Equivalence of Population Variances:

Synchronizing the Objective and Analysis

Constance Mara

James M. Anderson Center for Health Systems Excellence

Cincinnati Children's Hospital Medical Center

Robert A. Cribbie

Quantitative Methods Program, Department of Psychology

York University

Abstract

Researchers are often interested in testing for the equivalence of population variances. Traditional difference-based procedures are appropriate to answer questions about differences in some statistic (e.g., variances, etc.). However, if a researcher is interested in evaluating the equivalence of population 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 difference-based variance homogeneity tests under common data conditions. Results demonstrated that traditional difference-based tests assess equality of variances from the wrong perspective, and that the proposed Levene-Wellek-Welch test for equivalence of group variances using the absolute deviations from the median was the best performing test for detecting equivalence. An *R* function is provided in order to facilitate use of this test for equivalence of population variances.

keywords: equivalence testing, homogeneity of variances, ANOVA, dispersion

Equivalence of Population Variances:

Synchronizing the Objective and Analysis

Homogeneity of variances occurs when population distributions have similar dispersion. Researchers are becoming increasingly interested in the properties of their data aside from central tendency, such as dispersion. For instance, Borkenau, Hrebícková, Kuppens, Realo, and Allik (2013) hypothesized that self-reported personality scores would have similar variability across males and females. 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. A more well-known reason for exploring variances is to verify the homogeneity of variances assumption related to traditional parametric tests of mean differences. Regardless of the reason, researchers need a valid test for assessing questions related to variability.

There has been substantial research on different tests that can be used to test for differences in variances. This paper discusses whether traditional tests of variance homogeneity address the problem of variance equality from the wrong perspective. We argue that to test for variance homogeneity, one should use equivalence tests because the research hypothesis of variance equality is properly aligned with the alternate hypothesis, not the null hypothesis. To that end, we first 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 equivalence. Then, the main goal of this paper is to compare our newly developed tests

for equivalence of group variances to currently recommended variance homogeneity tests under data conditions common in educational and psychological research. A review of traditional variance homogeneity tests as well as equivalence testing is outlined before developing the new 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. In this case, the researcher would like to find that the variances are equal across groups. It is important to note that it is not necessary to use a preliminary test of variance homoscedasticity in order to justify the use of heteroscedastic procedures (e.g., Welch's heteroscedastic ANOVA instead of the traditional ANOVA) because these tests are generally effective regardless of whether variances are equal or unequal across groups. Many researchers have suggested abandoning non-robust parametric procedures completely in favor of robust procedures that do not require the homogeneity of variances assumption (e.g., Wilcox, Charlin, & Thompson, 1986; Zimmerman, 2004). However, researchers in the educational and behavioral sciences still widely use traditional parametric procedures and need to screen for the assumptions associated with these tests.

A more interesting reason for assessing equivalence of variances is that the primary research question is concerned with whether the dispersion of the dependent variable is similar across multiple groups. 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 among 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 for assessing equivalence of variances is quite relevant to the kinds of questions educational researchers (and researchers in related disciplines) are interested in and necessary if a researcher wants to justify the use of a traditional mean difference test. However, as we argue in this paper, 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 $z_{ij} = \left| X_{ij} - M_j \right|$, where X_{ij} is the score of the *i*th individual in the *j*th group and M_j is the mean of 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}. Previous simulation 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. 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. However, the Levene test based on the median has often been recommended because it performs well across a wide range of conditions.

Traditional Variance Homogeneity Procedures Evaluated in the Current Study. The current 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.

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. Instead of using the *j*th sample mean in the sample score transformation, this

modification uses the transformation, $\tilde{z}_{ij} = \left| X_{ij} - MDN_j \right|$, where MDN_j is the *j*th sample median. The transformed scores are analyzed using an ANOVA *F*- test.

Levene's original test with a Welch adjustment ("LevWelch_mean"). Welch's (1951) heteroscedastic 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 also has relevance to Levene's test for homogeneity of variances (and its modifications), given that Levene's test uses the ANOVA *F*-test and thus also assumes homogeneity of variances (more specifically, homogeneity of 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, Charlin, & Thompson, 1986).

