Testing for Equivalence of Group Variances

Constance A. Mara

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN PSYCHOLOGY
YORK UNIVERSITY
TORONTO, ONTARIO

June 2013

Abstract

The purpose of this paper was to situate a test for equality of group variances within the equivalence testing framework. Even though difference-based procedures are appropriate to answer questions about differences in some statistic (e.g., means, variances, etc.), these procedures are not appropriate to address questions related to variance homogeneity. Thus, if a researcher is interested in evaluating the similarity of group variances, it is more appropriate to use a procedure specifically designed to determine equivalence. A simulation study was used to compare newly developed equivalence-based tests to currently recommended difference-based variance homogeneity tests under data conditions common in psychological research. The results of this study provided evidence regarding the problems with assessing equality of variances with traditional difference-based tests. Most notably, traditional difference-based tests assess equality of variances from the wrong perspective, encouraging researchers to "accept" the null hypothesis. The results also demonstrated that the newly developed Levene-Wellek-Welch test for equivalence of group variances using the absolute deviations from the median was the best-performing equivalence-based test statistic in terms of accurate Type I error rates and highest power for detecting equivalence across the conditions evaluated. In addition, the use of the Levene-Wellek-Welch median-based test was demonstrated with an applied example, and an R function was provided in order to facilitate use of this newly developed equivalence of group variances test.

Acknowledgements

I would like to acknowledge the contributions of my mentor and supervisor, Dr. Robert Cribbie. He has provided me with the support, encouragement, enthusiasm, and positive attitude I needed to complete my doctoral work. Without his guidance and persistent help, this dissertation would not have been possible. His contributions to the success of my graduate training overall cannot be overstated. I hope to be even a fraction of the amazing teacher and mentor he is. I would also like to acknowledge my committee members, Dr. David Flora and Dr. Richard Lalonde, for their comments, revisions, promptness, and thoughtful questions throughout this process. Your feedback has been instrumental in the successful completion of this dissertation. Finally, I want to thank my husband for supporting my pursuit of higher education and for his unwavering confidence in me.

Table of Contents

List of Tables

List of Figures

Testing for Equivalence of Group Variances

Heterogeneity of variances occurs when one or more groups of sample scores have a wider dispersion of scores than other groups. Researchers are becoming increasingly interested in the properties of their data aside from central tendency, such as dispersion or variability. For instance, van Raalte et al. (2011) were interested in whether there was more variability in life expectancy in lower-educated groups in Europe versus those with advanced degrees. Pahnke et al. (2010) were interested in the variability of sweating rates of males versus females in an Ironman competition. Salgado (1995) examined whether the variability in validity coefficients in self-report tests for a specific construct was equivalent to the variability in validity coefficients in psychomotor tests evaluated by an external rater of the same construct. Finally, a more well-known reason for assessing group differences in variability is to verify the homogeneity of variances assumption related to traditional parametric tests of mean differences (i.e., ANOVA *F*-test; independent-samples *t*-test). Although the disciplines and research questions are varied, the fact remains that researchers need a valid test for assessing questions related to variability. More specifically, this paper addresses the need for a valid test of variance equivalence, or in other words, a test for homogeneity of variances.

There has been substantial research on different tests that can be used to test for variance equivalence, including Levene's (1960) test, which is the default test in popular statistical software packages (e.g., SPSS). This paper discusses whether traditional tests of variance homogeneity address the problem of variance equality from the wrong perspective. More specifically, although popular tests of variance homogeneity evaluate the null hypothesis that the variances are equal, it can be the case that the research hypothesis (not the null hypothesis) relates to the equality of the variances. In order to test for variance homogeneity, the use of equivalence tests is recommended. If one uses equivalence-based procedures, the research hypothesis of variance equality is properly aligned with the alternate hypothesis, not the null hypothesis.

The purpose of this paper is to situate a test for equality of group variances within the equivalence testing framework. As discussed later, even though difference-based procedures are appropriate to answer questions about differences in some statistic (e.g., means, variances, etc.), these procedures are not appropriate to address questions related to homogeneity. The main goal of this paper is to compare newly developed tests for equivalence of group variances to currently recommended variance homogeneity tests under data conditions common in psychological research. A review of traditional variance homogeneity tests as well as equivalence testing is outlined before developing equivalence testing procedures for detecting variance homogeneity.

Why Test for Equivalence of Variances?

One of the most common reasons that researchers want to test for equivalence of group variances is to justify the use of tests that assume variance homogeneity in their primary analysis (e.g., to meet the assumptions required by an independent-samples *t*-test or a one-way ANOVA *F*-test). In this case, the researcher would like to find that the variances are equal across groups, and, if using a traditional test for homogeneity of variances, would like to "accept" the null hypothesis for these tests,

$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_J^2$. In other words, the research hypothesis that the variances are equal is aligned with the null hypothesis rather than the alternate hypothesis.

It is important to note that it is not necessary to use a preliminary test of variance heteroscedasticity in order to justify the use of heteroscedastic procedures (e.g., Welch-adjusted ANOVA) since these tests are generally effective regardless of whether variances are equal across groups. Many researchers have suggested abandoning non-robust parametric procedures completely in favour of robust procedures that do not require the homogeneity of variances assumption (e.g., Wilcox, Charlin, & Thompson, 1986; Zimmerman, 2004). Specifically, these researchers recommend universal use of robust procedures when researchers are interested in comparing the central tendency of groups. These researchers emphasize that testing for homogeneity of variances is unnecessary given that robust procedures do not require homogeneous variances. Indeed, even papers proposing new procedures for testing for homogeneity of variances assert that this testing procedure can be abandoned if all researchers simply used a robust test statistic (e.g., Keselman, Wilcox, Algina, Othman, & Fradette, 2008). However, researchers in psychology still widely use non-robust parametric procedures and need to screen for the assumptions associated with these tests. In addition, other research has shown that using a preliminary test to screen for homogeneity of variances and then deciding to use a traditional ANOVA or a robust version (i.e., adaptive tests) is comparable to just using a Welch ANOVA at the first stage in terms of power (Gastwirth, Gel, & Miao, 2009).

A more interesting reason for assessing equivalence of variances is that the primary research question is concerned with whether the dependent variable variability of multiple groups or conditions is similar. As Parra-Frutos (2009) discusses, researchers are becoming more interested in the properties of their data aside from central tendency, such as dispersion or variability. For instance, research questions concerning "uniformity" or "similarity" of groups are increasingly common, which encompasses questions about the comparability of the dispersion of scores between groups. Bryk and Raudenbush (1988) argue that the presence of heterogeneity of variance across groups can have important implications for the research conclusions. Specifically, the presence of heterogeneity of variances in an experimental study indicates the presence of an interaction between person characteristics and treatment group membership. In other words, heterogeneity of variances can indicate that individuals vary in their response to the treatment (assuming the treatment group was a fixed effect). This could be an important consideration for researchers, and valid tests for evaluating heterogeneity or homogeneity of variances (depending on the researcher's expectations) would be important to evaluate within an experimental design. Indeed, in more complex modeling procedures, comparing the variability associated with a particular effect (e.g., variability around the intercept or slope in a latent growth curve model) between different groups is a common research goal (e.g., there are no differences between the groups on the variability around the slope).

Given these two reasons for testing for variance homogeneity, a valid test assessing equivalence of variances is quite relevant to the kinds of research questions psychologists (and researchers in other disciplines) are interested in, and preliminary tests for homogeneity of variances are necessary if a researcher wants to justify the use of a non-robust test. However, as this paper argues, the currently available procedures are incorrectly assessing variance equality, so new procedures need to be developed and evaluated.

Traditional Approaches to Testing for Variance Homogeneity

In order to assess variance homogeneity, Levene (1960) proposed transforming the sample scores to the absolute deviations of the sample scores from the sample mean with $z_{ij} = |X_{ij} - M_j|$, where X_{ij} is scores of the *i*th individual in the *j*th group, and then using a traditional ANOVA *F*-test on the z_{ij} to assess variance equality across groups. The null hypothesis for Levene's procedure is that the population variances of all *J* groups are equal, $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_J^2$. The alternate hypothesis states that at least one group variance is not equal to at least one other.

Since Levene's test was published, there have been numerous modifications proposed because the original version demonstrates some undesirable statistical properties, such as low power compared to other tests (especially when sample sizes are unequal), and non-robustness to non-normally distributed X_{ij}. For example, recommendations suggest that using the group median or trimmed mean, rather than the mean in Levene's test provides better Type I error control, even in asymmetric distributions (Brown & Forsythe, 1974; Keselman et al., 2008). There also has been an attempt to develop nonparametric procedures that evaluate variance homogeneity, such as rank-based Levene-type tests and bootstrapped versions of Levene's original test (e.g., Lim & Loh, 1996; Nordstokke & Zumbo, 2010).

Despite nearly 50 years of research, there does not seem to be a general consensus for a single test statistic for evaluating homogeneity of variances that works uniformly well across common data scenarios. Previous simulations studies (e.g., Conover, Johnson, & Johnson, 1981; Keselman, Games, & Clinch, 1979; Lim & Loh, 1996; Nordstokke & Zumbo, 2010) have made a wide range of recommendations regarding the optimal homogeneity of variance test that is also robust to non-normality. For instance, Conover et al. (1981) suggest that the original Levene test using the median is one of the best performing statistics across a wide range of analytic conditions. Lim and Loh (1996) similarly recommend the Levene test using the median, but suggest that a bootstrapped version improves the performance of this statistic. Nordstokke and Zumbo (2010) recommended a rank-based Levene test as the most robust test statistic across many data conditions, and rank-based Levene tests were also recommended in the Conover study as having some desirable properties under certain conditions. Keselman et al. (1979) report that no single test could be uniformly recommended, as the performance of many variance homogeneity statistics depended on the analytic condition. They did suggest, however, that the original Levene using the median or the Levene using the median with a Welch adjustment might be the best choices. In a later study, Keselman et al. (2008) looked at trimmed-means strategies and suggested that the original Levene with trimmed means or the Levene using trimmed means with a Welch adjustment performed the best across the conditions evaluated (based on Type I error rates only). They further suggest, contrary to the Lim and Loh study, that bootstrapping was not necessary because satisfactory Type I error rates can be obtained without bootstrapping.

Problems with Traditional Tests for Equivalence of Variances

Even though the results of previous simulation studies have found a number of homogeneity of variance tests to perform adequately under different data conditions, they are all fundamentally incorrect for the problem of determining the equality of population variances, in that these difference-based procedures aim to "accept" a point-null hypothesis regarding the exact equality of group variances. The probability of a Type I error when testing the null hypothesis, $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_J^2$, is the chance of incorrectly concluding there is a difference between the variances when, in fact, there are no differences in the variances. Type I error rate control is protection against incorrectly identifying a difference among two or more variances when they are the same. However, if one fails to reject the null hypothesis, one cannot conclude that the variances are equivalent; failure to reject the null hypothesis, $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_J^2$, only implies that there is not enough evidence to conclude that there is a difference among the variances.

Another issue with traditional tests is that rejection or non-rejection of the null hypothesis of homogeneity of variance conveys very little about the potential similarity of the group variances in question. Specifically, the point-null hypothesis evaluated by difference-based homogeneity of variance tests is too specific and impractical for assessing the equivalence of the group variances. For instance, if there is a large sample size and a very minor difference among group variances, it is likely that a difference-based variance homogeneity test will reject the null hypothesis and declare the group variances different. However, small differences in the variances are usually expected, and thus the results of the traditional homogeneity of variance test and subsequent conclusions regarding the similarity of the group variances in this case could be impractical. Conversely, smaller sample sizes may result in very little power to detect important differences in the variances, resulting in inaccurate conclusions about the group variances. More generally, the power of difference-based procedures to detect equality of variances decreases (rather than increases) as sample size increases. This property is clearly incongruent with typical null hypothesis testing expectations.

Equivalence Testing

Equivalence tests are appropriate for a research question that deals with a lack of association. For example, a researcher may be interested in demonstrating that the means of groups are equivalent or that no relationship exists between two variables (e.g., Cribbie, Gruman, & Arpin-Cribbie, 2004; Goertzen & Cribbie, 2010; Robinson, Duursma, & Marshall, 2005; Rogers, Howard, & Vessey, 1993), or that the variances of two or more populations are equal (as proposed in the current study). A "*complete* lack of association" is unrealistic, as it is mathematically impossible for entities to be completely unrelated if the dependent variable is truly continuous (e.g., difference between the means is exactly zero, as with the traditional null hypothesis H_0: $\mu_1 = \mu_2$). Instead, with equivalence testing, a lack of association implies that the relationship is so small that it can be considered inconsequential or meaningless. For example, if a researcher was interested in demonstrating that two population variances were equivalent, then the researcher must decide how large a difference in the variances can be considered inconsequential. This difference is defined *a priori* as the equivalence interval ($-\varepsilon$, $+\varepsilon$). In other words, an equivalence test assesses whether the relationship between two or more entities (e.g., difference between population variances) falls within a specified interval which defines an unimportant difference (e.g., $-\varepsilon \leq \sigma_1^2 - \sigma_2^2 \leq \varepsilon$)[1].

Specifying the equivalence interval is the most challenging aspect of equivalence testing because there are no concrete rules to help researchers choose the appropriate equivalence interval. The equivalence interval must be selected based on researchers' knowledge of their field, their expertise with the constructs and samples being used, and an understanding of how "meaningless" might be quantified for their particular research question.

Equivalence testing was first introduced to psychology by Rogers and colleagues (1993). Since then, there have been numerous papers recommending its use in many data analytic situations common to psychological research (e.g., Cribbie et al., 2004; Seaman & Serlin, 1998). The most common procedure was developed by Schuirmann (1987) and involves testing the equivalence of two independent sample means. For detecting the equivalence of more than two means, simulation research (e.g., Cribbie, Arpin-Cribbie, & Gruman, 2010) has recommended Wellek's (2003) one-way test of equivalence. Wellek's test simultaneously evaluates the equivalence of all *J* population means. The null

---

[1] Alternatively, equivalence tests could be used to assess the "similarity" of a particular value to a target value.

hypothesis for a one-way equivalence test is that the difference among the means of the groups falls within an equivalence interval such that:

$$H_0 : \Psi^2 \geq \varepsilon^2$$
$$H_1 : \Psi^2 < \varepsilon^2$$

where ε is the equivalence limit and

$$\Psi^2 = \frac{\Sigma_{i=1}^{n_j} \Sigma_{j=1}^{J} \left(\frac{n_j}{\bar{n}} \right) \left(M_j - \bar{X}_{..} \right)^2}{\sigma^2}$$

where $\bar{n}$ represents the mean sample size of the groups, M_j represents the mean of the jth group, $\bar{X}_{..}$ represents the average of the means for the J groups (i.e., the grand mean), and σ^2 represents the average within-group variance (assumed to be equal across groups). The null hypothesis for this test is rejected if $\Psi^2 < \Psi_{crit}$, where

$$\Psi_{crit} = \left(\frac{J-1}{\bar{n}} \right) F_{J-1, N-J, \alpha\left(\bar{n}\varepsilon^2\right)}$$

where $\bar{n}\varepsilon^2$ represents the noncentrality parameter. Wellek cautiously recommends adopting $\varepsilon = .25$ for a strict equivalence criterion and $\varepsilon = .50$ for a liberal equivalence criterion (discussed in more detail later).

Traditional Variance Homogeneity Procedures Evaluated in the Current Study

The current simulation study evaluated four traditional difference-based tests for homogeneity of variances, each of which is described below.

Levene's (1960) original test for homogeneity of variances ("Lev_mean"). Although Levene's (1960) test was not recommended in the literature (e.g., Conover et al., 1981; Lim & Loh, 1996), it is still regularly reported in popular statistical software programs, so it was included in this study. As mentioned previously, this test converts the sample scores, X_{ij}, with $z_{ij} = |X_{ij} - M_j|$, where M_j is the jth sample mean, $j = 1, \ldots, J$, and then uses the transformed scores, z_{ij}, in the following ANOVA test statistic to assess variance homogeneity:

$$F = \frac{(N-J)}{J-1} \frac{\sum_{j=1}^{J} n_j (\bar{Z}_j - \bar{Z}_{..})^2}{\sum_{j=1}^{J} \sum_{j=1}^{n_j} (z_{ij} - \bar{Z}_j)^2}$$

where n_j is the sample size of the jth group, $\bar{Z}_j$ is the mean of the z_{ij} for the jth group, and $\bar{Z}_{..}$ is the overall, grand mean for all z_{ij}. Critical values for F can be obtained from the F-distribution based on $J - 1$ and $N - J$ degrees of freedom.

