NORMES

A Cautionary Tale About Levene's Tests for Equal Variances

David W. Nordstokke and Bruno D. Zumbo
University of British Columbia

Abstract
The central messages of this paper are that (a) unequal variances may be more prevalent than typically imagined in educational and policy research, and (b) when considering tests of equal variances one needs to be cautious about what is being referred to as "Levene's test" because Levene's test is actually a family of techniques. Depending on which of the Levene tests that are being implemented, and particularly the Levene test based on means which is found in widely used software like SPSS, one may be using a statistical technique that is as bad (if not worse) than the F test which the Levene test was intended to replace.

Introduction

When comparing groups in educational, social, behavioral, and policy research a common tacit, yet essential, statistical assumption is that the variances of the dependent variable for each group are equal. This assumption is referred to as 'homogeneity of variances' when using statistics like the *t*-test or analysis of variance to compare group means. For example, as is widely seen in educational and policy research, one may use the independent samples *t*-test to compare boys and girls in terms of their average mathematics achievement test scores, and hence one is, sometimes unknowingly, assuming that the boys and girls have equal mathematics score variances.

The matter of unrecognized, or ignored, statistical assumptions and their impact on research practice are exaggerated during what one of the founding editors of this journal, Professor Sean Mulvenon, aptly describes as our "Era of Point-and-Click Statistics," wherein easy to use statistical software often masks and hides the complex statistical assumptions and realities of day-to-day research practice in educational, social, behavioral, and policy studies.

This matter of meeting complex statistical questions and procedures with deceptively easy to use statistical software, and in turn its impact on research practice, is a theme that runs throughout this paper.

Zumbo and Coulombe (1997) remind us that there are, at least, two situations in which one cannot assume equality of variances: (a) when the groups of participants (i.e., subjects or experimental units) are formed by domain differences such as age groups, gender, or educational level, and/or, (b) when the participants (knowingly or unknowingly to the researcher) differ on some important, possibly unmeasured variable. In either situation, one cannot necessarily assume that the 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 means or median – nonparametric tests are also susceptible to issues of unequal variances when testing for equal medians (Harwell, Rubinstein, Hayes, & Olds, 1992; Zimmerman & Zumbo, 1993a; 1993b). It can be easily argued that either of these situations occurs commonly in educational, behavioral, social, and policy research. One then cannot assume equal variances and hence needs to regularly test for equality of variances before testing for equal means (or medians).

Common understanding, as documented in statistical and methodological research papers, textbooks, and codified in widely used statistical software, is that the F test for equality of variances is problematic in terms of its inflated Type I error rate with non-normal population data. As a reminder, the hypothesis for the F test of variances is

$$\begin{aligned} H_0 &: \sigma_1^2 = \sigma_2^2 \\ H_1 &: \sigma_1^2 \neq \sigma_2^2 \end{aligned}. \qquad \text{(H1)}$$

The test statistic to test H_0 against H_1 is

$$F = \frac{s_1^2}{s_2^2}. \qquad \text{(T1)}$$

When the H_0 in (H1) is true, the sampling distribution of $F(\upsilon_1, \upsilon_2)$ from (T1) is the *F*-family of distributions with $\upsilon_1 = n_1 - 1$ and $\upsilon_2 = n_2 - 1$ degrees of freedom, and the sample variances and sample sizes are s_1^2, s_2^2, n_1 and n_2, respectively. The reader should see Glass & Hopkins (1984, p. 263) for a detailed description. It

has been known for over half a century, however, that the test of (H1) by (T1) is notoriously sensitive to and largely invalidated by non-normally distributed (population) dependent variable scores (Box, 1953).