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 $w_{z_j} = \frac{n_j}{s^2_{z_j}}$, n_j is the size of the *j*th group, $s^2_{z_j}$ is the variance of the transformed scores for the *j*th group, $\overline{Z}'_{..} = \frac{\sum w_{z_j} \overline{Z}_j}{\sum w_{z_j}}$, and $\overline{Z}_j$ is the mean of the z_{ij} for the *j*th group.

The *F*` 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's median-based test with Welch adjustment ("LevWelch_mdn"). This procedure uses the absolute deviations from the median, $\tilde{z}_{ij} = \left| X_{ij} - MDN_j \right|$, to conduct the Welch ANOVA *F*` to assess homogeneity of variances (outlined previously). In this case $\overline{Z}_j$ 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.

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 fail to reject a null hypothesis regarding the exact equality of group variances. Specifically, if one is using a traditional test for homogeneity of variances, the goal is to fail to reject the null hypothesis for these tests, $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_J^2$ (where *J* = number of groups). In other words, the research

hypothesis that the variances are equal is aligned with the null hypothesis rather than the alternate hypothesis. 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 a true 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 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 the group variances, it is likely that a difference-based variance homogeneity test will reject the null hypothesis and declare the population variances unequal. 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 population 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 population variances. More generally, traditional difference-based procedures are less

likely to detect equality of variances as sample size increases, which is 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, that no relationship exists between two variables, or that variables do not interact (e.g., Cribbie, Gruman, & Arpin-Cribbie, 2004; Cribbie, Ragoonanan & Counsell, in press; Goertzen & Cribbie, 2010; Robinson, Duursma, & Marshall, 2005; Rogers, Howard, & Vessey, 1993, Schuirmann, 1987; Wellek, 2010), or that the variances of two or more populations are equal (as proposed in the current study). In equivalence testing, a lack of association implies that the difference between two statistics is so small that it can be considered inconsequential or meaningless. This difference is defined *a priori* as the equivalence interval (e.g., *-δ, δ;* discussed in more detail later). 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., $-\delta \leq \sigma_1^2 - \sigma_2^2 \leq \delta$).

Novel Equivalence-Based Homogeneity of Variance Tests Evaluated in the Current Study

Given the fundamental problems with traditional tests for homogeneity of variances, we developed an equivalence-based test for homogeneity of variances along with several modifications. Previously, Wellek (2010) developed an approach for assessing the equivalence of variances for two groups that utilizes the ratio of the largest to smallest variance, which, as suggested by an anonymous reviewer, is similar in nature

to the F_{max} test developed by Hartley (1950). The current research explores an alternative approach that uses the raw difference rather than the ratio and is suitable for two or more independent groups. The null hypothesis for the one-way equivalence test for homogeneity of variances is that the difference among the variances of the groups is equal to or larger than an a priori determined cutoff (ε^2):

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

where Ψ^{*2} quantifies the difference among the variances of the groups and ε^2 represents the smallest difference in the variances that is considered meaningful. Note that the equivalence interval includes any value less than ε^2, and contains a lower bound of zero since we are working in squared units (which differs from the equivalence interval of many other equivalence tests which are symmetric around zero). More discussion regarding ε^2 is provided below.

Levene-Wellek test for equivalence of variances ("LW_mean"). This procedure is based on Wellek's (2010) original one-way equivalence test statistic (which simultaneously evaluates the equivalence of all *J* population means), substituting Levene's original transformation in place of the raw scores. This new hybrid test statistic can be presented as:

$$LW = \frac{\sum_{j=1}^{J} n_j \left(\overline{Z}_j - \overline{Z}_{..}^{*}\right)^2 \Big/ J-1}{\sum_{i=1}^{n}\sum_{j=1}^{J} (z_{ij} - \overline{Z}_j)^2 \Big/ (N-J)}$$

with Levene's original transformation, $z_{ij} = \left| X_{ij} - M_j \right|$, so that $\overline{Z}_j = \frac{\sum z_{ij}}{n_j}$, and