Levene's test using the median ("Lev_mdn"). This modification of Levene's test, originally proposed by Brown and Forsythe (1974), was considered the best procedure in Conover et al.'s (1981) simulation study, in terms of most accurate Type I error rates. Thus, I included this procedure in the study. Instead of using the jth sample mean in the sample score transformation, this modification uses the transformation, $\tilde{z}_{ij} = |X_{ij} - MDN_j|$, where MDN_j is the jth sample median. Again, the transformed scores are analyzed using an ANOVA F- test.

Levene's original test with a Welch adjustment ("LevWelch_mean"). Welch's (1951) adjusted degrees of freedom procedure has been proposed as a solution to unequal variance issues in independent groups design procedures like Student's t-test and the ANOVA F-test. However, the Welch adjustment to the ANOVA F-test has relevance to Levene's test for homogeneity of variances (and its modification), given that Levene's test uses the ANOVA F-test and also assumes homogeneity of variances (more specifically, the variances of absolute values of the deviation scores, z_{ij}). It seems illogical to have a test for homogeneity of variances that, itself, assumes homogeneity of variances. Thus, researchers have proposed using the Welch-adjusted statistic to test for homogeneity of variances (e.g., Keselman et al., 1979; Parra-Frutos, 2009; Wilcox et al., 1986).

The original one-way Welch-adjusted ANOVA $F`$-test is defined as:

$$F' = \frac{\sum w_j \left(M_j - \overline{X}'_{..}\right)^2 \Big/ J-1}{1 + \frac{2(J-2)}{J^2-1} \sum \left(\frac{1}{n_j - 1}\right)\left(1 - \frac{w_j}{\sum w_j}\right)^2},$$

where $w_j = \frac{n_j}{s_j^2}$, $\overline{X}'_{..} = \frac{\sum w_j M_j}{\sum w_j}$, and $df' = \frac{J^2 - 1}{3\sum \left(\frac{1}{n_j - 1}\right)\left(\frac{w_j}{\sum w_j}\right)^2}$. As with the original Levene test, one simply substitutes the transformed scores, $z_{ij} = \left|X_{ij} - M_j\right|$, into the $F`$ equation to assess homogeneity of variances (without requiring the homogeneity of variances assumption), so that the test statistic becomes:

$$F` = \frac{\sum w_{z_j} (\overline{Z}_j - \overline{Z}'_{..})^2 \Big/ J-1}{1 + \frac{2(J-2)}{J^2-1} \sum \left(\frac{1}{n_j - 1}\right)\left(1 - \frac{w_{z_j}}{\sum w_{z_j}}\right)^2},$$

where $w_{z_j} = \frac{n_j}{s_{z_j}^2}$, $\overline{Z}'_{..} = \frac{\sum w_j \overline{Z}_j}{\sum w_j}$ and $\overline{Z}_j$ is the mean of the z_{ij} for the *j*th group.

Levene's median-based test with Welch adjustment ("LevWelch_mdn"). This procedure uses the absolute deviations from the median, $\tilde{z}_{ij} = |X_{ij} - MDN_j|$, to calculate the Welch ANOVA F'- test to assess homogeneity of variances (outlined previously), so that $\overline{Z}_j$ is the mean of the $\tilde{z}_{ij}$ for the *j*th group. Given that the Brown-Forsythe version of the procedure is most widely recommended in the literature, a Welch-version of this test was included in this study.

It is important to remember that all of the difference-based tests discussed in this section test the null hypothesis that the population variances are equal, and rejection of this null hypothesis implies that the variances cannot be assumed equal. However, as discussed earlier, when the goal is to demonstrate that the population variances of the groups are equal, the alternate hypothesis, rather than the null hypothesis, should be expressed in terms of variance equality.

Equivalence-Based Homogeneity of Variance Tests

Given the fundamental problems with the traditional tests for homogeneity of variances, I developed an equivalence-based test for homogeneity of variances along with several modifications. The null hypothesis for a one-way equivalence test for homogeneity of variances is that the difference among the variances of the groups falls within an equivalence interval,

$$H_0 : \Psi^{2*} \geq \varepsilon^2$$
$$H_1 : \Psi^{2*} < \varepsilon^2$$

where Ψ^{2*} represents the equivalence of group variances test statistic, defined shortly.

Levene-Wellek test for equivalence of variances ("LW_mean"). This procedure is based on Wellek's (2003) original one-way equivalence test statistic, substituting Levene's original transformation in place of the raw scores. This new hybrid test statistic can be defined as:

$$\Psi^{2*} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{J}\left(\frac{n_j}{\bar{n}}\right)\left(\bar{Z}_j - \bar{Z}_{..}^{*}\right)^2}{s_{z_j}^2}$$

with Levene's original transformation, $z_{ij} = \left|X_{ij} - M_j\right|$, so that $\bar{Z}_j = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j} z_{ij}}{n_j}$,

$\bar{Z}_{..}^{*} = \frac{\left(\sum_{j=1}^{J}\sum_{i=1}^{n_j} z_{ij}\right)}{\sum_{j=1}^{J} n_j}$, and $s_{z_j}^2$ is the variance of the transformed sample scores.

As mentioned previously, both the original Levene test and Wellek's one-way test assume homogeneity of variances, which is an unreasonable assumption when these tests are used to evaluate homogeneity of variances. In addition, previous research on traditional difference-based homogeneity of variance tests have found that certain modifications of the original Levene test perform better. Thus, this study included three additional procedures based on modifications of this newly developed Levene-Wellek test, as described next.

Levene-Wellek using the median ("LW_median"). This procedure is an adaptation of the Levene-Wellek test (defined above) using the absolute deviations from the sample median instead of the absolute deviations from the sample mean (i.e., Brown-Forsythe transformation of the sample scores; "LW_median");

Levene-Wellek-Welch ("LWW_mean"). This version of the procedure is based on the Levene-Wellek test on the mean, but including a Welch adjustment to test for equivalence of group variances without assuming homogeneity of variances. As discussed previously, the ANOVA test statistic used to evaluate variance homogeneity assumes that the variances (of the transformed scores) are homogeneous. Thus, researchers have suggested using a robust version of the ANOVA test statistic (i.e., the Welch-adjusted ANOVA). The new equivalence-based robust test statistic can be defined as

$$\Psi_W^{2*} = \frac{\sum w_{z_j}(\bar{Z}_j - \bar{Z}'_{..})^2}{1 + \frac{2(J-2)}{J^2-1}\sum(\frac{1}{n_j-1})(1 - \frac{w_{z_j}}{\sum w_{z_j}})^2}\left(\frac{J-1}{\bar{n}}\right);$$

where $w_{z_j} = {n_j}/{s_{z_j}^2}$, $\bar{Z}'_{..} = {\sum w_{z_j}\bar{Z}_j}/{\sum w_{z_j}}$, n_j is the size of the jth group, $s_{z_j}^2$ is the variance of the transformed scores for the jth group, and $\bar{Z}_j$ is the mean of the transformed sample scores for each group (as defined previously). The test statistic is approximately distributed as F with J-1 numerator degrees of freedom and denominator degrees of freedom as:

$$df' = \frac{J^2 - 1}{3\sum\left(\frac{1}{n_j-1}\right)\left(1 - \frac{w_{z_j}}{\sum w_{z_j}}\right)^2}.$$

Levene-Wellek-Welch using the median ("LWW_median"). The final novel procedure developed for this study uses the previously defined Levene-Wellek-Welch test, but instead of the original Levene transformation, this procedure uses the Brown-Forsythe transformation of the absolute deviations of the sample scores from the median.

The Equivalence Interval

Wellek (2003) provides several broad recommendations in terms of selecting equivalence intervals. However, the nature of the research should be the determining factor in the selection of an appropriate equivalence interval. Indeed, Wellek and other equivalence testing researchers have cautioned that general recommendations or fixed general rules regarding the selection of an equivalence interval is not advisable, but should be a point of careful consideration that is specific to the individual study. Epsilon (ε) can be described as the maximum difference in the variances that one would consider unimportant. In general, Wellek suggests that entities differing by no more than 10% are very similar, while differences of more than 20% are practically significant. Thus, a 10% difference would be a strict equivalence criterion ($\varepsilon = .25$) and 20% would be a more liberal equivalence criterion ($\varepsilon = .50$; see Wellek, 2003, pp. 16, 17, & 22 for details).

Issues to Consider in Comparing Equivalence Tests and Difference Tests

It is important to discuss some difficulties with comparing the results of difference-based tests to those of equivalence-based tests. The major issue is that these two types of tests evaluate different hypotheses. Difference-based tests evaluate a point-null hypothesis that is very specific, and in the case of variance equality, quite impractical. For example, it is strictly impossible to find that variances are exactly equal, if one uses enough decimal places. In addition, the research hypothesis regarding variance equality is aligned with the null hypothesis, rather than the alternate hypothesis, so the researcher's goal is to “accept” the null hypothesis. Equivalence-based tests evaluate the null hypothesis that the difference among the variances falls outside a pre-specified equivalence interval. Thus, to determine that the variances are nearly equivalent, one wants to reject this null hypothesis and find instead that the difference among the variances falls within the equivalence interval. In this case, the research hypothesis is the alternate to the null hypothesis, which is congruent with normal null-hypothesis testing procedures. However, comparisons could be made regarding the overall pattern of results for detecting homogeneity of variances between these two testing methods. The outcome in this study was the proportion of declarations of equivalence. In other words, what was the probability of detecting equivalence? This outcome was defined by the proportion of non-rejections of the null hypothesis in the difference-based tests and by the proportion of rejections of the null hypothesis for the equivalence-based tests.

Method

Monte Carlo simulations were used to compare the probability of declaring equivalence for the four difference-based tests for homogeneity of variances to that of the four novel equivalence-based tests for equality of variances. In addition, Type I error rates and power for the equivalence procedures were assessed and compared. The performance of the eight homogeneity of variance tests was evaluated with a normal population distribution shape as well as a positively skewed distribution (χ^2 with 3 degrees of freedom; see Figure 1) because previous research (e.g., Brown & Forsythe, 1974; Keselman et al., 2008) has indicated that some homogeneity of variance tests are not robust under conditions of non-normality. In order to evaluate the Type I error rates of the equivalence-based procedures, the liberal bounds of $\pm$ 0.5α (Bradley, 1978) were used. Therefore, with an alpha level of .05, a procedure was considered to have an accurate empirical Type I error rate in a specific condition if the rate fell between .025 and .075. In each table, inaccurate Type I error rates are italicized, and the corresponding power rates for these conditions are also italicized since those power rates would also be inaccurate. The simulations were conducted with the open-source statistical software *R* (R Development Core Team, 2010).

The definition of "*power*" is different for the equivalence-based tests compared to the difference-based tests because, as discussed previously, these two types of tests have different null hypotheses. Therefore, instead of determining the probability of rejecting a false null hypothesis, that is, "*power*" for any particular test, this study determined the "probability of finding equivalence" for both the equivalence-based and the difference-based procedures. In other words, this study focused on the probability that a particular test declares the variances equivalent when they are in fact equivalent (where "*equivalent*" is defined by the null hypothesis for the difference-based tests and by the equivalence interval for the equivalence-based tests). Empirical Type I error rates for the equivalence-based tests were obtained by deriving the differences in the variances that matched the bounds of the equivalence interval (i.e., $\Psi^2 = \varepsilon^2$) in conditions where the population variances differed across groups.

Several variables were manipulated in this study, including distribution shapes, balanced versus unbalanced designs (i.e., equal versus unequal group sizes), number of groups, sample sizes, variances, and pairings of unequal sample sizes with unequal variances. The conditions examined in this study can be found in Tables 1 and 2. For the equivalence-based tests, both of the recommended equivalence limits of $\varepsilon = .25$ and $\varepsilon = .50$ were used (Wellek, 2003).

For the normally distributed conditions, n_j standard normal observations were generated for the jth group, where $j = 1, \ldots, J$, and the resulting values were multiplied by $\sqrt{\sigma_j^2}$ so that the observations would have variances, σ^2_j, as outlined in Table 1. In order to examine the effects of positively skewed distributions on the performance of the test statistics, n_j observations were generated for each of the J groups from a χ^2 distribution with 3 degrees of freedom. In order to ensure the observations from the χ^2 distribution had the variances specified in Table 1, first the mean and variance of the distribution had to be set to 0 and 1, respectively. This was accomplished by subtracting 3 (the degrees of freedom, which is equal to the mean of the χ^2 distribution) from the observations, which centers the distribution of scores at zero, and then dividing by $\sqrt{2 * 3}$ (the variance of the χ^2 distribution is $2df$; in this case, $df = 3$, so dividing the observations by the square root of 2*3 sets the standard deviation to 1). The resulting values were then multiplied by $\sqrt{\sigma_j^2}$, to produce a distribution of observations with the variances outlined in Table 1.

Once the observations were generated for each replication, the four difference-based procedures and the four equivalence-based procedures were performed on the data of each replication. As discussed previously, to determine the probability of declaring equivalence for the difference-based tests, it was noted when the null hypothesis was not rejected. In order to determine the probability of declaring equivalence for the equivalence-based tests (i.e., power), it was noted when the null hypothesis was rejected. This process was repeated across 10,000 replications per condition to obtain the probability of declaring equivalence for each condition.

Unbalanced designs, defined as unequal sample sizes, that are paired with unequal variances can affect Type I and Type II error control of ANOVA-type procedures (Keselman et al., 1998; Othman et al., 2004). Thus, the current study examined both positive and negative pairings of the variances and sample sizes. Positive pairing occurs when the largest group size is paired with the largest variance and the smallest group size is paired with the smallest variance. Negative pairing occurs when the largest group size is paired with the smallest variance and the smallest group size is paired with the largest variance. Previous research on the robustness of ANOVA-type procedures (Othman et al., 2004; Yin & Othman, 2009) has found that positive pairings result in conservative Type I error rates and negative pairings result in liberal Type I error rates. A procedure is considered liberal if its Type I error rate is greater than the nominal alpha level and considered conservative if its Type I error rate is less than the nominal alpha level.

In summary, four novel equivalence-based procedures were evaluated in a simulation study based on Type I error rates and power across 384 conditions (12 sample size conditions × 2 group size conditions × 8 variance ratios × 2 distributions shapes). The difference-based procedures were evaluated based on the probability of declaring equivalence in 216 conditions (there are fewer conditions evaluated because the difference based tests were not subject to the same Type I error conditions as the equivalence procedures, such that the equivalence interval conditions did not apply).

Results

Difference-Based Procedures

Normal Distributions.

When the population variances of the groups were exactly equal, this was a Type I error condition for the difference-based procedures. Therefore, the probability of declaring equivalence (i.e., "accepting" the null hypothesis) in this condition should have been approximately $1 - \alpha$ (in this case, .95), regardless of sample size. Although in most cases the rates were close to .95, with positive and negative pairings of unequal sample sizes and variances and small sample sizes, the rates were sometimes too conservative or too liberal (see Tables 3 and 4).

The 1.3:1 variance ratio condition was a power condition for the difference-based tests, so the probability of declaring equivalence (i.e., failing to reject the point-null hypothesis) equals the Type II error rate. Thus, as expected, the probability of declaring equivalence decreased as sample sizes increased (see Tables 5 and 6). Specifically, for $J = 2$, the probability of declaring equivalence was between 85% and 95% when $\bar{n} = 10$, and decreased slightly to between 76% and 80% when $\bar{n} = 100$. For $J = 4$, probability of declaring equivalence was highest for $\bar{n} = 10$, ranging between 86% and 98%, then slightly decreased to between 82% and 90% when $\bar{n} = 100$.

For the 2:1 variance ratio condition, the difference-based tests had a very high probability of declaring equivalence at $\bar{n} = 10$. In the largest sample size conditions ($\bar{n} = 100$), the probability of declaring equivalence was much lower and ranged between 11% and 21% in the equal sample size conditions (see Tables 7 and 8). It is important to note that the 2:1 variance ratio in this condition meant the point-null hypothesis of the difference-based procedures was false, and thus these results were not unexpected. However, the backward nature of using difference-based tests for addressing questions of equivalence was apparent, as equivalence is found up to 97% of the time at small sample sizes, but this same difference in the variances was statistically different the majority of the time in the largest sample sizes.

For a 6:1 variance ratio and $J = 2$ (see Table 9) in the smallest sample size conditions, the probability of declaring equivalence was as high as 99% for the negative pairing conditions (specifically, for the Levene-Welch test using the median), and was as high as 57% in the equal sample size conditions. In the largest sample size conditions, as expected, the probability of declaring equivalence was zero. When $J = 4$ (see Table 10) in the smallest sample size conditions, the probability of declaring equivalence was as high as 85% in the negative pairing conditions, and was as high as 72% in equal sample size conditions.

Positively Skewed Distributions (χ^2, 3 df).