Building on the work of Box, Scheffe, and others, Levene (1960) introduced a methodological approach that was meant to resolve Box's concern for the *F*-test being so sensitive to population non-normality when investigating equality of variance. In short, Levene's approach involves using the usual *F*-test for equality of means computed on what we will refer to as intermediary scores, which one defines as the absolute deviations of the data points from an estimate of the center of the group – i.e., a one-way ANOVA of the centered original data. Levene's original proposal was to compute these intermediary (centered) scores by centering at the sample mean. In short, the original Levene's test involves one conducting a one-way, j-group, ANOVA of the transformed original data, $\left|X_{ij} - \bar{X}_j\right|$, for each i individual in the j groups, where $\bar{X}_j$ denotes the mean of the j^{th} group; and for our purposes, Levene's original test will be denoted as

$$\text{ANOVA}(\left|X_{ij} - \bar{X}_j\right|). \qquad \text{(T2)}$$

The original Levene's test, (T2), was initially found to be quite robust to departures from normality (Levene, 1960). It was this initial finding that drew attention to (T2) as a useful alternative to the *F*-test, (T1). It has, however, been shown using computer simulation that violations of normality increases the Type I error rate of the Levene's test (T2) (e.g., Shoemaker, 2003; Zimmerman, 2004). Carroll and Schneider (1985) showed mathematically that Levene's test involving means, (T2), maintains its nominal Type I error rate only for symmetric distributions – distributions that are non-normal but yet still symmetric obviously fall within this category; for example, the uniform distribution. They also described a modified Levene's test (Brown and Forsythe, 1974) incorporating the sample median, rather than the mean,

$$\text{ANOVA}(\left|X_{ij} - Mdn_j\right|), \qquad \text{(T3)}$$

where Mdn_j denotes the sample median for the j^{th} group and the remaining notation is the same as above. They went on to show that (T3) maintains its Type I error rate for asymmetric distributions. That is, Carroll and Schneider

show that, asymptotically, Levene's approach has the correct Type I error rate whenever the estimate of group 'center' is an estimate of group median, (T3). They went on to show that this explains why published Monte-Carlo studies have found that Levene's original proposal of centering at the sample mean, (T2), has the correct Type I error rate only for symmetric distributions, while centering at the sample median has correct Type I error rate both for symmetric and for asymmetric distributions (Brown and Forsythe, 1974). Interestingly, it was this median-based approach, (T3), and not the mean-based approach, (T2), that was found to be the most robust and useful of 56 possible tests for homogeneity in extensive simulations done by Conover, Johnson, and Johnson (1981).

What becomes evident from the simulations and mathematical work is that one needs to be precise about which Levene-like test is being used, (T2) or (T3). In fact, Levene introduced a strategy for data analysis, centering then applying the ANOVA, so there really is no one Levene test, per se, but instead an approach or strategy to the problem. Curiously, research papers and textbooks, as well as the codified methods in widely used statistical software, such as SPSS, continue to use the original Levene's test, (T2), without even mentioning that alternatives have been developed, or warning the data analyst that (T2) may be problematic. In many textbooks and software documentation it is stated that the (unspecified) "Levene test" is robust to non-normality and should be used instead of the notorious *F*-test from (T1). For example, SPSS calls its test the Levene *F*-test and states that their Levene test is reported in place of the traditional *F*-test in (T1).

To take our discussion a step further, textbooks going back 20 years, including recently published introductory statistics and research methodology textbooks for the social and educational sciences, were consulted to obtain information regarding the assumption of equal variances, for two independent groups, and how to test that assumption for one's data. Nearly all of the textbooks recommended using what they refer to as Levene's test for equality of variances and most suggested the use of SPSS (e.g., Cohen & Lea, 2004, Cramer, 1996, Tabachnick & Fidell, 2007, Vaughan, 1998). What is even more troubling is that one widely used and influential textbook suggested that if the sample sizes are equal then the assumption of equal variances can be disregarded (Hays, 1988), and yet another, Ferguson and Takane (1989) suggested to conduct the *F*-test of (T1), without reference to the over half

a century old finding by Box. In fact, as Keyes and Levy (1997) note, the Levene's test involving means, (T2), is available in many widely used statistical software packages such as BMDP, MINITAB, and SPSS, and in some cases (e.g., SPSS *t*-test) it is the only test made available to the software user.