$\overline{Z}^*_{..} = \frac{\left(\sum_{j=1}^{J} \sum_{i=1}^{n_j} z_{ij} \right)}{\sum_{j=1}^{J} n_j}$. $H_0 : \Psi^{*2} \geq \varepsilon^2$ is rejected if LW< $F_{\alpha, J-1, N-J, \bar{n}\varepsilon^2}$, where $\bar{n}\varepsilon^2$ represents the noncentrality parameter, which is computed by multiplying the average group size by the squared upper bound of the equivalence interval. It is important to note that for simplicity we have framed the LW as a noncentral *F* statistic as opposed to the traditional formulation in the metric of ψ^2 (Wellek, 2010).

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).

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 (transformed score) variances. The new equivalence-based robust test statistic can be presented as:

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

As with the LW test, the test statistic is approximately distributed as noncentral *F* with the noncentrality parameter $\overline{n}\varepsilon^2$, *J*-1 numerator degrees of freedom and denominator degrees of freedom of:

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

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 (2010) 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, 2010, 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 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 ("power" to detect 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 using a normal population distribution as well as a positively skewed population distribution (χ^2 with 3 degrees of freedom). In order to evaluate the Type I error rates of the equivalence-based procedures, the liberal bounds of $\alpha \pm 0.5\alpha$ (Bradley, 1978) were used. Therefore, with an α 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. It is important to note that if an inaccurate Type I error rate is obtained for a specific condition, the corresponding power rate for that condition should also be interpreted with caution since it could be artificially inflated or deflated as a result of the imprecise Type I error control. The simulations were conducted with the open-source statistical software *R* (R Development Core Team, 2016).

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 alternate hypothesis 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. See Figure 1 for further clarification.

We looked at $J = 4$ groups for all conditions. Several variables were manipulated in this study including sample size ($\bar{n}$ = 10, 25, 50, 100), balanced versus unbalanced designs (i.e., equal versus unequal group sizes), equality/inequality of population variances, and pairings of unequal sample sizes with unequal population variances. The summary of the conditions tested in this study can be found in Table 1. For the equivalence-based tests, the conservative and liberal equivalence limits of ε = .25 and ε = .50, respectively, were used (Wellek, 2010). However, the pattern of results for both equivalence limits were similar, so only the results for ε = .50 are presented. As expected, the power rates for ε = .25 were lower across all conditions.

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 σ_j 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 three 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 the mean (mean = df = 3) and dividing by the standard deviation (sd = $\sqrt{2df} = \sqrt{(2)(3)} \approx 2.45$) of the χ^2

distribution. The resulting values were then multiplied by σ_j to produce a distribution of observations with the variances outlined in Table 1.

Unbalanced designs (i.e., unequal sample sizes across the groups) that are paired with unequal variances can severely 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 (directly) and negative (inverse) 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. The sample size pairings can be found 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. 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.

Results

As noted previously, due to similar patterns of results, only results for ε = .50 are presented[1].

Equivalence-Based Procedures

Empirical Type I error rates.

A summary of the Type I error rate results can be found in Table 2; recall that these are the Type I error rates that were obtained when $\Psi^{*2} = \varepsilon^2$. It is hoped that this table will help provide an easily interpretable summary of the Type I error control of the procedures. More specific observations are also discussed below.

Normal Distributions. Type I error rates in the equal sample size conditions were maintained close to the nominal level, ranging from .0381 to .0702. For the positive pairing conditions, the Levene-Wellek-Welch procedures (LWW_mean, LWW_median) had acceptable Type I error rates for all sample sizes. However, both of the Levene-Wellek procedures (i.e. LW_mean, LW_median) had overly liberal Type I error rates at the highest sample size (.084 and .0869). For the negative pairing conditions, LWW_median had a very liberal Type I error rate at the smallest sample size condition (.1014 at $\bar{n}$ = 10). However, at the larger sample sizes in the negative pairing conditions, the Type I error rates were acceptable for the LWW_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.