As discussed previously, when the variances of the groups were exactly equal, this condition evaluated Type I error rates for the difference-based procedures. Therefore, the probability of declaring equivalence in this condition should have been approximately $1 - \alpha$ (.95) for the difference-based procedures. As found in Tables 11 and 12, this result was obtained for most replications with the median-based tests, but the mean-based procedures demonstrated rates that were often very conservative. The rates across the procedures ranged from approximately 95% with the Levene test using the median, but were as low as 80% for the other procedures. Thus, the probability of declaring equivalence was less than what was found in the normally distributed conditions. Note that, as before, sample size did not impact the probability of declaring equivalence in this condition for the difference-based tests.

For a 1.3:1 variance ratio, the variances were slightly different, so the point null hypothesis for the difference-based tests was false. The probability of declaring equivalence (i.e., failing to reject the point-null hypothesis) decreased as sample sizes increased (see Tables 13 and 14), as expected. Specifically, for $J = 2$, the probability of declaring equivalence was between 78% and 97% when $\bar{n} = 10$, and decreased to between 74% and 87% when $\bar{n} = 100$. For $J = 4$, the probability of declaring equivalence was highest for $\bar{n} = 10$, ranging between 73% and 96%, and remained somewhat unchanged at approximately 82% to 95% when $\bar{n} = 100$.

When there was a 2:1 variance ratio, again, the point-null hypothesis for the difference-based procedures was false. Consequently, the probability of declaring equivalence (i.e., not rejecting the null hypothesis) decreased as sample sizes increased.

In the smallest sample size conditions, the probability of declaring equivalence ranged from 64% to 97% (see Tables 15 and 16). In the largest sample size conditions, the probability of declaring equivalence ranged from 24% to 59%.

For a 6:1 variance ratio, many false declarations of equivalence were observed for the difference-based procedures. In the smallest sample size condition (see Tables 17 and 18), false declarations of equivalence occurred between 33% and 99% of the time. When $\bar{n} = 25$, false declarations of equivalence ranged from 9% to 50% across conditions. In the largest sample size conditions, the rate was approximately zero.

Equivalence-Based Procedures

Normal Distributions.

Empirical Type I error rates. For $J = 2$ groups and $\varepsilon = .25$, the Type I error rates for the equivalence procedures are in Table 19. All four equivalence-based procedures maintained the Type I error rates very close to the nominal level when the group sample sizes were equal, with error rates ranging from .0475 to .0557. When unequal variances and sample sizes were positively paired, the Type I error rates were acceptably close to the nominal level, ranging from .0387 to .0565. For negative pairings of sample sizes to variances, the Type I error rates for the largest sample size condition ($\bar{n} = 100$) ranged from .0437 to .0505. However, with small sample sizes in the negative pairing conditions, the Type I error rates for the Levene-Wellek tests were slightly liberal (e.g., .0847, .0762). The Levene-Wellek-Welch procedures were less affected by negative pairings and maintained the Type I error rates within the bounds of .025 and .075, with empirical Type I error rates ranging from .0464 to .0687.

For $J = 2$ groups and $\varepsilon = .50$, the Type I error rates are in Table 20. The same pattern of results reported for the more strict equivalence criterion also holds for this more liberal equivalence criterion. When sample sizes were equal, the Type I error rates of all four equivalence-based procedures were close to the nominal level, ranging from .0357 to .0531. For positive pairing conditions, the Type I error rates were also very accurate, ranging from .0394 to .0552. However, for the negative pairing conditions, the Type I error rates were too liberal at smaller sample sizes for the Levene-Wellek procedure using the median (e.g., .0886, .0759). The Levene-Wellek using the mean had better Type I error rates, ranging from .0306 in the largest sample size condition to .0686 in the smallest sample size condition. The Levene-Wellek-Welch procedures (both mean and median) maintained the Type I error rates within the bounds of .025 to .075 in all conditions, but were also slightly more liberal when sample sizes were smaller.

For $J = 4$ groups and $\varepsilon = .25$, the empirical Type I error rates for all four equivalence procedures were acceptably close to the nominal level, within the bounds of .025 and .075 (see Table 21). This result occurred for equal sample size conditions as well as the positive pairing and negative pairing conditions.

For $J = 4$ groups and $\varepsilon = .50$, the Type I error rates (see Table 22) in the equal sample size conditions were maintained at the nominal level, ranging from .0381 to .0702. For the positive pairing conditions, the Levene-Wellek-Welch procedures had acceptable Type I error rates in all sample sizes, ranging from .0388 to .0523. However, both of the Levene-Wellek procedures (i.e., based on the mean and the median) had overly liberal Type I error rates at the highest sample size (i.e., .0824 and .0869). For the negative pairing conditions, the Levene-Wellek-Welch using the median had a very liberal Type I error rate (i.e., .1014) at the smallest sample size condition ($\bar{n}$ = 10). However, at the larger sample sizes in the negative pairing conditions, the Type I error rates were acceptable for the Levene-Wellek-Welch test using the median. The other three equivalence procedures maintained the Type I error rates within the bounds of .025 to .075 in all of the negative pairing conditions.

Power. When variances were exactly equal, the difference in the variances (equal to zero) fell within the equivalence interval, thus this was a power condition for the equivalence-based procedures. For $J = 2$ groups and a strict equivalence criterion (ε = .25), power for all of the equivalence procedures was quite low at the smallest samples sizes, but increased to approximately 60% power at the largest sample size (See Table 23). The same pattern was found when sample sizes were equal, and for positive and negative pairing conditions. There was a slight power advantage for the median-based equivalence procedures. When $J = 4$ and ε = .25, a similar trend occurred, with low power in the smallest sample size condition and approximately 43% power in the largest sample size conditions (see Table 24). Again, there was a slight power advantage for the median-based equivalence tests.

When variances were exactly equal and ε = .50, power, as expected, improved. When $J = 2$, over 90% power for detecting equivalence was achieved when $\bar{n}$ = 50, and reached nearly 100% in the largest sample size conditions. This result occurred for equal sample sizes as well as positive and negative pairing conditions (see Table 25). The same pattern of results was observed when $J = 4$ (see Table 26).

With a 1.3:1 variance ratio and $\varepsilon = .25$, the test statistic, ψ^2, was less than ε^2; thus the differences in the variances was less than the equivalence interval. Thus, this condition also evaluated power for the equivalence procedures (i.e., probability of declaring equivalence when the variances were not meaningfully different). When $J = 2$ and small sample sizes, power for the equivalence procedures was low, but increased as sample sizes increased, reaching approximately 35% (see Table 27). For $J = 4$ (see Table 28), the same pattern of results was observed, with power reaching approximately 40% to 43% in the largest sample size condition.

For a 2:1 variance ratio and $\varepsilon = .50$, this difference in the variances was within the equivalence interval for the equivalence procedures, such that $\psi^2 < \varepsilon$; therefore, this condition was another test of the power of these procedures for the more liberal equivalence limit. For $J = 2$ and when sample sizes were equal, power was quite low at the smaller sample sizes, and increased to between 57% and 59% in the largest sample size condition ($\bar{n} = 100$). For the positive pairing condition, power in the largest sample size condition ranged from 54% to 62%. For the negative pairing condition at the largest sample size, power ranged between 52 and 67% (see Table 29). For $J = 4$, power was slightly lower than it was in the two group condition. In the largest sample size condition, power was approximately 41% to 61% (see Table 30). All four equivalence procedures had comparable power rates across all sample size and variance combinations.

False declarations of equivalence. For a 6:1 variance ratio, ψ^2 was greater than ε^2; thus, the differences in the variances exceeded the equivalence interval and the equivalence procedures should not reject the null hypothesis of variance heterogeneity.

This was also another evaluation of the Type I error rates of the equivalence procedures, given that the null hypothesis of variance heterogeneity was true in this condition. Specifically, the difference among the group variances exceeded the equivalence interval. Note, however, that the error rates in this variance ratio condition should be less than the Type I error rates obtained when the differences among the variances matched the bounds of the equivalence interval. When $J = 2$ groups and $\varepsilon = .25$ (see Table 31), the probability of declaring equivalence for the equivalence-based procedures was zero in the larger sample size conditions. When $J = 2$ and $\varepsilon = .50$ (see Table 32), the probability of declaring equivalence for the equivalence procedures was also zero in the largest sample size conditions.

For $J = 4$ and $\varepsilon = .25$, the probability of declaring equivalence when there was a 6:1 variance ratio, as desired, was very low and was zero in the highest sample size conditions (see Table 33). For $J = 4$ and $\varepsilon = .50$, the probability of declaring equivalence was also low at small sample sizes and was zero in the larger sample size conditions.

Positively Skewed Distributions (χ^2, 3 df).

Empirical Type I error rates. For the equivalence tests when $J = 2$ and $\varepsilon = .25$, both the Levene-Wellek using the median and Levene-Wellek-Welch using the median maintained the Type I error rates within the bound of .025 to .075. However, the mean-based versions of these procedures had Type I error rates that were too liberal at the largest sample sizes for equal sample sizes conditions, and the positive pairing conditions. For the negative pairing conditions, the Levene-Wellek using the mean was the only procedure to have inaccurate Type I error rates, exceeding the nominal level (i.e., too liberal) at the larger sample sizes (see Table 35).

When $J = 2$ and $\varepsilon = .50$, the Type I error rates for all of the equivalence procedures were accurate when sample sizes were equal. When unequal sample sizes were positively paired with the variances, the Levene-Wellek using the mean and the Levene-Wellek-Welch using the mean had Type I error rates that were too liberal at $\bar{n} = 50$. When variances were negatively paired with unequal sample sizes, the Levene-Wellek using the mean had Type I error rates that were too liberal in some conditions, and the Levene-Wellek-Welch using the median had a Type I error rate that was slightly conservative at $\bar{n} = 100$. See Table 36.

For $J = 4$ and $\varepsilon = .25$, the Type I error rates were accurate across all conditions for the median-based Levene-Wellek and the median-based Levene-Wellek-Welch equivalence tests. However, the mean-based versions of these tests had inaccurate Type I error rates at the smallest sample size condition, $\bar{n} = 10$, when sample sizes were equal or positively paired. See Table 37.

For $J = 4$ and $\varepsilon = .50$, all of the equivalence procedures, once again, maintained accurate Type I error rates when variances were negatively paired with unequal sample sizes. However, when variances were positively paired with the largest unequal sample size, $\bar{n} = 100$, only the Levene-Wellek-Welch using the median had an accurate Type I error rate. Additionally, the mean-based Levene-Wellek-Welch test had a Type I error rate that was too conservative when $\bar{n} = 10$ and positively paired with the variances.

Finally, the Levene-Wellek-Welch using the mean had a Type I error rate too liberal in the largest equal sample size condition. See Table 38.

Power. When variances were exactly equal, this was a power condition for the equivalence-based procedures because the combined difference in the variances was less than the equivalence interval. For $J = 2$ and $\varepsilon = .25$, power was approximately 60% at the largest sample size (see Table 39). Thus, power for these procedures was not affected by the positive skewness of the group distributions. When $J = 4$ and $\varepsilon = .25$ (see Table 40), power was 40% to 41% for the median-based equivalence procedures and 25% to 26% for the mean-based counterparts in the largest sample size conditions. For $J = 2$ and $\varepsilon = .50$, the pattern was similar (and nearly identical to the results of our normally distributed groups) and reached over 99% power in the largest sample size condition (see Table 41). For $J = 4$ and $\varepsilon = 50$ (see Table 42), power approached 99% for the median based-procedures when variances were exactly equal. However, for the mean-based procedures, power was slightly lower, at approximately 95%.

As evaluated with normally distributed groups, with a 1.3:1 variance ratio and $\varepsilon = .25$, the test statistic, ψ^2 was less than ε^2; therefore, the combined difference in the variances was less than the equivalence interval. Thus, this condition evaluated power for the equivalence procedures (i.e., probability of declaring equivalence when the variances were considered equivalent). When $J = 2$, power for the equivalence procedures was low at the small sample sizes, but increased as sample sizes increased, ranging from 37% to 45% in the largest sample size conditions (see Table 43). For $J = 4$ (see Table 44), the same pattern of results was observed, with power ranging between 19% and 34% in the largest sample size conditions.

For a 2:1 variance ratio and $\varepsilon = .50$, the combined difference in the variances was less than the equivalence interval; thus, this condition also assessed power for the four equivalence procedures. When $J = 2$ and sample sizes were equal, the mean-based procedures had comparable power rates, reaching 68% at the largest sample size condition. The median-based procedures had higher power than their mean-based counterparts, reaching almost 85% in the largest sample size condition. When sample sizes were unequal and positively paired with the variances, again, the mean-based procedures had lower power than the median-based procedures. However, the Levene-Wellek versions had slightly higher power than the Welch-adjusted versions. Conversely, when unequal sample sizes were negatively paired with the variances, the Welch-adjusted procedures had a power advantage, and the Levene-Wellek-Welch using the median had the highest power at approximately 89%. See Table 45. When $J = 4$, a similar pattern of results was obtained. When sample sizes were equal, the median-based procedures had the highest power at all sample sizes, reaching between 73% and 75% in the largest sample size condition. This power advantage for the median-based tests was also observed when unequal sample sizes were positively paired with variances (71% to 82% at $\bar{n} = 100$), and when unequal sample sizes were negatively paired with variances (71% to 84% at $\bar{n} = 100$). See Table 46.

False declarations of equivalence. When the variance ratio was 6:1, the combined difference in the variances was greater than the equivalence interval so that ψ^2

> ε^2. Thus, any rejections of the null hypothesis were errors for this condition. As seen in Table 47 and 48 for $\varepsilon = .25$, there were no false declarations of equivalence for the equivalence procedures in the largest sample size conditions. Additionally, the error rates for the smaller sample size conditions never exceeded the empirical Type I error rates reported previously in this study. This result was obtained for both $J = 2$ and $J = 4$. When $\varepsilon = .50$, the error rates were fairly low when sample sizes were equal, or unequal sample sizes were positively paired with variance, and lowest when $J = 4$. The error rates were slightly higher for the negative pairing conditions, although they remained close to the empirical Type I error rates reported previously in this study. In the largest sample sizes conditions, the error rates across all conditions were at or nearly zero (see Tables 49 and 50).

Summary of Equivalence Procedures' Results

Give the scope of the conditions covered in the current study, a summary of the empirical Type I error rates for the equivalence procedures is given in Table 51, and a summary of the power conditions is in Table 52. The new Levene-Wellek procedure based on the absolute deviations from the mean was the poorest performing procedure in terms of Type I error rates. The proposed Levene-Wellek-Welch test based on the absolute deviations from the median was the best performing procedure in terms of Type I error rates. With regard to power, when the distributions of the groups were normal, all four of the procedures had comparable power rates. When the distributions were positively skewed, the procedures based on the absolute deviations from the median had a power advantage, and the Levene-Wellek-Welch based on the absolute deviations from the median had a slight power advantage over the non-Welch counterpart. These two procedures (the Levene-Wellek using the median and the Levene-Wellek-Welch using the median) had comparable power rates for the equal sample size conditions, but the median-based Levene-Wellek-Welch test had better power rates when unequal variances were positively paired or negatively paired with unequal sample sizes. It is also important to note that the Levene-Wellek using the median had more instances of inaccurate Type I error rates, which makes the corresponding power conditions less reliable[2]. Therefore, the general conclusion from the current study is that the new Levene-Wellek-Welch procedure based on the absolute deviations from the median was the best performing test statistic for assessing homogeneity of variances across the conditions tested. An applied example demonstrating use of this procedure is presented next.

Applied Example

This section presents a demonstration of how to use the best-performing equivalence-based homogeneity of variance test (in terms of power and Type I error rate) using a substantive example from psychological research and contrasts these results to the performance of the original Levene median-based test using the same data. This comparison achieves two goals: 1) to demonstrate the use of the new equivalence-based homogeneity of variance procedure; and 2) to further highlight the fundamental flaws of the original Levene-type difference-based tests for homogeneity of variances.

---

[2] As noted previously, iitalicized values in the tables indicate inaccurate Type I error rates and corresponding inaccurate power rates.

Data were taken from Arpin-Cribbie, Irvine, and Ritvo (2011). Participants were randomly assigned to one of three groups: no treatment, general stress management, or cognitive behavioural therapy (CBT). Participants were measured on various outcomes at pre-test and again following the intervention 11 weeks later (posttest). The overall sample size was 83. Of interest was ensuring that the three randomly assigned groups did not differ on baseline measures in terms of central tendency, but also to ensure that the dispersion of scores within each group was comparable between groups. The original study looked at equivalence of the groups on all pre-test measures, but the current example just tests for the equivalence of variances on the baseline measure of the Perfectionism Cognitions Inventory (PCI; Flett et al., 1998) for the purpose of demonstration. Descriptive statistics for the three groups on this measure are in Table 53. The variances for the stress management group and the no treatment group were similar, but the CBT group variance was more than two times larger than the Stress Management group. A visual depiction of the spread of the scores for each group is in Figure 2 and shows that the spread of scores on the PCI measure was greater in the CBT group than in the stress management group.