To provide a concrete example of the analytic results noted above, we conducted a simulation study of the Type I error rate of the Levene test, (T2), provided by software packages like SPSS. In addition, we also included the *F*-test, (T1), to show its comparative performance to (T2) – i.e., how does the Levene's test compare to using the notoriously bad *F*-test? This comparison of (T1) to (T2) is somewhat novel and really meant to be a pointed contrast of the much-advocated use of Levene's test; by which is typically meant (T2). Therefore, the purpose of the simulation is to document the Type I error rates (and, if appropriate the statistical power) of Levene's test, (T2), with an eye toward cautioning researchers who implement tests of equal variances using Levene's test – by which is meant (T2) – in their day-to-day research activities. In fact, much to our chagrin in our era of "point-and-click" statistics, (T2) is embodied in day-to-day research activities by default in statistical software packages.

It should be noted that Carroll and Schneider's (1985) results make a simulation study, per se, unnecessary for the mathematically (and statistically) inclined who can decode those findings and incorporate them into their research practice. However, as we show above, given that 20 years after its publication Carroll and Schneider's results have evidently yet to enter the consciousness of textbook writers and statistical software designers in the social and behavioral sciences. Hence, this simulation study was, in its essence, intended to be a persuasive demonstration of why we should tend to the warnings in Carroll and Schneider (1985) and others in the statistical and methodological literature, and a reminder that when one hears reference to the "Levene test" one should then ask: which one?

Methods

Data Generation

Given our study purposes, a computer simulation was performed using SPSS software. Throughout the remainder of this paper, we will use the term "SPSS Levene's test" as shorthand for the original Levene test in (T2). Following standard simulation methodology (e.g., Zimmerman, 1987; 2004),

population distributions were generated using a pseudo random number sampling method to produce χ^2 distributions. The design of the simulation study was a 4 x 3 x 3 x 9 completely crossed design with: (a) four levels of skewness of the population distribution, (b) three levels of sample size, (c) three levels of sample size ratio, n_1/n_2, and (d) nine levels of ratios of variances – the dependent variables in the simulation design are the Type I error rates (when the variances are equal), and power under the eight conditions of unequal variances. Of course, we will only investigate statistical power in those conditions wherein the nominal Type I error rate (in our study 0.05) is maintained.

Shape of the population distribution. We investigated four levels of skewness, 0, 1, 2, and 3. We used the family of χ^2 distributions to simulate the population data. As is well known, as the degrees of freedom of a χ^2 distribution increases it more closely approximates a normal distribution[1]. The skew of the distributions for both groups were always in the same direction in all replications.

Sample Sizes. Three different sample sizes, $N = n_1 + n_2$, were investigated: 24, 48, and 96. Three levels of ratio of group sizes (n_1/n_2 : 1/1, 2/1, and 3/1) were also investigated.

Population variance ratios. Nine levels of variance ratios were investigated (σ_1^2/σ_2^2 : 5/1, 4/1, 3/1, 2/1, 1/1, 1/2, 1/3, 1/4, 1/5). 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 Levene's test in additive models for variances. As a whole, the complex multivariate

1 It should be noted that the population skewness was determined empirically for large sample sizes of 100,000 simulees with 10,000, 7.4, 2.2, and 0.83 degrees of freedom resulting in skewness values of 0.03, 1.03, 1.92, and 3.06, respectively.

variable space represented by our simulation design captures many of the possibilities found in day-to-day research practice.

Determining Type I Error Rates & Power

The frequency of Type I errors was tabulated for each cell in the design. In all there were 324 cells in the simulation design. As a description of our methodology, the following will describe the procedure for completing the steps for one cell in the design. First, two similarly distributed populations are produced; for this example it is two normally distributed populations that are sampled to create two groups. In this case each group has twelve members, and the population variances of the two groups are equal. An independent samples *t*-test using SPSS is then performed on the two groups; a Levene's test for equality of variances, by which we mean (T2), is reported in this procedure as a default test to determine if the variances are significantly different at the nominal alpha value of 0.05. Again, note, that we intend to mimic day-to-day research practice. This procedure was replicated 5,000 times for each cell in the design.

