We acknowledge that Cohen's criterion is somewhat arbitrary, and dependent on effect size magnitude, but nonetheless it provides us a working criterion and reference point. Therefore, even though these data were quite skewed, by Cohen's standard for statistical power, the nonparametric Levene test performed well by maintaining its nominal Type I error rate and exhibited substantial statistical power. An explanation for the performance of the nonparametric Levene test on this data is related to its continuous nature. When sampling from these data occurred, there were very few ties, and it is has been shown that a large number of tied scores result in the breakdown of rank transformation procedures because of the assumption of continuity central to many nonparametric tests (Bradley, 1968). When data come from a continuous population distribution, the *method of ranks* (Friedman, 1937) is quite efficient because the likelihood of tied scores is small.

The second, third and fourth data sets (language, math, and memory tests, respectively) from the CBC Test the Nation program represented typical assessment data, where several items are scored on a Likert scale and combined to yield a scale score representing a psychological construct. In the cases of the language and memory measures, the nonparametric Levene test maintained its nominal Type I error rates and demonstrated statistical power values of at least 80% in several cases involving a variance ratio of three to one. These results demonstrate the usefulness of the nonparametric Levene test on 'typical' assessment data that is used throughout psychological, social, and educational research. It should be noted, however, that the statistical power for the math test data never reached Cohen's standard of 80% even in the case of a variance ratio of three to one. In addition, two of the data sets used in the simulation (i.e., data set #3 and data set #4, math and memory tests) had distributions with marked outliers – test score values in the range of 5.0 in Figure 1. This demonstrates that the nonparametric Levene test is able to efficiently deal with distributions that not only

deviate from normality in terms of skewness, but also possess marked outliers.

The fifth data set (i.e., Demographic variable from the CBC Test the Nation data) used in this simulation study resulted in elevated Type I error rates making the results of the nonparametric Levene test invalid. This is likely due to the nature of the variable used in this data. The population distribution had many ties, thus, upon sampling the distributions, many ties were present in the sample data. As mentioned earlier, the rank transformation procedure breaks down when there are many ties. It should be noted, that a distribution with a large number of tied scores results in a highly kurtotic distribution that is difficult to analyze using either parametric (i.e., t-test/ANOVA) or nonparametric (Mann-Whitney U/Kruskal-Wallis) techniques due to a lack of variability in the data, so this issue is not just an artifact of the *method of ranks*. In essence, crude scaling and measurement procedures impedes researchers' ability to discern between the 'true' scores of the study participants resulting in tied scores, thus reducing the variability in the data due to imprecise measurement.

Ties are a problem in all rank-based non-parametric statistics because, with ties, the set of N observations does not correspond to the set of ranks 1, 2, 3, ..., N; where N denotes the total sample size. That is, the original scores do not uniquely map on to a set of ranks. In essence, in the case of ties, the particular set of ranks depends on the pattern and number of ties – which in turn depends on the reasons for ties (e.g., population characteristics, appropriateness of the scaling, skewness and kurtosis of the variable). This makes each case of ties somewhat idiosyncratic and a valid and uniformly most powerful test difficult, and at times impossible, to derive. Based on the current knowledge in statistics, our recommendation to practitioners is that if ties are a concern then one should consider calculating the critical values of the test statistic using exact nonparametric tests involving computer intensive methods (e.g., permutation, randomization, bootstrapping or jackknife methods). These methods allow one to calculate the critical value for the test that are tailored to one's particular case of ties; and also, very importantly, allows one to investigate the discrete nature of the sample distribution of the statistic. As Zumbo and Coulombe (1997, p. 141) remind us, discrete sampling distributions, by their very nature, constrain the significance levels that one can reasonably use to test a hypothesis and hence only partially solve the problem of ties. Unfortunately, at the moment, specialized software is needed for this computation. In addition, further research is needed to investigate the operating characteristics of these methods.