The original Levene test indicated that there were no statistically significant differences among the group variances, $F = 2.50, p = .09$. The Levene test using the median (i.e., the Brown-Forsythe modification of the Levene test) also indicated that there were no statistically significant differences in group variances, $F = 2.10, p = .13$. Next, the newly developed median-based Levene-Wellek-Welch equivalence test was used, setting an equivalence interval such that approximately 20% difference in the variances was considered trivial (i.e., $\varepsilon = .50$). This equivalence test found that the variances were not significantly equivalent ($\psi^{2*} = 0.16$) > ($\psi^{2*}_{critical} = .09$). Thus, the difference-based tests found that the group variances were not different, but the equivalence test indicated that the group variances were not equivalent. The reason this occurred was discussed in the introduction: Because the sample sizes of the groups were relatively small, power to detect even non-trivial differences in the variances was reduced. Consequently, the difference-based procedures declared non-trivial differences between the group variances equivalent, whereas the equivalence test found that the difference in these group variances exceeded the pre-specified equivalence limit. In other words, if a researcher were to use the traditional difference-based procedure like the Levene test to evaluate the equality of the group variances, they would come to the wrong conclusion in this situation. Using the new equivalence-based procedure ensures that researchers who are evaluating variance equality have a valid test for assessing this problem, and will, therefore, reach accurate conclusions regarding the equality of their group variances.

Discussion

Results of the simulation study demonstrated the backward nature of the traditional difference-based procedure for assessing equality of group variances. Specifically, power for detecting equivalence was in the wrong direction such that increased sample sizes resulted in decreased power for detecting equivalence of the variances. Additionally, the simulation results helped demonstrate that the point-null hypothesis is impractical, which is important because small differences in the variances are often inconsequential and are expected. Even though the difference-based tests often failed to reject the null hypothesis when there were small differences in the variances, this was because they were not performing correctly. As sample sizes increased, the chances of declaring small differences in the variances as important differences increased. Conversely, large and arguably important differences in the group variances were often declared equivalent by the difference-based tests when sample sizes were small.

Given these problems with the traditional difference-based procedures, equivalence-based procedures are more appropriate if the research goal is to evaluate variance equality. Equivalence tests align the research hypothesis of variance equality with the alternate hypothesis, so that power to detect equivalence and reject the null hypothesis increases with sample size, as expected when using null-hypothesis testing procedures. Additionally, the use of an interval hypothesis, rather than a point-null hypothesis, allows researchers to dictate how much or little overlap in the variances might be important. In general, small differences in the variances are expected and usually are inconsequential, so a test designed to assess approximate equality is far more practical than tests that evaluate exact equivalence (i.e., zero difference among the group variances). However, no such procedures existed had been developed prior to this research. This study developed four procedures, combining existing procedures for variance equality and equivalence testing logic.

Empirical Type I error rates for the equivalence procedures indicated the probability of that procedure rejecting the null hypothesis of variance heterogeneity when the null hypothesis was actually true. In other words, the differences in the variances were at or outside the bounds of the equivalence interval. The current study found that when sample sizes were equal and the dependent variable was normally distributed, all of the new equivalence procedures maintained the Type I error rates close to the nominal level (i.e., within Bradley's (1978) liberal limits for robustness). In addition, positively paired unequal variances and sample sizes (i.e., largest variance paired with largest sample size) and negatively paired unequal variances and sample sizes (i.e., largest variance paired with smallest sample size) under normally distributed conditions had minimal effect on the Type I error rates of two versions of the Levene-Wellek-Welch tests (i.e., one version based on the absolute deviations from the mean and one version based on the absolute deviations from the median). There was a slight tendency for the mean-based procedures (i.e., the Levene-Wellek and the Levene-Wellek-Welch on the absolute deviations from the mean) to have Type I error rates that were more conservative than their median-based counterparts (i.e., the Levene-Wellek and the Levene-Wellek-Welch using the absolute deviations from the median).

When the distribution of the dependent variable was positively skewed, in general the median-based equivalence procedures outperformed the mean-based versions of the tests in terms of Type I error rates. This finding was expected, as it is well-known that the median is a more accurate measure of central tendency than the mean for non-normally distributed data.

Power for the equivalence procedures indicated the probability of rejecting the null hypothesis of variance heterogeneity when the null hypothesis was false in the population. As expected, power for all of the equivalence procedures was higher when the equivalence limit was larger (i.e., power was higher when $\varepsilon = .50$ than when $\varepsilon = .25$), and power rates for all procedures increased as group sample sizes increased. In addition, power for all procedures wesa generally higher when $J = 2$ versus $J = 4$. Unequal sample sizes and positive or negative pairings of the sample sizes and unequal variances had little effect on power rates compared to the equal sample size conditions.

Regarding the power rates of specific procedures, the two median-based procedures outperformed both of the mean-based equivalence procedures in most conditions. The Levene-Wellek-Welch using the median had a power advantage when sample sizes were unequal, but was comparable to, or outperformed by the Levene-Wellek using the median when sample sizes were equal. Finally, in some conditions the Levene-Wellek using the median was outperformed by the Levene-Wellek-Welch using the mean in terms of power rates in normally distributed conditions.

Based on the Type I error rates and power results, the median-based Levene-Wellek-Welch equivalence test was the most robust procedure across the conditions tested, with consistently higher power over the other procedures. Therefore, it is recommended to researchers who wish to assess equality of group variances.

Limitations

Although this study attempted to be as comprehensive as possible, there are many other conditions that could be tested to further evaluate the new equality of variances equivalence procedures. It is difficult to test every data scenario a researcher might encounter. However, the results supported the objectives of this study, in that the fundamental flaws of traditional difference-based tests were revealed, and the newly developed equivalence-based procedures were subjected to various data conditions to evaluate their robustness. In addition, the conditions selected for this study represent common data analytic conditions in psychology, and the pattern of results should generalize across other data scenarios.

A broader limitation of the current research concerns the ease of implementation of the recommended equivalence procedure. Typically, modern robust statistical procedures are not readily available at the time of development in the popular statistical software programs, such as SPSS and SAS. However, the open-source software program, R, allows researchers to implement their own user-generated functions. Thus, to address this limitation, a function for the Levene-Wellek-Welch procedure based on the absolute deviations from the median was developed for R to facilitate the use of this procedure, and can be found in Appendix A.

Future Directions

Future simulation research should expand on the data conditions tested in this study to further evaluate the newly developed procedures. For example, different distribution shapes, more group sizes, and different variance conditions and sample size to variance combinations should be evaluated. In addition, evaluating a trimmed means or boostrapped version of the Levene-Wellek-Welch equivalence test could be a useful extension of this research.

More broadly, future research should include discussions regarding the importance of examining the variances associated with one's data and the implications of homogeneity or heterogeneity of group variances. For example, Bryk and Raudenbush

(1988) suggest that heterogeneity within groups can indicate the presence of an interaction between person characteristics and group membership. Alternatively, homogeneity of group variances in the presence of mean differences might indicate that, even though the groups may represent different populations, they do share similarities in composition that might be interesting to explore. However, discussions regarding variance homogeneity or heterogeneity from a theoretical perspective are not as popular in psychology as other disciplines. For example, Sagrestano, Heavey, and Christensen (1998) argue that different perspectives in social psychology tend to focus on different aspects of variability. An individual differences approach focuses on between-group variability while neglecting within-group variability, whereas a social structural approach focuses on within-group variability but may neglect between-group differences. Future research might be focused on unifying these approaches, such that comparing the within-group variability between groups becomes an important research consideration, thus, methodological support for these research goals will be needed.

Finally, the equivalence-based tests for homogeneity of variances can be expanded to test the equivalence of variances among groups in more complex designs. For instance, instead of using nested chi-square procedures to evaluate the equivalence of the variance parameters across multiple groups in structural equation models or latent growth curve models, one could use equivalence-based procedures.

Conclusions

This study provided evidence to researchers regarding the problems with assessing equality of variances with difference-based tests. Most notably, difference-based tests assess equality of variances from the wrong perspective, encouraging researchers to "accept" the null hypothesis. However, null-hypothesis testing is not meant to be used in this way, and this misuse results in power for detecting equivalence to decrease as sample sizes increase. In addition, large differences in the variances are often declared equivalent when group sample sizes are small. Previous research in the equivalence testing literature recommends that researchers should not use a test developed to evaluate differences when the primary research question deals with equivalence. Thus, four novel equivalence procedures to assess equality of variances were proposed. Of these procedures, the Levene-Wellek-Welch equivalence of variances test based on the absolute deviations from the median was the best-performing test statistic in terms of accurate Type I error rates and highest power for detecting equivalence across the conditions evaluated. Therefore, researchers should evaluate hypotheses of equivalent variances using this median-based Levene-Wellek-Welch equivalence test.

References

Arpin-Cribbie, C., Irvine, J., & Ritvo, P. (2011). Web-based cognitive-behavioral therapy for perfectionism: A randomized controlled trial. *Psychotherapy Research, 22,* 194-207.

Bradley, J.V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 81,* 144-152.

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

Bryk, A. S. & Raudenbush, S. W. (1988). Heterogeneity of variances in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin, 104,* 396-404.

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,* 351-361.

Cribbie, R. A., Arpin-Cribbie, C. A., & Gruman, J. A. (2010). Equivalence tests for one-way independent groups designs. *Journal of Experimental Education, 78,* 1-13.

Cribbie, R. A., Gruman, J. A., & Arpin-Cribbie, C. A. (2004). Recommendations for applying tests of equivalence. *Journal of Clinical Psychology, 60,* 1-10.

Gastwirth, J.L., Gel, Y.R., & Miao, W. (2009). The impact of Levene's test of equality of variances on statistical theory and practice. *Statistical Science, 24,* 343-360.

Goertzen, J. R. & Cribbie, R. A. (2010). Detecting a lack of association: An equivalence testing approach. *British Journal of Mathematical and Statistical Psychology, 63*, 527-537.

Keselman, H.J., Games, P.A., & Clinch, J.J. (1979). Tests for homogeneity of variance. *Communications in Statistics – Simulations and Computation, B8,* 113-129.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C. & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68,* 350-386.

Keselman, H.J., Wilcox, R.R., Algina, J., Othman, A.R., & Fradette, K. (2008). A comparative study of robust test for spread: Asymmetric trimming strategies. *Bristish Journal of Mathematical and Statistical Psychology, 61,* 235-253.

Levene, H. (1960). Robust tests of equality of variances. In *Contributions to Probability and Statistics, Essays in Honor of Harold Hotelling*, I. Olkin et al. (Eds.). Stanford University Press, pp. 278-292.

Lim, T-S. & Loh, W-Y. (1996). A comparison of tests of equality of variances. *Computational Statistics and Data Analysis, 22,* 287-301.

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

Othman, A. R., Keselman, H. J., Padmanabhan, A.R., Wilcox, R. R., Algina, J., Fradette, K. (2004). Comparing measures of the "typical" score across treatment groups. *British Journal of Mathematical and Statistical Psychology, 57,* 215-234.

Pahnke, M. D., Trinity, J. D., Zachwieja, J. J., Stofan, J. R., Hiller, D. H., & Coyle, E. F. (2010). Serum sodium concentration changes are related to fluid balance and sweat sodium loss. *Medicine and Science in Sports and Exercise.* 42, 1669-1680.

Parra-Frutos, I. (2009). The behaviour of the modified Levene's test when data are not normally distributed. *Computational Statistics, 24,* 671-693.

R Development Core Team (2010). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna: Austria. ISBN 3-900051-07-0, http://www.R-project.org.

Robinson, A.P., Duursma, R.A., & Marshall, J.D. (2005). A regression-based equivalence test for model validation: Shifting the burden of proof. *Tree Physiology 25*, 903–913.

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113,* 553-565.

Sagrestano, L.M., Heavey, C. L., Christensen, A. (1998). Theoretical approaches to understanding sex differences and similarities in conflict behavior. In *Sex differences and similarities in communication: Critical essays and empirical investigations of sex and gender in interaction.* Canary, D. J. & Dindia, K. (Eds). Mahwah, NJ: Lawrence Erlbaum Associates Publishers, pp. 287-302.

Salgado, J. F. (1995). Situational specificity and within-setting validity variability. *Journal of Occupational and Organizational Psychology, 68,* 123-132.

Schuirmann, D.J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics, 15,* 657-680.

Seaman, M. A. & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods, 3,* 403-411.

van Raalte, A. A., Kunst, A. E., Deboosere, P., Leinsalu, M., Lundberg. O., Martikainen, P., Strand, B. H., Artnik, B., Wojtyniak, B., & Mackenbach, J. P. (2011). More variation in lifespan in lower educated groups: Evidence from 10 European countries. *International Journal of Epidemiology, 40*, 1703-14.

Welch, B.L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika, 38,* 330-336.

Wellek, S. (2003). *Testing statistical hypotheses of equivalence.* New York: Chapman & Hall/CRC

Wilcox, R.R., Charlin, V.L., & Thompson, K.L. (1986). New Monte Carlo results on the robustness of the ANOVA F, W, and F* statistics. *Communications in Statistics Simulation and Computation, 15,* 933-943.

Yin, T. S. & Othman, A. R. (2009). When does the pooled variances *t*-test fail? *African Journal of Mathematics and Computer Science Research, 2,* 56-62.

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

Table 1.

Distribution shapes and samples sizes for the simulation study

λ	n_j ($J = 2$)	n_j ($J = 4$)
Normal	10,10	10,10,10,10
χ^2 (3 *df*)	5,15	5,8,12,15
	15,5	15,12,8,5
	25,25	25,25,25,25
	18,32	18,22,28,32
	32,18	32,28,22,18
	50,50	50,50,50,50
	25,75	25,40,60,75
	75,25	75,60,40,25
	100,100	100,100,100,100
	80,120	50,80,120,150
	120,80	150,120,80,50

Table 2.

Equivalence intervals and population variances used in the simulation study.

EI	σ^2 ($J = 2$)	σ^2 ($J = 4$)
	Normal Distribution	
.25	1, 1.721	1, 1.224, 1.448, 1.672
	1, 1	1, 1, 1, 1
	1, 6	1, 3, 4, 6
	1, 1.3	1, 1.1, 1.2, 1.3
.50	1, 3.1	1, 1.642, 2.284, 2.926
	1, 1	1, 1, 1, 1
	1, 6	1, 3, 4, 6
	1, 2	1, 1.33, 1.66, 2
	χ^2 Distribution (3 *df*)	
.25	1, 1.89	1, 1.28, 1.56, 1.84
	1, 1	1, 1, 1, 1
	1, 6	1, 3, 4, 6
	1, 1.3	1, 1.1, 1.2, 1.3
.50	1, 3.7	1, 1.85, 2.70, 3.55
	1, 1	1, 1, 1, 1
	1, 6	1, 3, 4, 6
	1, 2	1, 1.33, 1.66, 2

Table 3.

Probability of declaring equivalence (1-α) for difference-based procedures; Normal distribution; J = 2; σ_j^2 = 1, 1

n_j	Lev_mean	Lev_mdn	LevWelch_mean	LevWelch_mdn
		Equal Sample Sizes		
10, 10	.9398	.9646	.9444	.9685
25, 25	.9246	.9583	.9432	.9589
50, 50	.9452	.9509	.9452	.9511
100, 100	.9489	.9508	.9489	.9508
		Unequal Sample Sizes – Positive Pairings		
5, 15	.9404	.9770	.8881	.9348
18, 32	.9444	.9576	.9412	.9536
25, 75	.9466	.9532	.9375	.9474
80, 120	.9522	.9557	.9518	.9544
		Unequal Sample Sizes – Negative Pairings		
15, 5	.9449	.9812	.8920	.9364
32, 18	.9453	.9575	.9420	.9551
75, 25	.9512	.9588	.9436	.9517
120, 80	.9495	.9526	.9491	.9522

Note: Lev_mean = Levene's original mean-based test; Lev_mdn = Brown-Forsythe version of Levene's test using median; LevWelch_mean = Welch-version of Levene's original mean-based test; LevWelch_mdn = Welch version of the Levene test based on median

Table 4.