In the cells that maintained their Type I error rates, statistical power is represented by the percentage of times that the Levene's test, (T2), correctly rejected the null hypothesis.

Results and Conclusions

Type I error rates for the Levene's mean test is presented in Table 1. Table 1 has four columns: (i) total sample size, *N*, (ii) ratio of sample sizes, n_1/n_2, (iii) Type I error rate of SPSS's Levene test, (T2), and (iv) the Type I error rate for the *F*-test, (T1). Within the table there are the four levels of skewness of the population distribution. As an example, the Type I error rate of SPSS's Levene test for a skewness of zero, a total sample size of 24 (with 12 per group) is 6.0%.

For symmetric distributions (i.e., skewness of zero) the Type I error rates, for both the SPSS Levene's test and the *F*-test, were near the nominal alpha level of 0.05. Furthermore, for these symmetric distributions, the SPSS

Table 1

Empirical Type I Error Rates for the SPSS Levene's and the F Tests, for Various Sample Sizes, and Skewness of the Population Distribution

N	*n*1/*n*2	SPSS's Levene's Test	*F*-test
Skew = 0			
24	**1/1**	6.0	5.1
24	**2/1**	5.9	5.6
24	**3/1**	5.8	5.2
48	**1/1**	5.3	5.4
48	**2/1**	5.6	4.9
48	**3/1**	5.5	4.9
96	**1/1**	4.8	4.6
96	**2/1**	5.1	4.8
96	**3/1**	4.7	4.9
Skew = 1			
24	**1/1**	8.1	8.5
24	**2/1**	8.0	8.1
24	**3/1**	8.3	8.5
48	**1/1**	8.0	8.7
48	**2/1**	7.7	9.6
48	**3/1**	8.5	9.1
96	**1/1**	8.3	10.8
96	**2/1**	8.2	10.2
96	**3/1**	7.1	10.0
Skew = 2			
24	**1/1**	14.4	16.4
24	**2/1**	13.5	16.2
24	**3/1**	13.7	15.3
48	**1/1**	14.6	17.5
48	**2/1**	13.0	17.3
48	**3/1**	13.0	18.9
96	**1/1**	12.8	18.8
96	**2/1**	13.3	18.4
96	**3/1**	12.8	20.1
Skew =3			
24	**1/1**	22.8	24.4
24	**2/1**	23.4	27.7
24	**3/1**	19.7	28.0
48	**1/1**	21.0	24.8
48	**2/1**	20.4	27.8
48	**3/1**	19.2	29.9
96	**1/1**	20.3	27.5
96	**2/1**	20.2	29.2
96	**3/1**	19.5	29.9

Levene's test and the *F*-test were not influenced by either total sample sizes or unequal group sizes.

When the distribution had a skewness of one, two, or three, (i.e., the non-normal distributions) the Type I error rate of both the SPSS Levene's test and the *F*-test were inflated above the nominal level of 0.05. In fact, one finds that the skewness and sample size inequalities lead to even further Type I error rate inflation. Although both are quite inflated above their nominal Type I error rates, SPSS Levene's test appears to be less effected by unequal group sizes.

The statistical power results of the SPSS Levene's test and *F*-test under zero skewness (symmetric distribution) conditions are presented in Table 2. Note that power was only reported for those cells in the simulation design for which the nominal Type I error rate was protected. Table 2 is structured so that the first column lists the two statistical tests, either SPSS Levene's test or the *F*-test. Furthermore, columns two and three list the total sizes and the ratio of sample sizes, respectively. The ratio of samples sizes, n_1/n_2 : 1/1, 2/1, and 3/1, are also paired with the ratio of population variances, σ_1^2/σ_2^2, resulting in 1/2, 1/3, 1/4, 1/5 being inversely paired, and 5/1, 4/1, 3/1, 2/1 are directly paired. Therefore, as an example, in the case of a total sample size of 24, with 16 in group one and 8 in group two (i.e., a 2/1 sample size ratio), the statistical power of the SPSS Levene's test is 62.0% and the *F*-test 67.5% in the variance ratio of one to five (group one to group 2, hence an inverse pairing).