To summarize, when data come from non-normal population distributions, the nonparametric Levene test maintains its Type I error rates and possesses moderate to high statistical power for detecting differences in variances. However, when there are a large number of ties present in the data, the ranking procedure is not appropriate for detecting differences in variances.

References

Bradley, J.V. (1968). *Distribution-Free Statistical Tests*. Toronto: Prentice-Hall, Inc.

Bradley, J.V. (1977). A common situation conductive to bizarre distribution shapes. *American Statistician*, *31*(*4*), 147-150.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144–152.

Brown, M.B. & Forsythe, A.B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, *69*(*2*), 364-367.

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences (2nd ed.)*. New Jersey: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *American Statistician, 35,* 124-129.

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.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, *32*, 675-701.

Glass, G.V. (1966). Testing homogeneity of variances. *American Education Research Journal*, *3*(*3*), 187-190.

Glass, G.V., Peckham, P.D., & Sanders, J.R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. *Review of Education Research*, *42*, 237-288.

Harwell, M.R., Rubinstein. E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed

effects ANOVA cases. *Journal of Educational Statistics, 17,* 315-339.

Keyes, T. M., & Levy, M. S. (1997). Analysis of Levene's test under design imbalance. *Journal of Educational and Behavioral Statistics, 22,* 227-236.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156-166.

Nordstokke, D.W. & Zumbo, B.D. (2007). A cautionary tale about Levene's tests for equality of variances. *Journal of Educational Research and Policy Studies*, *7*(*1*), 1-14.

Nordstokke, D.W. & Zumbo, B.D. (2010). A new nonparametric test for equal variances. *Psicologica*, *31*, 401-430.

O'Brien, R. G. (1978). Robust Techniques for Testing Heterogeneity of Variance Effects in Factorial Designs. *Psychometrika, 43*, 327-344.

O'Brien, R. G. (1979). A General ANOVA Method for Robust Tests of Additive Models for Variances. *Journal of the American Statistical Association, 74*, 877-880.

Shoemaker, L. H. (2003). Fixing the *F* Test for Equal Variances. *American Statistician*, *57*(*2*), 105-114.

Tomarken, A.J. & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific non-centrality structures. *Psychological Bulletin, 99*(*1*), 90-99.

Zimmerman, D.W. (2004). A note on preliminary test of equality of variances. *British Journal of Mathematical and Statistical Society*, *57*, 173-181.

Zimmerman, D. W., & Zumbo, B.D. (1993a). Rank transformations and the power of the Student t-test and Welch's t-test for non-normal populations with unequal variances. *Canadian Journal of Experimental Psychology, 47,* 523-539.

Zimmerman, D. W., & Zumbo, B.D. (1993b). The relative power of parametric and nonparametric statistical methods. In G. Keren and C. Lewis (Eds.) *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues* (pp. 481-517). Hillsdale, NJ: Erlbaum.

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*, *51*, 139-150.

Author Note:

An earlier version of this paper was presented at the 7th International Test Commission conference, July 2010, in Hong Kong.

Citation:

Nordstokke, David W., Zumbo, Bruno D., Cairns, Sharon L., Saklofske, Donald H. (2011). The operating characteristics of the nonparametric Levene test for equal variances with assessment and evaluation data. *Practical Assessment, Research & Evaluation*, 16(5). Available online: http://pareonline.net/getvn.asp?v=16&n=5.

Corresponding Authors:

David W. Nordstokke, Ph.D.
Division of Applied Psychology
University of Calgary
2500 University Drive NW
Education Tower 302
Calgary, Alberta,
T2N 1N4

Email: dnordsto [at] ucalgary.ca

Bruno D. Zumbo, Ph.D.
University of British Columbia
Scarfe Building, 2125 Main Mall
Department of ECPS
Vancouver, B.C.
CANADA
V6T 1Z4

e-mail: bruno.zumbo [at] ubc.ca