Probability of declaring equivalence (1-α) for difference-based procedures; Normal distribution; J = 4; σ_j^2 = 1, 1, 1, 1

n_j	Lev_mean	Lev_mdn	LevWelch_mean	LevWelch_mdn
		Equal Sample Sizes		
10, 10, 10, 10	.9326	.9666	.9263	.9580
25, 25, 25, 25	.9437	.9641	.9375	.9583
50, 50, 50, 50	.9473	.9574	.9449	.9557
100, 100, 100, 100	.9440	.9493	.9427	.9469
		Unequal Sample Sizes – Positive Pairings		
5, 8, 12, 15	.9360	.9736	.8876	.9453
18, 22, 28, 32	.9448	.9615	.9371	.9551
25, 40, 60, 75	.9450	.9559	.9381	.9491
50, 80, 120, 150	.9488	.9527	.9433	.9480
		Unequal Sample Sizes – Negative Pairings		
15, 12, 8, 5	.9353	.9741	.8844	.9393
32, 28, 22, 18	.9473	.9574	.9449	.9557
75, 60, 40, 25	.9477	.9581	.9344	.9477
150, 120, 80, 50	.9490	.9546	.9443	.9500

Note: Lev_mean = Levene's original mean-based test; Lev_mdn = Brown-Forsythe version of Levene's test using median; LevWelch_mean = Welch-version of Levene's original mean-based test; LevWelch_mdn = Welch version of the Levene test based on median

Table 5.

Probability of declaring equivalence for difference-based procedures; Normal distribution; $J = 2$; $\sigma_j^2 = 1,\ 1.3$

n_j	Lev_mean	Lev_mdn	LevWelch_mean	LevWelch_mdn
		Equal Sample Sizes		
10, 10	.9282	.9564	.9326	.9595
25, 25	.9030	.9250	.9035	.9262
50, 50	.8622	.8764	.8626	.8766
100, 100	.7778	.7876	.7779	.7877
		Unequal Sample Sizes – Positive Pairings		
5, 15	.9403	.9730	.8520	.9010
18, 32	.9116	.9262	.8844	.9011
25, 75	.8975	.9020	.8397	.8504
80, 120	.7844	.7898	.7635	.7694
		Unequal Sample Sizes – Negative Pairings		
15, 5	.9199	.9723	.9114	.9548
32, 18	.9078	.9292	.9308	.9524
75, 25	.8751	.8995	.9175	.9419
120, 80	.7708	.7826	.7908	.8030

Note: Lev_mean = Levene's original mean-based test; Lev_mdn = Brown-Forsythe version of Levene's test using median; LevWelch_mean = Welch-version of Levene's original mean-based test; LevWelch_mdn = Welch version of the Levene test based on median

Table 6.

Probability of declaring equivalence for difference-based procedures; Normal distribution; J = 4; σ_j^2 = 1, 1.1, 1.2, 1.3

n_j	Lev_mean	Lev_mdn	LevWelch_mean	LevWelch_mdn
		Equal Sample Sizes		
10, 10, 10, 10	.9249	.9659	.9195	.9591
25, 25, 25, 25	.9173	.9438	.9122	.9387
50, 50, 50, 50	.8912	.9060	.8899	.9053
100, 100, 100, 100	.8301	.8422	.8310	.8418
		Unequal Sample Sizes – Positive Pairings		
5, 8, 12, 15	.9389	.9757	.8610	.9247
18, 22, 28, 32	.9263	.9458	.8944	.9202
25, 40, 60, 75	.9146	.9243	.8752	.8879
50, 80, 120, 150	.8648	.8685	.8245	.8322
		Unequal Sample Sizes – Negative Pairings		
15, 12, 8, 5	.9182	.9696	.8990	.9529
32, 28, 22, 18	.9201	.9463	.9317	.9546
75, 60, 40, 25	.8966	.9204	.9228	.9436
150, 120, 80, 50	.8511	.8660	.8862	.9009

Note: Lev_mean = Levene's original mean-based test; Lev_mdn = Brown-Forsythe version of Levene's test using median; LevWelch_mean = Welch-version of Levene's original mean-based test; LevWelch_mdn = Welch version of the Levene test based on median

Table 7.

Probability of declaring equivalence for difference-based procedures; Normal distribution; $J = 2$; $\sigma_j^2 = 1,\ 2$

n_j	Lev_mean	Lev_mdn	LevWelch_mean	LevWelch_mdn
		Equal Sample Sizes		
10, 10	.8498	.9012	.8599	.9112
25, 25	.6661	.7123	.6691	.7156
50, 50	.4020	.4250	.4030	.4267
100, 100	.1104	.1162	.1104	.1163
		Unequal Sample Sizes – Positive Pairings		
5, 15	.9068	.9469	.7309	.8051
18, 32	.7043	.7334	.6148	.6507
25, 75	.5530	.5606	.4137	.4310
80, 120	.1229	.1273	.1037	.1087
		Unequal Sample Sizes – Negative Pairings		
15, 5	.8474	.9435	.9188	.9792
32, 18	.6659	.7168	.7529	.8074
75, 25	.4909	.5381	.6396	.7035
120, 80	.1211	.1290	.1448	.1546

Note: Lev_mean = Levene's original mean-based test; Lev_mdn = Brown-Forsythe version of Levene's test using median; LevWelch_mean = Welch-version of Levene's original mean-based test; LevWelch_mdn = Welch version of the Levene test based on median

Table 8.

Probability of declaring equivalence for difference-based procedures; Normal distribution; J = 4; σ_j^2 = 1, 1.33, 1.66, 2

n_j	Lev_mean	Lev_mdn	LevWelch_mean	LevWelch_mdn
		Equal Sample Sizes		
10, 10, 10, 10	.8711	.9314	.8669	.9230
25, 25, 25, 25	.7560	.8138	.7508	.8054
50, 50, 50, 50	.5236	.5581	.5131	.5509
100, 100, 100, 100	.1939	.2064	.1824	.1958
		Unequal Sample Sizes – Positive Pairings		
5, 8, 12, 15	.9052	.9552	.7758	.8673
18, 22, 28, 32	.7705	.8170	.7032	.7500
25, 40, 60, 75	.6257	.6481	.5121	.5375
50, 80, 120, 150	.3040	.3127	.2233	.2318
		Unequal Sample Sizes – Negative Pairings		
15, 12, 8, 5	.8557	.9501	.8967	.9675
32, 28, 22, 18	.7384	.7976	.7920	.8515
75, 60, 40, 25	.5383	.5904	.6253	.6847
150, 120, 80, 50	.2166	.2400	.2675	.2976

Note: Lev_mean = Levene's original mean-based test; Lev_mdn = Brown-Forsythe version of Levene's test using median; LevWelch_mean = Welch-version of Levene's original mean-based test; LevWelch_mdn = Welch version of the Levene test based on median

Table 9.

Probability of declaring equivalence for difference-based procedures; Normal distribution; J = 2; σ_j^2 = 1, 6

n_j	Lev_mean	Lev_mdn	LevWelch_mean	LevWelch_mdn
		Equal Sample Sizes		
10, 10	.4024	.5305	.4351	.5728
25, 25	.0272	.0398	.0288	.0422
50, 50	.0001	.0001	.0001	.0001
100, 100	.0000	.0000	.0000	.0000
		Unequal Sample Sizes – Positive Pairings		
5, 15	.6690	.7832	.2750	.3756
18, 32	.0473	.0608	.0188	.0235
25, 75	.0014	.0016	.0000	.0000
80, 120	.0000	.0000	.0000	.0000
		Unequal Sample Sizes – Negative Pairings		
15, 5	.4393	.6569	.8164	.9952
32, 18	.0411	.0567	.0956	.1377
75, 25	.0028	.0036	.0140	.0215
120, 80	.0000	.0000	.0000	.0000

Note: Lev_mean = Levene's original mean-based test; Lev_mdn = Brown-Forsythe version of Levene's test using median; LevWelch_mean = Welch-version of Levene's original mean-based test; LevWelch_mdn = Welch version of the Levene test based on median

Table 10.

Probability of declaring equivalence for difference-based procedures; Normal distribution; J = 4; $\sigma_j^2 = 1, 3, 4, 6$

n_j	Lev_mean	Lev_mdn	LevWelch_mean	LevWelch_mdn
		Equal Sample Sizes		
10, 10, 10, 10	.5569	.7216	.4756	.6316
25, 25, 25, 25	.0788	.1165	.0372	.0555
50, 50, 50, 50	.0004	.0004	.0001	.0001
100, 100, 100, 100	.0000	.0000	.0000	.0000
		Unequal Sample Sizes – Positive Pairings		
5, 8, 12, 15	.7458	.8523	.3934	.5557
18, 22, 28, 32	.1366	.1826	.0424	.0580
25, 40, 60, 75	.0122	.0146	.0007	.0013
50, 80, 120, 150	.0000	.0000	.0000	.0000
		Unequal Sample Sizes – Negative Pairings		
15, 12, 8, 5	.4622	.7015	.6286	.8537
32, 28, 22, 18	.0576	.0870	.0475	.0710
75, 60, 40, 25	.0004	.0007	.0006	.0008
150, 120, 80, 50	.0000	.0000	.0000	.0000

Note: Lev_mean = Levene's original mean-based test; Lev_mdn = Brown-Forsythe version of Levene's test using median; LevWelch_mean = Welch-version of Levene's original mean-based test; LevWelch_mdn = Welch version of the Levene test based on median

Table 11.

Probability of declaring equivalence (1-α) for difference-based procedures; χ^2 *distribution (3 df); J = 2;* $\sigma_j^2 = 1,\ 1$

n_j	Lev_mean	Lev_mdn	LevWelch_mean	LevWelch_mdn
		Equal Sample Sizes		
10, 10	.8726	.9536	.8821	.9578
25, 25	.8820	.9531	.8840	.9545
50, 50	.8792	.9503	.8799	.9510
100, 100	.8872	.9504	.8873	.9507
		Unequal Sample Sizes – Positive Pairings		
5, 15	.8764	.9698	.8185	.9172
18, 32	.8816	.9529	.8765	.9473
25, 75	.8909	.9553	.8744	.9348
80, 120	.8892	.9517	.8899	.9491
		Unequal Sample Sizes – Negative Pairings		
15, 5	.8814	.9674	.8135	.9110
32, 18	.8841	.9522	.8771	.9456
75, 25	.8946	.9579	.8731	.9344
120, 80	.8875	.9508	.8858	.9502

Note: Lev_mean = Levene's original mean-based test; Lev_mdn = Brown-Forsythe version of Levene's test using median; LevWelch_mean = Welch-version of Levene's original mean-based test; LevWelch_mdn = Welch version of the Levene test based on median

Table 12.

Probability of declaring equivalence (1-α) for difference-based procedures; χ^2 distribution (3 df); J = 4; σ_j^2 = 1, 1, 1, 1

n_j	Lev_mean	Lev_mdn	LevWelch_mean	LevWelch_mdn
		Equal Sample Sizes		
10, 10, 10, 10	.8178	.9562	.8056	.9357
25, 25, 25, 25	.8280	.9619	.8129	.9419
50, 50, 50, 50	.8355	.9529	.8242	.9388
100, 100, 100, 100	.8355	.9535	.8297	.9469
		Unequal Sample Sizes – Positive Pairings		
5, 8, 12, 15	.8123	.9637	.7580	.9111
18, 22, 28, 32	.8307	.9566	.8079	.9303
25, 40, 60, 75	.8407	.9575	.8135	.9306
50, 80, 120, 150	.8436	.9536	.8219	.9374
		Unequal Sample Sizes – Negative Pairings		
15, 12, 8, 5	.8204	.9612	.7637	.9129
32, 28, 22, 18	.8395	.9567	.8157	.9354
75, 60, 40, 25	.8380	.9549	.8059	.9240
150, 120, 80, 50	.8396	.9521	.8183	.9336

Note: Lev_mean = Levene's original mean-based test; Lev_mdn = Brown-Forsythe version of Levene's test using median; LevWelch_mean = Welch-version of Levene's original mean-based test; LevWelch_mdn = Welch version of the Levene test based on median

Table 13.

Probability of declaring equivalence for difference-based procedures; χ^2 distribution (3 df); $J = 2$; $\sigma_j^2 = 1,\ 1.3$

n_j	Lev_mean	Lev_mdn	LevWelch_mean	LevWelch_mdn
		Equal Sample Sizes		
10, 10	.8644	.9480	.8754	.9532
25, 25	.8460	.9349	.8491	.9366
50, 50	.8210	.9075	.8218	.9081
100, 100	.7597	.8561	.7598	.8563
		Unequal Sample Sizes – Positive Pairings		
5, 15	.8929	.9717	.7774	.8830
18, 32	.8626	.9403	.8260	.9035
25, 75	.8545	.9329	.7754	.8556
80, 120	.7671	.8616	.7437	.8338
		Unequal Sample Sizes – Negative Pairings		
15, 5	.8664	.9638	.8528	.9431
32, 18	.8434	.9245	.8704	.9498
75, 25	.8252	.9084	.8599	.9463
120, 80	.7530	.8483	.7755	.8727

Note: Lev_mean = Levene's original mean-based test; Lev_mdn = Brown-Forsythe version of Levene's test using median; LevWelch_mean = Welch-version of Levene's original mean-based test; LevWelch_mdn = Welch version of the Levene test based on median

Table 14.

Probability of declaring equivalence for difference-based procedures; χ^2 *distribution (3 df); J = 4;* $\sigma_j^2 = 1,\ 1.1,\ 1.2,\ 1.3$

n_j	Lev_mean	Lev_mdn	LevWelch_mean	LevWelch_mdn
		Equal Sample Sizes		
10, 10, 10, 10	.8138	.9552	.8017	.9356
25, 25, 25, 25	.8039	.9457	.7856	.9240
50, 50, 50, 50	.8355	.9529	.8242	.9388
100, 100, 100, 100	.8355	.9535	.8297	.9469
		Unequal Sample Sizes – Positive Pairings		
5, 8, 12, 15	.8266	.9646	.7299	.8943
18, 22, 28, 32	.8178	.9499	.7631	.9026
25, 40, 60, 75	.8407	.9575	.8135	.9306
50, 80, 120, 150	.8436	.9536	.8219	.9374
		Unequal Sample Sizes – Negative Pairings		
15, 12, 8, 5	.8026	.9581	.7812	.9278
32, 28, 22, 18	.7965	.9388	.8034	.9358
75, 60, 40, 25	.8380	.9549	.8059	.9240
150, 120, 80, 50	.8396	.9521	.8183	.9336

Note: Lev_mean = Levene's original mean-based test; Lev_mdn = Brown-Forsythe version of Levene's test using median; LevWelch_mean = Welch-version of Levene's original mean-based test; LevWelch_mdn = Welch version of the Levene test based on median

Table 15.

Probability of declaring equivalence for difference-based procedures; χ^2 *distribution (3 df); J = 2;* $\sigma_j^2 = 1,\ 2$

n_j	Lev_mean	Lev_mdn	LevWelch_mean	LevWelch_mdn
		Equal Sample Sizes		
10, 10	.8030	.9090	.8175	.9205
25, 25	.6740	.8080	.6797	.8127
50, 50	.4998	.6267	.5009	.6292
100, 100	.2476	.3345	.2480	.3349
		Unequal Sample Sizes – Positive Pairings		
5, 15	.8731	.9568	.6770	.8122
18, 32	.7205	.8455	.6375	.7515
25, 75	.5991	.7453	.4663	.5672
80, 120	.2631	.3738	.2366	.3275
		Unequal Sample Sizes – Negative Pairings		
15, 5	.7930	.9342	.8704	.9712
32, 18	.6783	.8006	.7556	.8854
75, 25	.5552	.6742	.6758	.8280
120, 80	.2516	.3426	.2779	.3896

Note: Lev_mean = Levene's original mean-based test; Lev_mdn = Brown-Forsythe version of Levene's test using median; LevWelch_mean = Welch-version of Levene's original mean-based test; LevWelch_mdn = Welch version of the Levene test based on median

Table 16.

Probability of declaring equivalence for difference-based procedures; χ^2 distribution (3 df); J = 4; σ_j^2 = 1, 1.33, 1.66, 2

n_j	Lev_mean	Lev_mdn	LevWelch_mean	LevWelch_mdn
Equal Sample Sizes				
10, 10, 10, 10	.7640	.9300	.7523	.9068
25, 25, 25, 25	.6768	.8746	.6560	.8451
50, 50, 50, 50	.5210	.7394	.5067	.7105
100, 100, 100, 100	.2860	.4673	.2750	.4464
Unequal Sample Sizes – Positive Pairings				
5, 8, 12, 15	.8044	.9516	.6437	.8359
18, 22, 28, 32	.7070	.8877	.6186	.7963
25, 40, 60, 75	.6142	.8184	.4845	.6590
50, 80, 120, 150	.3869	.5949	.2982	.4537
Unequal Sample Sizes – Negative Pairings				
15, 12, 8, 5	.7253	.9336	.7932	.9492
32, 28, 22, 18	.6560	.8567	.6940	.8761
75, 60, 40, 25	.5252	.7419	.6097	.8244
150, 120, 80, 50	.3119	.4997	.3695	.5852

Note: Lev_mean = Levene's original mean-based test; Lev_mdn = Brown-Forsythe version of Levene's test using median; LevWelch_mean = Welch-version of Levene's original mean-based test; LevWelch_mdn = Welch version of the Levene test based on median

Table 17.