It is evident from Table 2 that when comparing the SPSS Levene's test to the corresponding *F*-test, in 66 of the possible 72 such comparisons in Table 2 the *F*-test is more powerful than the SPSS Levene's test. In fact, the *F*-test is more powerful than the corresponding SPSS Levene's test for all cases of direct pairings (i.e., when the larger sample size comes from a population with the larger variance). The power superiority of the *F*-test for normal distributions is expected from mathematical statistics (i.e., the *F*-test is most powerful for the normal population distribution).

Table 2

Statistical Power for SPSS Levene's Test and the F-Test for Varying Sample Size

			Population Variance Ratio, σ_1^2/σ_2^2							
Test	*N*	*n*1/*n*2	**1/5**	**1/4**	**1/3**	**1/2**	**2/1**	**3/1**	**4/1**	**5/1**
			Inverse Pairings				**Direct Pairings**			
Levene	24	1/1	60.6	50.7	35.7	18.2	18.2	35.7	50.7	60.6
F	24	1/1	81.0	73.1	54.4	29.3	29.3	54.4	73.1	81.0
Levene	24	2/1	62.0	50.8	36.1	17.7	13.8	26.0	40.1	49.7
F	24	2/1	67.5	54.6	38.5	17.3	35.5	58.2	75.6	84.0
Levene	24	3/1	57.0	45.7	33.3	16.4	11.7	19.5	28.4	36.6
F	24	3/1	49.1	37.0	23.0	9.0	37.9	57.5	73.4	83.7
Levene	48	1/1	92.4	84.3	64.9	31.3	31.3	64.9	84.3	92.4
F	48	1/1	98.6	95.0	81.2	48.5	48.5	81.2	95.0	98.6
Levene	48	2/1	90.3	80.3	63.1	31.9	27.1	55.3	76.4	87.7
F	48	2/1	95.1	88.2	72.2	37.7	51.7	82.8	94.7	98.3
Levene	48	3/1	85.2	75.0	57.3	29.4	21.4	45.7	66.2	78.0
F	48	3/1	88.0	78.8	60.2	27.8	51.0	80.2	93.2	97.6
Levene	96	1/1	99.9	98.9	92.4	58.6	58.6	92.4	98.9	99.9
F	96	1/1	100.0	99.8	97.8	76.4	76.4	97.8	99.8	100.0
Levene	96	2/1	99.5	98.0	89.4	56.9	51.0	89.0	98.4	99.8
F	96	2/1	99.8	99.4	94.8	68.3	74.6	97.8	99.9	100.0
Levene	96	3/1	98.7	96.1	85.9	48.4	41.0	81.9	96.1	99.1
F	96	3/1	99.4	98.2	90.7	56.2	70.6	97.0	99.8	100.0

Discussion

Several points are important to take away from this study.