Probability of declaring equivalence for difference-based procedures; χ^2 distribution (3 df); J = 2; $\sigma_j^2 = 1,\ 6$

n_j	Lev_mean	Lev_mdn	LevWelch_mean	LevWelch_mdn
		Equal Sample Sizes		
10, 10	.4672	.6840	.4982	.7190
25, 25	.1093	.1938	.1125	.2019
50, 50	.0051	.0107	.0052	.0109
100, 100	.0000	.0000	.0000	.0000
		Unequal Sample Sizes – Positive Pairings		
5, 15	.6877	.8774	.3307	.5036
18, 32	.1446	.2644	.0916	.1473
25, 75	.0302	.0622	.0114	.0190
80, 120	.0000	.0000	.0000	.0000
		Unequal Sample Sizes – Negative Pairings		
15, 5	.4874	.7259	.8154	.9909
32, 18	.1160	.1856	.1961	.3599
75, 25	.0234	.0374	.0640	.1383
120, 80	.0000	.0000	.0000	.0000

Note: Lev_mean = Levene's original mean-based test; Lev_mdn = Brown-Forsythe version of Levene's test using median; LevWelch_mean = Welch-version of Levene's original mean-based test; LevWelch_mdn = Welch version of the Levene test based on median

Table 18.

Probability of declaring equivalence for difference-based procedures; χ^2 *distribution (3 df); J = 4;* σ_j^2 = 1, 3, 4, 6

n_j	Lev_mean	Lev_mdn	LevWelch_mean	LevWelch_mdn
		Equal Sample Sizes		
10, 10, 10, 10	.5150	.8037	.4544	.7129
25, 25, 25, 25	.1582	.3780	.1098	.2441
50, 50, 50, 50	.0119	.0442	.0075	.0218
100, 100, 100, 100	.0000	.0000	.0000	.0000
		Unequal Sample Sizes – Positive Pairings		
5, 8, 12, 15	.6825	.9047	.3623	.6049
18, 22, 28, 32	.2387	.4988	.1203	.2430
25, 40, 60, 75	.0671	.1834	.0244	.0518
50, 80, 120, 150	.0007	.0041	.0001	.0007
		Unequal Sample Sizes – Negative Pairings		
15, 12, 8, 5	.4431	.7770	.6080	.8863
32, 28, 22, 18	.1257	.3168	.1302	.2848
75, 60, 40, 25	.0085	.0263	.0099	.0305
150, 120, 80, 50	.0000	.0001	.0000	.0001

Note: Lev_mean = Levene's original mean-based test; Lev_mdn = Brown-Forsythe version of Levene's test using median; LevWelch_mean = Welch-version of Levene's original mean-based test; LevWelch_mdn = Welch version of the Levene test based on median

Table 19.

Type I error rates for equivalence procedures; Normal distribution; J = 2; $\varepsilon \approx .25$; $\sigma_j^2 = 1,\ 1.721\ (\psi^2 = \varepsilon^2)$

n_j	LW_mean	LW_median	LWW_mean	LWW_median
Equal Sample Sizes				
10, 10	.0475	.0551	.0475	.0551
25, 25	.0531	.0557	.0531	.0557
50, 50	.0485	.0509	.0485	.0509
100, 100	.0516	.0521	.0516	.0521
Unequal Sample Sizes – Positive Pairings				
5, 15	.0387	.0431	.0402	.0518
18, 32	.0450	.0466	.0458	.0478
25, 75	.0565	.0522	.0497	.0484
80, 120	.0511	.0511	.0518	.0531
Unequal Sample Sizes – Negative Pairings				
15, 5	.0463	.0585	.0483	.0687
32, 18	.0556	.0603	.0533	.0590
75, 25	*.0762*	*.0847*	.0551	.0617
120, 80	.0505	.0527	.0437	.0464

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 20.

Type I error rates for equivalence procedures; Normal distribution; J = 2; $\varepsilon \approx .50$; $\sigma_j^2 = 1,\ 3.1\ (\psi^2 = \varepsilon^2)$

n_j	LW_mean	LW_median	LWW_mean	LWW_median
Equal Sample Sizes				
10, 10	.0450	.0531	.0450	.0531
25, 25	.0423	.0467	.0423	.0467
50, 50	.0376	.0434	.0375	.0431
100, 100	.0358	.0423	.0357	.0420
Unequal Sample Sizes – Positive Pairings				
5, 15	.0394	.0390	.0409	.0459
18, 32	.0404	.0424	.0439	.0477
25, 75	.0543	.0552	.0463	.0495
80, 120	.0444	.0488	.0405	.0440
Unequal Sample Sizes – Negative Pairings				
15, 5	.0686	*.0886*	.0485	.0736
32, 18	.0612	.0720	.0445	.0527
75, 25	.0628	*.0759*	.0346	.0443
120, 80	.0306	.0349	.0304	.0350

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 21.

Type I error rates for equivalence procedures; Normal distribution; J = 4; $\varepsilon \approx .25$; $\sigma_j^2 = 1,\ 1.224,\ 1.448,\ 1.672\ (\psi^2 = \varepsilon^2)$

n_j	LW_mean	LW_median	LWW_mean	LWW_median
Equal Sample Sizes				
10, 10, 10, 10	.0429	.0606	.0437	.0611
25, 25, 25, 25	.0466	.0571	.0503	.0601
50, 50, 50, 50	.0461	.0497	.0479	.0525
100, 100, 100, 100	.0490	.0520	.0523	.0547
Unequal Sample Sizes – Positive Pairings				
5, 8, 12, 15	.0386	.0503	.0378	.0535
18, 22, 28, 32	.0463	.0517	.0449	.0513
25, 40, 60, 75	.0524	.0551	.0455	.0488
50, 80, 120, 150	.0624	.0623	.0476	.0490
Unequal Sample Sizes – Negative Pairings				
15, 12, 8, 5	.0403	.0629	.0443	.0749
32, 28, 22, 18	.0483	.0575	.0516	.0620
75, 60, 40, 25	.0478	.0591	.0503	.0615
150, 120, 80, 50	.0558	.0605	.0554	.0605

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 22.

Type I error rates for equivalence procedures; Normal distribution; J = 4; $\varepsilon \approx .50$;

$\sigma_j^2 = 1,\ 1.642,\ 2.284,\ 2.926\ (\psi^2 = \varepsilon^2)$

n_j	LW_mean	LW_median	LWW_mean	LWW_median
		Equal Sample Sizes		
10, 10, 10, 10	.0421	.0590	.0495	.0702
25, 25, 25, 25	.0407	.0511	.0525	.0675
50, 50, 50, 50	.0381	.0456	.0536	.0627
100, 100, 100, 100	.0394	.0441	.0533	.0608
		Unequal Sample Sizes – Positive Pairings		
5, 8, 12, 15	.0434	.0515	.0388	.0518
18, 22, 28, 32	.0448	.0514	.0493	.0541
25, 40, 60, 75	.0622	.0641	.0492	.0521
50, 80, 120, 150	*.0824*	*.0869*	.0488	.0523
		Unequal Sample Sizes – Negative Pairings		
15, 12, 8, 5	.0445	.0738	.0511	*.1014*
32, 28, 22, 18	.0409	.0507	.0559	.0736
75, 60, 40, 25	.0350	.0448	.0536	.0695
150, 120, 80, 50	.0259	.0324	.0510	.0610

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 23.

Probability of declaring equivalence (power) for equivalence procedrues; Normal distribution; J = 2; $\varepsilon \approx .25$; $\sigma_j^2 = 1,\ 1\ (\psi^2 < \varepsilon^2)$

n_j	LW_mean	LW_median	LWW_mean	LWW_median
		Equal Sample Sizes		
10, 10	.0661	.0719	.0661	.0720
25, 25	.1048	.1118	.1048	.1118
50, 50	.2254	.2308	.2254	.2308
100, 100	.5938	.6010	.5938	.6009
		Unequal Sample Sizes - Positive Pairings		
5, 15	.0512	.0622	.0569	.0780
18, 32	.1111	.1163	.1152	.1224
25, 75	.1923	.1952	.2207	.2302
80, 120	.5970	.6060	.6047	.6142
		Unequal Sample Sizes - Negative Pairings		
15, 5	.0551	.0623	.0614	.0785
32, 18	.1049	.1110	.1099	.1158
75, 25	*.1937*	*.1989*	.2203	.2292
120, 80	.5985	.6047	.6068	.6141

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 24.

Probability of declaring equivalence (power) for equivalence procedures; Normal distribution; $J = 4$; $\varepsilon \approx .25$; $\sigma_j^2 = 1,\ 1,\ 1,\ 1$ ($\psi^2 < \varepsilon^2$)

n_j	LW_mean	LW_median	LWW_mean	LWW_median
Equal Sample Sizes				
10, 10, 10, 10	.0553	.0765	.0544	.0750
25, 25, 25, 25	.0935	.1094	.0935	.1103
50, 50, 50, 50	.1776	.1874	.1758	.1868
100, 100, 100, 100	.4188	.4309	.4181	.4299
Unequal Sample Sizes - Positive Pairings				
5, 8, 12, 15	.0500	.0659	.0537	.0826
18, 22, 28, 32	.0893	.1024	.0916	.1068
25, 40, 60, 75	.1583	.1705	.1693	.1847
50, 80, 120, 150	.3973	.4080	.4204	.4329
Unequal Sample Sizes - Negative Pairings				
15, 12, 8, 5	.0447	.0671	.0484	.0795
32, 28, 22, 18	.0908	.1057	.0930	.1075
75, 60, 40, 25	.1633	.1740	.1736	.1887
150, 120, 80, 50	.4013	.4153	.4214	.4366

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 25.

Probability of declaring equivalence (power) for equivalence procedures; Normal distribution; J = 2; $\varepsilon \approx .50$; $\sigma_j^2 = 1,\ 1$ $(\psi^2 < \varepsilon^2)$

n_j	LW_mean	LW_median	LWW_mean	LWW_median
Equal Sample Sizes				
10, 10	.1571	.1767	.1575	.1767
25, 25	.5936	.6241	.5936	.6240
50, 50	.9369	.9447	.9368	.9446
100, 100	.9990	.9992	.9990	.9992
Unequal Sample Sizes – Positive Pairings				
5, 15	.1250	.1498	.1413	.1893
18, 32	.5739	.5977	.5880	.6161
25, 75	.8901	.8987	.9177	.9284
80, 120	.9991	.9992	.9991	.9994
Unequal Sample Sizes – Negative Pairings				
15, 5	.1322	*.1518*	.1498	.1948
32, 18	.5625	.5894	.5773	.6064
75, 25	.8880	.8967	.9216	.9330
120, 80	.9986	*.9988*	.9988	.9989

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 26.

Probability of declaring equivalence (power) for equivalence procedures; Normal distribution; J = 4; $\varepsilon \approx .50$; $\sigma_j^2 = 1,\ 1,\ 1,\ 1\ (\psi^2 < \varepsilon^2)$

n_j	LW_mean	LW_median	LWW_mean	LWW_median
		Equal Sample Sizes		
10, 10, 10, 10	.1323	.1762	.1319	.1776
25, 25, 25, 25	.3978	.4442	.3928	.4387
50, 50, 50, 50	.8174	.8362	.8098	.8314
100, 100, 100, 100	.9925	.9937	.9900	.9914
		Unequal Sample Sizes – Positive Pairings		
5, 8, 12, 15	.1035	.1501	.1075	.1729
18, 22, 28, 32	.3946	.4379	.3910	.4404
25, 40, 60, 75	.7986	.8199	.8036	.8273
50, 80, 120, 150	*.9908*	*.9919*	.9898	.9909
		Unequal Sample Sizes – Negative Pairings		
15, 12, 8, 5	.1095	.1554	.1131	*.1761*
32, 28, 22, 18	.3905	.4377	.3906	.4375
75, 60, 40, 25	.7924	.8159	.7953	.8196
150, 120, 80, 50	.9904	.9921	.9897	.9908

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 27.

Probability of declaring equivalence (power) for equivalence procedures; Normal distribution; J = 2, $\varepsilon \approx .25$; $\sigma_j^2 = 1,\ 1.3$ ($\psi^2 < \varepsilon^2$)

n_j	LW_mean	LW_median	LWW_mean	LWW_median
Equal Sample Sizes				
10, 10	.0594	.0675	.0594	.0675
25, 25	.0889	.0956	.0889	.0957
50, 50	.1537	.1556	.1537	.1556
100, 100	.3412	.3466	.3412	.3466
Unequal Sample Sizes – Positive Pairings				
5, 15	.0480	.0528	.0542	.0665
18, 32	.0865	.0897	.0893	.0921
25, 75	.1453	.1449	.1625	.1640
80, 120	.3389	.3423	.3395	.3425
Unequal Sample Sizes – Negative Pairings				
15, 5	.0489	.0636	.0571	.0828
32, 18	.0910	.0972	.0969	.1030
75, 25	*.1554*	*.1641*	.1840	.1986
120, 80	.3403	.3488	.3543	.3635

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 28.

Probability of declaring equivalence (power) for equivalence procedures; Normal distribution; $J = 4$; $\varepsilon \approx .25$; $\sigma_j^2 = 1,\ 1.1,\ 1.2,\ 1.3$ ($\psi^2 < \varepsilon^2$)

n_j	LW_mean	LW_median	LWW_mean	LWW_median
Equal Sample Sizes				
10, 10, 10, 10	.0528	.0765	.0544	.0750
25, 25, 25, 25	.0935	.1094	.0935	.1103
50, 50, 50, 50	.1776	.1874	.1758	.1868
100, 100, 100, 100	.4188	.4309	.4181	.4299
Unequal Sample Sizes – Positive Pairings				
5, 8, 12, 15	.0500	.0659	.0537	.0826
18, 22, 28, 32	.0893	.1024	.0916	.1068
25, 40, 60, 75	.1583	.1705	.1693	.1847
50, 80, 120, 150	.3973	.4080	.4204	.4329
Unequal Sample Sizes – Negative Pairings				
15, 12, 8, 5	.0447	.0671	.0484	.0795
32, 28, 22, 18	.0908	.1057	.0930	.1075
75, 60, 40, 25	.1633	.1740	.1736	.1887
150, 120, 80, 50	.4013	.4153	.4214	.4366

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 29.

Probability of declaring equivalence (power) for equivalence procedures; Normal distribution; $J = 2$; $\varepsilon \approx .50$; $\sigma_j^2 = 1,\ 2\ (\psi^2 < \varepsilon^2)$

n_j	LW_mean	LW_median	LWW_mean	LWW_median
		Equal Sample Sizes		
10, 10	.1011	.1137	.1014	.1138
25, 25	.2222	.2409	.2222	.2409
50, 50	.3607	.3835	.3604	.3831
100, 100	.5709	.5900	.5696	.5890
		Unequal Sample Sizes – Positive Pairings		
5, 15	.0823	.0846	.0875	.1018
18, 32	.2156	.2218	.2084	.2157
25, 75	.3848	.3891	.3683	.3837
80, 120	.6059	.6218	.5410	.5574
		Unequal Sample Sizes – Negative Pairings		
15, 5	.0985	*.1295*	.1159	.1710
32, 18	.2413	.2674	.2698	.3018
75, 25	.3725	*.4068*	.5643	.6296
120, 80	.5291	.5518	.6426	.6673

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 30.

Probability of declaring equivalence (power) for equivalence procedures; Normal distribution; J = 4; $\varepsilon \approx .50$; $\sigma_j^2 = 1,\ 1.33,\ 1.66,\ 2$ $(\psi^2 < \varepsilon^2)$

n_j	LW_mean	LW_median	LWW_mean	LWW_median
		Equal Sample Sizes		
10, 10, 10, 10	.0792	.1103	.0771	.1102
25, 25, 25, 25	.1644	.1929	.1644	.1912
50, 50, 50, 50	.2764	.3098	.2645	.2932
100, 100, 100, 100	.4873	.5098	.4556	.4783
		Unequal Sample Sizes – Positive Pairings		
5, 8, 12, 15	.0706	.0950	.0674	.1007
18, 22, 28, 32	.1646	.1824	.1523	.1730
25, 40, 60, 75	.2984	.3098	.2713	.2875
50, 80, 120, 150	*.5436*	*.5514*	.4917	.5053
		Unequal Sample Sizes – Negative Pairings		
15, 12, 8, 5	.0715	.1160	.0845	*.1508*
32, 28, 22, 18	.1523	.1826	.1696	.2057
75, 60, 40, 25	.2728	.3076	.3542	.4058
150, 120, 80, 50	.4141	.4484	.5723	.6103

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 31.

Probability of declaring equivalence for equivalence procedures; Normal distribution;

J = 2; $\varepsilon \approx .25$; $\sigma_j^2 = 1,\ 6\ (\psi^2 > \varepsilon^2)$

n_j	LW_mean	LW_median	LWW_mean	LWW_median
		Equal Sample Sizes		
10, 10	.0030	.0033	.0030	.0033
25, 25	.0000	.0000	.0000	.0000
50, 50	.0000	.0000	.0000	.0000
100, 100	.0000	.0000	.0000	.0000
		Unequal Sample Sizes – Positive Pairings		
5, 15	.0024	.0013	.0022	.0014
18, 32	.0000	.0000	.0000	.0000
25, 75	.0000	.0000	.0000	.0000
80, 120	.0000	.0000	.0000	.0000
		Unequal Sample Sizes – Negative Pairings		
15, 5	.0111	.0151	.0130	.0199
32, 18	.0000	.0001	.0000	.0001
75, 25	.0000	.0000	.0000	.0000
120, 80	.0000	.0000	.0000	.0000

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 32.

Probability of declaring equivalence for equivalence procedures; Normal distribution; $J = 2$; $\varepsilon \approx .50$; $\sigma_j^2 = 1,\ 6\ (\psi^2 > \varepsilon^2)$

n_j	LW_mean	LW_median	LWW_mean	LWW_median
		Equal Sample Sizes		
10, 10	.0087	.0090	.0087	.0090
25, 25	.0006	.0007	.0006	.0007
50, 50	.0002	.0003	.0002	.0003
100, 100	.0000	.0000	.0000	.0000
		Unequal Sample Sizes – Positive Pairings		
5, 15	.0064	.0065	.0064	.0075
18, 32	.0011	.0012	.0008	.0011
25, 75	.0002	.0004	.0001	.0001
80, 120	.0000	.0000	.0000	.0000
		Unequal Sample Sizes – Negative Pairings		
15, 5	.0273	.0385	.0336	.0548
32, 18	.0032	.0038	.0040	.0052
75, 25	.0017	.0022	.0076	.0117
120, 80	.0000	.0000	.0000	.0000

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 33.

Probability of declaring equivalence for equivalence procedures; Normal distribution;

$J = 4$; $\varepsilon \approx .25$; $\sigma_j^2 = 1,\ 3,\ 4,\ 6$ ($\psi^2 > \varepsilon^2$)

n_j	LW_mean	LW_median	LWW_mean	LWW_median
		Equal Sample Sizes		
10, 10, 10, 10	.0018	.0032	.0018	.0026
25, 25, 25, 25	.0000	.0000	.0000	.0000
50, 50, 50, 50	.0000	.0000	.0000	.0000
100, 100, 100, 100	.0000	.0000	.0000	.0000
		Unequal Sample Sizes – Positive Pairings		
5, 8, 12, 15	.0018	.0019	.0015	.0023
18, 22, 28, 32	.0000	.0000	.0000	.0000
25, 40, 60, 75	.0000	.0000	.0000	.0000
50, 80, 120, 150	.0000	.0000	.0000	.0000
		Unequal Sample Sizes – Negative Pairings		
15, 12, 8, 5	.0018	.0040	.0024	.0059
32, 28, 22, 18	.0001	.0000	.0001	.0000
75, 60, 40, 25	.0000	.0000	.0000	.0000
150, 120, 80, 50	.0000	.0000	.0000	.0000

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 34.

Probability of declaring equivalence for equivalence procedures; Normal distribution;

$J = 4$; $\varepsilon \approx .50$; $\sigma_j^2 = 1,\ 3,\ 4,\ 6\ (\psi^2 > \varepsilon^2)$

n_j	LW_mean	LW_median	LWW_mean	LWW_median
		Equal Sample Sizes		
10, 10, 10, 10	.0040	.0086	.0034	.0078
25, 25, 25, 25	.0005	.0006	.0004	.0006
50, 50, 50, 50	.0001	.0001	.0000	.0000
100, 100, 100, 100	.0000	.0000	.0000	.0000
		Unequal Sample Sizes – Positive Pairings		
5, 8, 12, 15	.0080	.0092	.0056	.0090
18, 22, 28, 32	.0014	.0016	.0006	.0007
25, 40, 60, 75	.0002	.0003	.0001	.0001
50, 80, 120, 150	.0000	.0000	.0000	.0000
		Unequal Sample Sizes – Negative Pairings		
15, 12, 8, 5	.0057	.0118	.0077	.0191
32, 28, 22, 18	.0003	.0005	.0004	.0006
75, 60, 40, 25	.0000	.0000	.0000	.0000
150, 120, 80, 50	.0000	.0000	.0000	.0000

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 35.

Type I error rates for equivalence procedures; χ^2 distribution (3 df); J = 2; $\varepsilon \approx .25$; $\sigma_j^2 = 1,\ 1.89\ (\psi^2 = \varepsilon^2)$

n_j	LW_mean	LW_median	LWW_mean	LWW_median
		Equal Sample Sizes		
10, 10	.0396	.0509	.0396	.0509
25, 25	.0531	.0479	.0532	.0479
50, 50	.0645	.0456	.0645	.0456
100, 100	*.0818*	.0450	*.0818*	.0450
		Unequal Sample Sizes – Positive Pairings		
5, 15	.0289	.0374	.0302	.0456
18, 32	.0440	.0471	.0455	.0485
25, 75	.0619	.0459	.0609	.0475
80, 120	*.0796*	.0469	*.0841*	.0502
		Unequal Sample Sizes – Negative Pairings		
15, 5	.0376	.0477	.0372	.0540
32, 18	.0549	.0552	.0498	.0522
75, 25	*.0801*	.0644	.0535	.0509
120, 80	*.0843*	.0496	.0724	.0428

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 36.

Type I error rates for equivalence procedures; χ^2 distribution (3 df); J = 2; $\varepsilon \approx .50$; $\sigma_j^2 = 1,\ 3.7\ (\psi^2 = \varepsilon^2)$

n_j	LW_mean	LW_median	LWW_mean	LWW_median
		Equal Sample Sizes		
10, 10	.0518	.0449	.0518	.0449
25, 25	.0622	.0387	.0622	.0387
50, 50	.0566	.0297	.0564	.0297
100, 100	.0614	.0308	.0609	.0307
		Unequal Sample Sizes – Positive Pairings		
5, 15	.0392	.0293	.0437	.0402
18, 32	.0637	.0374	.0750	.0445
25, 75	*.0859*	.0441	*.0859*	.0520
80, 120	.0738	.0380	.0719	.0406
		Unequal Sample Sizes – Negative Pairings		
15, 5	.0650	.0677	.0448	.0564
32, 18	*.0779*	.0499	.0552	.0337
75, 25	*.0934*	.0673	.0486	.0258
120, 80	.0529	.0289	.0487	*.0237*

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 37.

Type I error rates for equivalence procedures; χ^2 distribution (3 df); J = 4; $\varepsilon \approx .25$;

$\sigma_j^2 = 1,\ 1.28,\ 1.56,\ 1.84\ (\psi^2 = \varepsilon^2)$

n_j	LW_mean	LW_median	LWW_mean	LWW_median
		Equal Sample Sizes		
10, 10, 10, 10	*.0222*	.0433	*.0228*	.0431
25, 25, 25, 25	.0347	.0530	.0338	.0524
50, 50, 50, 50	.0417	.0479	.0441	.0504
100, 100, 100, 100	.0502	.0477	.0548	.0522
		Unequal Sample Sizes – Positive Pairings		
5, 8, 12, 15	*.0200*	.0380	*.0182*	.0402
18, 22, 28, 32	.0320	.0514	.0312	.0474
25, 40, 60, 75	.0431	.0476	.0358	.0431
50, 80, 120, 150	.0596	.0531	.0479	.0470
		Unequal Sample Sizes – Negative Pairings		
15, 12, 8, 5	*.0228*	.0489	*.0225*	.0560
32, 28, 22, 18	.0305	.0512	.0314	.0483
75, 60, 40, 25	.0434	.0550	.0431	.0522
150, 120, 80, 50	.0572	.0549	.0562	.0523

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 38.

Type I error rates for equivalence procedures; χ^2 distribution (3 df); J = 4; ε ≈ .50;

$\sigma_j^2 = 1,\ 1.85,\ 2.70,\ 3.55\ (\psi^2 = \varepsilon^2)$

n_j	LW_mean	LW_median	LWW_mean	LWW_median
Equal Sample Sizes				
10, 10, 10, 10	*.0246*	.0397	.0311	.0475
25, 25, 25, 25	.0427	.0426	.0605	.0580
50, 50, 50, 50	.0442	.0343	.0698	.0548
100, 100, 100, 100	.0520	.0359	*.0850*	.0609
Unequal Sample Sizes – Positive Pairings				
5, 8, 12, 15	*.0228*	.0362	*.0196*	.0356
18, 22, 28, 32	.0454	.0442	.0519	.0483
25, 40, 60, 75	.0723	.0530	.0664	.0510
50, 80, 120, 150	*.1113*	*.0802*	*.0814*	.0630
Unequal Sample Sizes – Negative Pairings				
15, 12, 8, 5	.0256	.0489	.0309	.0643
32, 28, 22, 18	.0369	.0376	.0551	.0536
75, 60, 40, 25	.0401	.0340	.0678	.0610
150, 120, 80, 50	.0330	*.0223*	.0716	.0514

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 39.

Probability of declaring equivalence (power) for equivalence procedures; χ^2 distribution (3 df); J = 2; $\varepsilon \approx .25$; $\sigma_j^2 = 1,\ 1\ (\psi^2 < \varepsilon^2)$

n_j	LW_mean	LW_median	LWW_mean	LWW_median
		Equal Sample Sizes		
10, 10	.0474	.0666	.0475	.0667
25, 25	.0806	.1115	.0807	.1117
50, 50	.1736	.2176	.1736	.2176
100, 100	*.5130*	.6039	*.5130*	.6039
		Unequal Sample Sizes - Positive Pairings		
5, 15	.0382	.0553	.0411	.0683
18, 32	.0809	.0992	.0834	.1038
25, 75	.1530	.1910	.1750	.2211
80, 120	*.5031*	.6008	*.5116*	.6100
		Unequal Sample Sizes - Negative Pairings		
15, 5	.0415	.0535	.0454	.0643
32, 18	.0809	.1034	.0856	.1061
75, 25	*.1606*	.1988	.1808	.2249
120, 80	*.5026*	.5978	.5100	.6077

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 40.

Probability of declaring equivalence (power) for equivalence procedures; χ^2 distribution (3 df); J = 4; $\varepsilon \approx .25$; $\sigma_j^2 = 1,\ 1,\ 1,\ 1$ ($\psi^2 < \varepsilon^2$)

n_j	LW_mean	LW_median	LWW_mean	LWW_median
Equal Sample Sizes				
10, 10, 10, 10	*.0257*	.0591	*.0246*	.0547
25, 25, 25, 25	.0501	.0964	.0470	.0901
50, 50, 50, 50	.1069	.1848	.1045	.1791
100, 100, 100, 100	.2644	.4170	.2610	.4085
Unequal Sample Sizes - Positive Pairings				
5, 8, 12, 15	*.0217*	.0548	*.0232*	.0599
18, 22, 28, 32	.0479	.0916	.0455	.0911
25, 40, 60, 75	.0910	.1640	.0947	.1695
50, 80, 120, 150	.2485	.4037	.2587	.4175
Unequal Sample Sizes - Negative Pairings				
15, 12, 8, 5	*.0248*	.0556	*.0230*	.0602
32, 28, 22, 18	.0479	.0953	.0465	.0947
75, 60, 40, 25	.0947	.1702	.1007	.1776
150, 120, 80, 50	.2486	.3955	.2538	.4105

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 41.

Probability of declaring equivalence (power) for equivalence procedures; χ^2 distribution (3 df); J = 2; $\varepsilon \approx .50$; $\sigma_j^2 = 1,\ 1\ (\psi^2 < \varepsilon^2)$

n_j	LW_mean	LW_median	LWW_mean	LWW_median
		Equal Sample Sizes		
10, 10	.1259	.1660	.1259	.1662
25, 25	.4853	.6009	.4853	.6009
50, 50	.8657	.9377	.8655	.9375
100, 100	.9911	.9993	.9910	.9993
		Unequal Sample Sizes – Positive Pairings		
5, 15	.1009	.1359	.1096	.1691
18, 32	.4774	.5848	.4865	.5977
25, 75	*.7965*	.8914	*.8336*	.9103
80, 120	.9908	.9988	.9913	.9987
		Unequal Sample Sizes – Negative Pairings		
15, 5	.1057	.1388	.1152	.1687
32, 18	*.4696*	.5733	.4809	.5874
75, 25	*.8081*	.8983	.8431	.9169
120, 80	.9926	.9990	.9921	*.9990*

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 42.

Probability of declaring equivalence (power) for equivalence procedures; χ^2 distribution (3 df); J = 4; $\varepsilon \approx .50$; $\sigma_j^2 = 1,\ 1,\ 1,\ 1\ (\psi^2 < \varepsilon^2)$

n_j	LW_mean	LW_median	LWW_mean	LWW_median
Equal Sample Sizes				
10, 10, 10, 10	*.0608*	.1334	.0599	.1277
25, 25, 25, 25	.2470	.4308	.2346	.4057
50, 50, 50, 50	.6364	.8264	.6200	.8039
100, 100, 100, 100	.9584	.9949	*.9517*	.9909
Unequal Sample Sizes – Positive Pairings				
5, 8, 12, 15	*.0529*	.1205	*.0524*	.1288
18, 22, 28, 32	.2457	.4182	.2334	.3944
25, 40, 60, 75	.6094	.8097	.6024	.7895
50, 80, 120, 150	*.9495*	*.9928*	*.9432*	.9880
Unequal Sample Sizes – Negative Pairings				
15, 12, 8, 5	.0543	.1243	.0530	.1267
32, 28, 22, 18	.2433	.4026	.2345	.3905
75, 60, 40, 25	.6195	.8137	.6098	.7890
150, 120, 80, 50	.9533	*.9937*	.9431	.9865

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 43.

Probability of declaring equivalence (power) for equivalence procedures; χ^2 distribution (3 df); J = 2; $\varepsilon \approx .25$; $\sigma_j^2 = 1,\ 1.3\ (\psi^2 < \varepsilon^2)$

n_j	LW_mean	LW_median	LWW_mean	LWW_median
		Equal Sample Sizes		
10, 10	.0476	.0648	.0476	.0648
25, 25	.0749	.0984	.0750	.0984
50, 50	.1540	.1825	.1540	.1825
100, 100	*.3654*	.4353	*.3654*	.4353
		Unequal Sample Sizes – Positive Pairings		
5, 15	.0374	.0495	.0407	.0605
18, 32	.0716	.0881	.0741	.0913
25, 75	.1305	.1560	.1455	.1805
80, 120	*.3691*	.4341	*.3705*	.4383
		Unequal Sample Sizes – Negative Pairings		
15, 5	.0396	.0561	.0432	.0677
32, 18	.0757	.0925	.0773	.0977
75, 25	*.1403*	.1722	.1584	.1997
120, 80	*.3686*	.4371	.3795	.4517

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 44.

Probability of declaring equivalence (power) for equivalence procedures; χ^2 distribution (3 df); J = 4; $\varepsilon \approx .25$; $\sigma_j^2 = 1,\ 1.1,\ 1.2,\ 1.3$ $(\psi^2 < \varepsilon^2)$

n_j	LW_mean	LW_median	LWW_mean	LWW_median
		Equal Sample Sizes		
10, 10, 10, 10	*.0254*	.0632	*.0266*	.0598
25, 25, 25, 25	.0477	.0942	.0473	.0899
50, 50, 50, 50	.0838	.1526	.0836	.1456
100, 100, 100, 100	.1910	.3132	.1862	.3080
		Unequal Sample Sizes – Positive Pairings		
5, 8, 12, 15	*.0218*	.0522	*.0214*	.0514
18, 22, 28, 32	.0445	.0855	.0434	.0842
25, 40, 60, 75	.0743	.1388	.0762	.1411
50, 80, 120, 150	.1906	.3038	.1932	.3003
		Unequal Sample Sizes – Negative Pairings		
15, 12, 8, 5	*.0225*	.0584	*.0213*	.0617
32, 28, 22, 18	.0434	.0845	.0435	.0861
75, 60, 40, 25	.0822	.1525	.0862	.1637
150, 120, 80, 50	.1916	.3111	.2070	.3401

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 45.