1. When speaking of "Levene's" test it is important to be precise about which of the family of possible Levene tests one is referring to. Furthermore, most elementary textbooks in the social and behavioral sciences, as well as some of the widely used statistical software packages, are referring to the original test proposed by Levene in 1960, (T2), based on means.
2. The widely discussed Levene's test, (T2), has been shown to be sensitive to non-normality of the population distribution (e.g., Carroll & Schneider, 1985). Not surprisingly our simulation study also showed this inflation in Type I error rate. However, it should be noted that the inflation is not minimal and depends, to some degree, on degree of skewness and sample size ratio. In fact, what is interesting is that the Levene's test, (T2), actually performs on par, in terms of invalidity, with the notorious *F*-test of (T1), known widely since 1950 to be problematic. However, although both are quite inflated above their nominal Type I error rates, Levene's test, (T2), appears to be less affected by unequal group sizes – hardly a consolation when the Type I error rates tend to be between two to four times the nominal alpha!
3. In those situations wherein the Levene's test (and hence the *F*-test) maintain their nominal alpha, as expected from results in mathematical statistics, the statistical power findings show that in most situations the *F*-test is more powerful than the Levene's test, (T2).
4. Given the current state of knowledge, following Brown and Forsythe (1974) and Conover, Johnson, and Johnson (1981) we recommend that day-to-day researchers use the median-based Levene's test, (T3). Unfortunately, this median-based Levene test is not currently available in the widely used software packages such as SPSS. An easily implemented new statistical technique we have developed entitled the 'nonparametric Levene test' is, however, showing very promising results in terms of maintaining its nominal Type I error rate and having substantial statistical power in all the conditions studied in this current paper. 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 mean-based Levene test (T2) to the ranks. This can be easily accomplished using widely available software such as SPSS.

In closing, this paper, therefore, is a cautionary tale about Levene's test for homogeneity of variances. If one is using the original variation of Levene's test, a mean-based test, (T2), such as that found in SPSS, one may be doing as poorly (or worse) than the notorious F test of equal variances. We hope that Carroll and Schneider's caution about the Levene's test will soon become as widely recognized, and adopted in textbooks and statistical software, as was Box's tale in 1953 about the *F*-test, and that the median-based Levene's test will be more widely used. In closing, then, one needs to keep in mind Ted Micceri's observation that in real data situations the normal curve appears nearly as often as the mythical unicorn (Micceri, 1989). Therefore, to George Box's well-known quip in his influential paper on tests on variances that the preliminary test on variances is rather like putting to sea in a row boat to find out whether conditions are sufficiently calm for an ocean liner to leave port, we would add that using the mean-based Levene's, (T2), is akin to sending out a dinghy instead.

References

Box, G. E. P. (1953). Non-normality and tests on variance. *Biometrika, 40*, 318–335.

Brown, M. B., & Forsythe, A. B. (1974a). The small sample behavior of some statistics which test for the equality of several means. *Technometrics*, *16*, 129-132.

Brown, M. B., & Forsythe, A. B. (1974b). Robust tests for the equality of variances. *Journal of the American Statistical Association*, *69*(346), 364-367.

Carroll, R. J., & Schneider, H. (1985). A note on Levene's tests for equality of variances. *Statistics and Probability Letters*, *3*, 191-194.

Cohen, B. H., & Lea, R. B. (2004). *Essentials of statistics for the social and behavioral sciences*. Hoboken: John Wiley & Sons.

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.

Cramer, D. (1996). *Basic statistics for social research*. New York: Routeledge.

Ferguson, G. A., & Takane, Y. (1989). *Statistical analysis in psychology and education (6th ed*.). Toronto: McGraw-Hill Book Company.

Glass, G. V., & Hopkins, B. K. (1984). *Statistical methods in education and psychology (2nd ed.)*. New York: Prentice-Hall.

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.

Hayes, W. L. (1988). *Statistics (4th ed.)*. Toronto: Holt, Rinehart and Winston, Inc.

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

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin et al. (Eds.), *Contributions to probability and statistics: Essay in honor of Harold Hotelling* (pp. 278-292). Stanford, CA: Stanford University Press.

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

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.

Tabachnick, B. G., & Fidell, L. S. (2007). *Experimental designs using ANOVA*. Belmont, CA: Thomson, Brooks-Cole.

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

Vaughan, E. D. (1998). *Statistics: Tools for understanding data in the behavioral sciences*. Upper Saddle River: Prentice-Hall.

Zimmerman, D. W. (1987). Comparative power of Student t test and Mann-Whitney U test for unequal sample sizes and variances. *Journal of Experimental Education*, *55*, 171-174.

Zimmerman, D. W. (2004). A note on preliminary test of equality of variances. *British Journal of Mathematical and Statistical Psychology*, *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.