Probability of declaring equivalence (power) for equivalence procedures; χ^2 *distribution (3 df); J = 2;* $\varepsilon \approx .50$; $\sigma_j^2 = 1,\ 2\ (\psi^2 < \varepsilon^2)$

n_j	LW_mean	LW_median	LWW_mean	LWW_median
		Equal Sample Sizes		
10, 10	.0961	.1232	.0963	.1233
25, 25	.2827	.3431	.2827	.3431
50, 50	.4688	.5840	.4683	.5831
100, 100	.6830	.8462	.6819	.8445
		Unequal Sample Sizes – Positive Pairings		
5, 15	.0776	.0945	.0834	.1176
18, 32	.2762	.3349	.2708	.3281
25, 75	*.4590*	.5753	*.4318*	.5267
80, 120	.7064	.8690	.6433	.8069
		Unequal Sample Sizes – Negative Pairings		
15, 5	.0980	.1357	.1065	.1626
32, 18	*.2948*	.3631	.3136	.3998
75, 25	*.4524*	.5513	.6150	.7662
120, 80	.6416	.8096	.7321	*.8898*

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 46.

Probability of declaring equivalence (power) for equivalence procedures; χ^2 distribution (3 df); $J = 4$; $\varepsilon \approx .50$; $\sigma_j^2 = 1,\ 1.33,\ 1.66,\ 2$ ($\psi^2 < \varepsilon^2$)

n_j	LW_mean	LW_median	LWW_mean	LWW_median
		Equal Sample Sizes		
10, 10, 10, 10	*.0467*	.1082	.0458	.1017
25, 25, 25, 25	.1388	.2448	.1350	.2341
50, 50, 50, 50	.3100	.4887	.2975	.4657
100, 100, 100, 100	.5558	.7512	*.5185*	.7397
		Unequal Sample Sizes – Positive Pairings		
5, 8, 12, 15	*.0402*	.0884	*.0361*	.0863
18, 22, 28, 32	.1444	.2545	.1329	.2252
25, 40, 60, 75	.3378	.5150	.2889	.4361
50, 80, 120, 150	*.6027*	*.8190*	*.5201*	.7116
		Unequal Sample Sizes – Negative Pairings		
15, 12, 8, 5	.0493	.1103	.0503	.1268
32, 28, 22, 18	.1385	.2485	.1432	.2649
75, 60, 40, 25	.2973	.4634	.3616	.5708
150, 120, 80, 50	.5018	*.7195*	.6407	.8414

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 47.

Probability of declaring equivalence for equivalence procedures; χ^2 distribution (3 df);

$J = 2$; $\varepsilon \approx .25$; $\sigma_j^2 = 1,\ 6\ (\psi^2 > \varepsilon^2)$

n_j	LW_mean	LW_median	LWW_mean	LWW_median
Equal Sample Sizes				
10, 10	.0116	.0115	.0117	.0115
25, 25	.0015	.0015	.0015	.0015
50, 50	.0000	.0000	.0000	.0000
100, 100	.0000	.0000	.0000	.0000
Unequal Sample Sizes – Positive Pairings				
5, 15	.0067	.0070	.0071	.0088
18, 32	.0015	.0010	.0015	.0010
25, 75	.0000	.0000	.0000	.0000
80, 120	.0000	.0000	.0000	.0000
Unequal Sample Sizes – Negative Pairings				
15, 5	.0190	.0243	.0209	.0311
32, 18	.0032	.0020	.0034	.0021
75, 25	.0009	.0007	.0010	.0008
120, 80	.0000	.0000	.0000	.0000

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 48.

Probability of declaring equivalence for equivalence procedures; χ^2 distribution (3 df);

$J = 4$; $\varepsilon \approx .25$; $\sigma_j^2 = 1, 3, 4, 6$ ($\psi^2 > \varepsilon^2$)

n_j	LW_mean	LW_median	LWW_mean	LWW_median
		Equal Sample Sizes		
10, 10, 10, 10	.0055	.0100	.0054	.0089
25, 25, 25, 25	.0002	.0005	.0002	.0003
50, 50, 50, 50	.0000	.0000	.0000	.0000
100, 100, 100, 100	.0000	.0000	.0000	.0000
		Unequal Sample Sizes – Positive Pairings		
5, 8, 12, 15	.0053	.0105	.0054	.0116
18, 22, 28, 32	.0008	.0010	.0006	.0008
25, 40, 60, 75	.0003	.0002	.0002	.0002
50, 80, 120, 150	.0000	.0000	.0000	.0000
		Unequal Sample Sizes – Negative Pairings		
15, 12, 8, 5	.0054	.0122	.0048	.0150
32, 28, 22, 18	.0007	.0009	.0007	.0007
75, 60, 40, 25	.0000	.0000	.0000	.0000
150, 120, 80, 50	.0000	.0000	.0000	.0000

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 49.

Probability of declaring equivalence for equivalence procedures; χ^2 distribution (3 df); $J = 2$; $\varepsilon \approx .50$; $\sigma_j^2 = 1,\ 6\ (\psi^2 > \varepsilon^2)$

n_j	LW_mean	LW_median	LWW_mean	LWW_median
Equal Sample Sizes				
10, 10	.0261	.0294	.0261	.0294
25, 25	.0145	.0146	.0145	.0146
50, 50	.0048	.0082	.0048	.0081
100, 100	.0017	.0038	.0017	.0038
Unequal Sample Sizes – Positive Pairings				
5, 15	.0203	.0239	.0228	.0297
18, 32	.0135	.0152	.0135	.0149
25, 75	.0101	.0188	.0087	.0136
80, 120	.0021	.0108	.0011	.0047
Unequal Sample Sizes – Negative Pairings				
15, 5	.0452	.0638	.0507	.0799
32, 18	.0194	.0236	.0220	.0280
75, 25	.0124	.0201	.0405	.0912
120, 80	.0010	.0037	.0036	.0158

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 50.

Probability of declaring equivalence for equivalence procedures; χ^2 *distribution (3 df);*

$J = 4$; $\varepsilon \approx .50$; $\sigma_j^2 = 1, 3, 4, 6$ $(\psi^2 > \varepsilon^2)$

n_j	LW_mean	LW_median	LWW_mean	LWW_median
		Equal Sample Sizes		
10, 10, 10, 10	.0124	.0259	.0117	.0222
25, 25, 25, 25	.0063	.0124	.0054	.0094
50, 50, 50, 50	.0013	.0047	.0012	.0027
100, 100, 100, 100	.0003	.0026	.0003	.0006
		Unequal Sample Sizes – Positive Pairings		
5, 8, 12, 15	.0117	.0243	.0107	.0227
18, 22, 28, 32	.0072	.0145	.0048	.0099
25, 40, 60, 75	.0099	.0261	.0054	.0133
50, 80, 120, 150	.0059	.0281	.0018	.0068
		Unequal Sample Sizes – Negative Pairings		
15, 12, 8, 5	.0138	.0281	.0149	.0394
32, 28, 22, 18	.0047	.0102	.0051	.0103
75, 60, 40, 25	.0015	.0037	.0024	.0062
150, 120, 80, 50	.0002	.0010	.0003	.0008

Note: LW_mean = Levene-Wellek test based on the mean; LW_median = Levene-Wellek test based on the median; LWW_mean = Levene-Wellek-Welch test based on the mean; LWW_median = Levene-Wellek-Welch test based on the median

Table 51.

Type I error rates summary: Minimum and maximum empirical Type I error rates and number of times the Type I error rates exceeded the bounds of .025 - .075 for the equivalence procedures over the 96 null conditions.

Test	Minimum Empirical Type I Error Rate	Maximum Empirical Type I Error Rate	Number of Times Type I Error Rate Exceeded the Bounds of .025-.075
Levene-Wellek mean	.0200	.1113	12
Levene-Wellek median	.0223	.0886	6
Levene-Wellek-Welch mean	.0182	.0859	9
Levene-Wellek-Welch median	.0237	.1014	2

Table 52.

Power summary: Proportion of conditions (out of 192 conditions) in which a specific equivalence procedure had the highest power, including ties (i.e., conditions where the null hypothesis was false).

Test	Proportion Test had Highest Power in Equal Sample Size Conditions (out of 64)	Proportion Test had Highest Power in Positive Pairing Conditions (out of 64)	Proportion Test had Highest Power in Positive Pairing Conditions (out of 64)	Proportion of Ties (out of the 192 conditions)*
Levene-Wellek mean	0%	0%	0%	0%
Levene-Wellek median	85.9%	34.4%	10.9%	9.4%
Levene-Wellek-Welch mean	0%	0%	0%	0%
Levene-Wellek-Welch median	42.2%	65.6%	90.6%	9.4%

* 18 ties total out of the 192 power conditions, and 17 of those ties were in an equal sample size condition.

Table 53.

Descriptive statistics for the three groups on The Perfectionism Cognitions Inventory at pre-test for the applied example, N = 83.

Group	Mean	Median	Variance	N
CBT	66.14	65	241.79	30
Stress	68.83	68	110.79	29
No Treatment	69.75	74	156.28	24

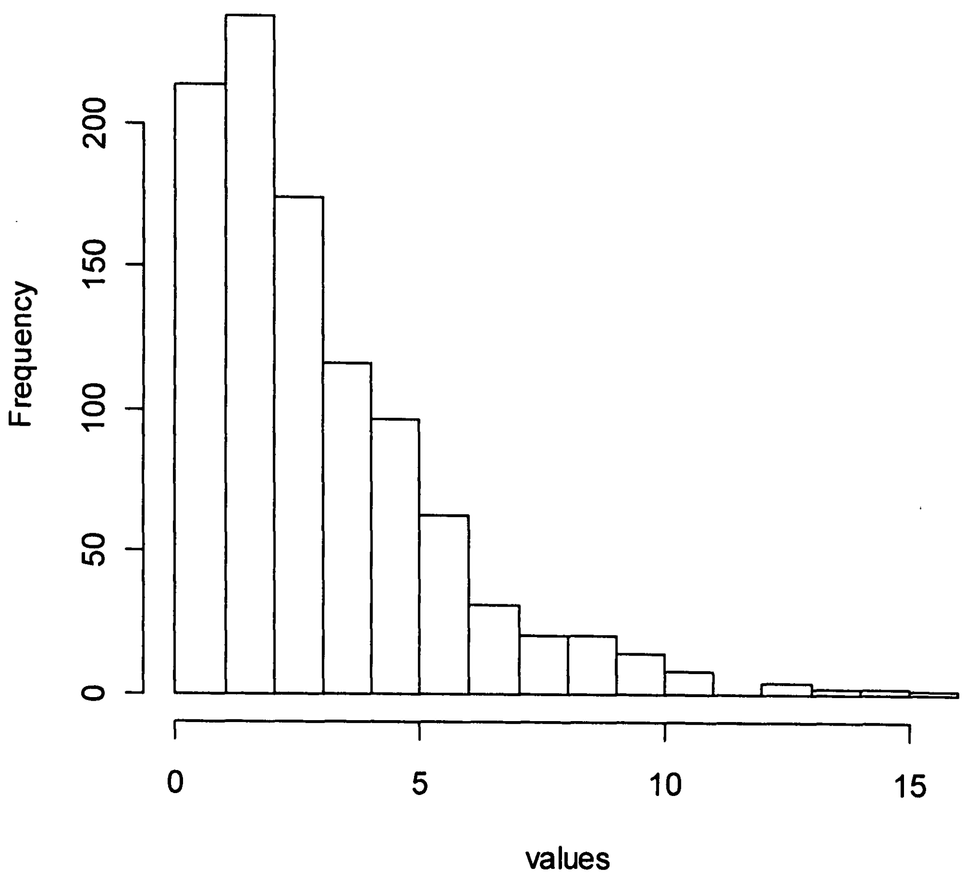

Figure 1. Histogram illustrating the shape of a chi-square distribution with 3 degrees of freedom.

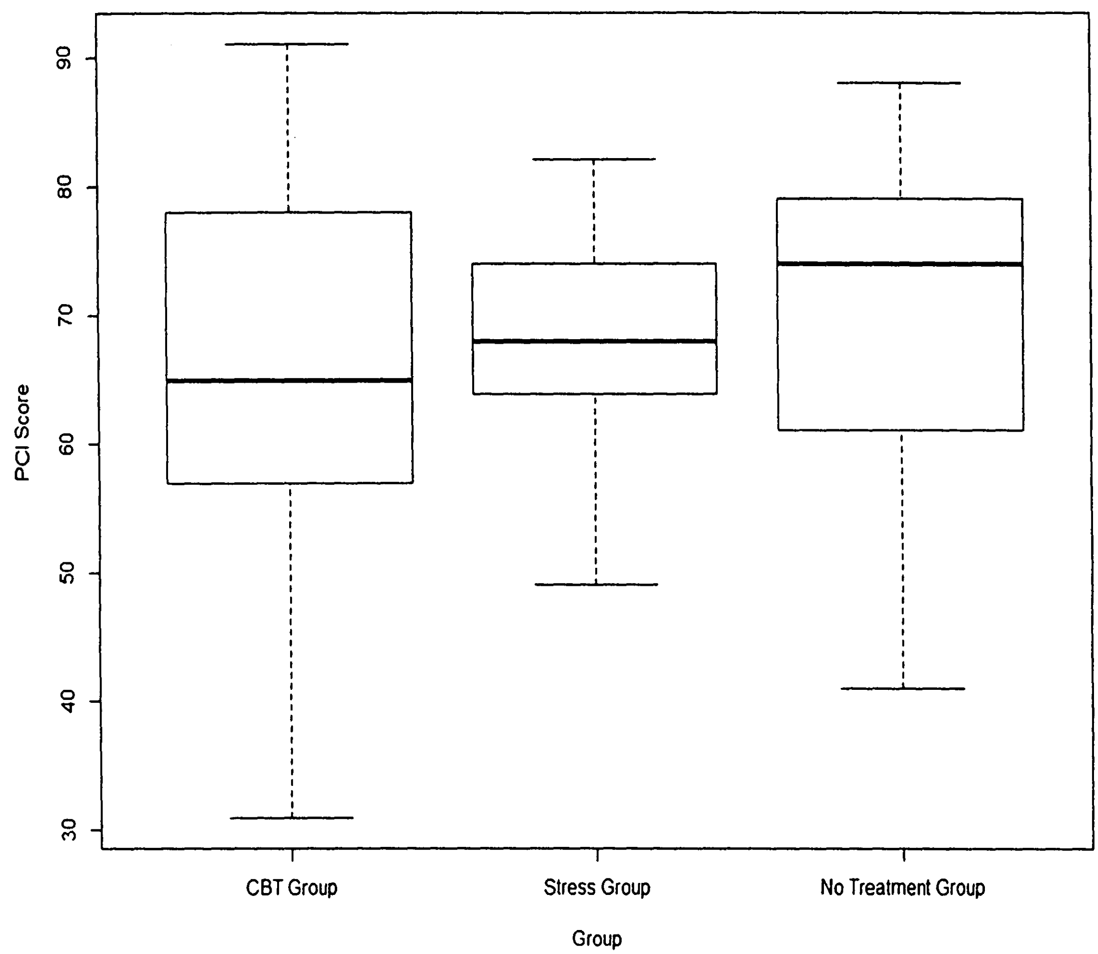

Figure 2. Boxplot of the 3 groups in the applied example.

Appendix A

```
Equiv_Vars<-function(x, group, eps, alpha=.05, na.rm=TRUE, ...) {

dv<-x
gr <- as.factor(group)
medians <- tapply(dv, gr, median)
n <- tapply(dv, gr, length)
resp.median <- abs(dv - medians[gr])
ngroup<-length(group)
alpha<-.05
eps<-eps
vars<-(tapply(dv, gr, var))

## Equivalence test for Equivalence of Variances ##

LWW_md<-oneway.test(resp.median~gr)$statistic*((ngroup-1)/(mean(n)))

crit_LWW_md<-((ngroup-1)/((mean(n))*qf(p=alpha, df1=ngroup-1,
      df2=oneway.test(resp.median~gr)$parameter[2],
      ncp=(mean(n))*eps^2)))

ifelse (LWW_md <= crit_LWW_md, decis_equiv<-"The null hypothesis that
      the differences vetween the group variances falls outside the
      equivalence interval can be rejected.", decis_equiv<-"The null
      hypothesis that the differences between the group variances falls
      outside of the equivalence interval cannot be rejected")

## Summary ##
title1<-"Variances of the Groups"
title2<-"Equivalence Based Equality of Variances Test"
stats_equiv<-c(eps,LWW_md,crit_LWW_md,decis_equiv)
names(stats_equiv)<-c("Equivalence Interval", "Equivalence Test
      Statistic", "critical value", "Decision")

out<-list (title1, vars, title2, stats_equiv)
out

}
```