Positively Skewed Distributions (χ^2, 3 df). All of the equivalence procedures maintained accurate Type I error rates when variances were negatively paired with

[1] Tables/Figures of the full results of all conditions evaluated in this study can be obtained by emailing the corresponding author.

unequal sample sizes. However, when variances were positively paired with the largest unequal sample size, $\bar{n}$ = 100, only the LWW_median had an accurate Type I error rate. Additionally, the LWW_mean had a Type I error rate that was too conservative (.0196) when $\bar{n}$ = 10 and sample sizes were positively paired with the variances, and a Type I error rate too liberal (.0814) in the largest equal sample size condition.

Power

A summary of the power results for the equivalence procedures can be found in Table 3. When variances were exactly equal, the difference in the variances (i.e., zero difference) fell within the equivalence interval, so this was a power condition for the equivalence-based procedures. Additionally, for a 2:1 variance ratio this difference in the variances was within the equivalence interval for the equivalence procedures, such that $\psi^{*2} < \varepsilon^2$; therefore, this condition was another test of the power of these procedures.

Normal Distributions. 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.

For a 2:1 variance ratio in the largest sample size condition, power was approximately 41% to 61%. All four equivalence procedures had comparable power rates across all sample size and variance combinations under normal distribution conditions.

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

For a 2:1 variance ratio, when sample sizes were equal, the median-based procedures had the highest power at all sample sizes. This power advantage for the

median-based tests was also observed when unequal sample sizes were positively paired with variances and when unequal sample sizes were negatively paired with variances. See Figure 2.

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.

Normal Distributions. As expected, the probability of declaring equivalence was low at small sample sizes and was zero in the larger sample size conditions.

Positively Skewed Distributions (χ^2, 3 df). The error rates were almost zero when sample sizes were equal, or unequal sample sizes were positively paired with the variances. The error rates were slightly higher for the negative pairing conditions, although they remained close to the empirical Type I error rates.In the largest sample sizes conditions, the error rates across all conditions were at or nearly zero (see Figure 3).

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., failing to reject the null hypothesis) in this

condition should have been approximately 1 – α (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, as expected, sometimes too conservative or too liberal.

For the 2:1 variance ratio condition, the difference-based tests had a very high probability of declaring equivalence at $\bar{n}$ = 10 (note that this is an incorrect decision, i.e., a Type II error). In the largest sample size conditions ($\bar{n}$ = 100), the probability of declaring equivalence was much lower in the equal sample size conditions. It is important to note that the 2:1 variance ratio in this condition meant the 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 significant most of the time in the largest sample size conditions.

For a 6:1 variance ratio 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 – α (.95) for the difference-based procedures. 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.

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. See Figure 2.

For a 6:1 variance ratio, many false declarations of equivalence were observed for the difference-based procedures in the smaller sample size conditions (see Figure 3). However, in the largest sample size conditions, the rate was approximately zero.

Discussion

Results of the simulation study demonstrated the backward nature of the traditional difference-based procedures for assessing equality of population 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 traditional 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 nonsignificant 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 population variances). This study developed four procedures, combining existing procedures for variance equality and equivalence testing logic.

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.

In order to facilitate use of our newly developed procedure, a function for the Levene-Wellek-Welch procedure based on the absolute deviations from the median was developed in *R* (*R* Project, 2015) and is available at http://cribbie.info.yorku.ca

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 educational and psychological research. However, research into the performance of the proposed equivalence-based tests of homogeneity of variance across a wider range of conditions is definitely recommended.

A broader limitation of equivalence testing procedures in general involves the decision around appropriate equivalence intervals. 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. While this could be construed as a limitation, we challenge researchers to think carefully about meaningless differences among their groups when selecting equivalence intervals rather than relying on rules of thumb or generic guidelines.

Applied Example

This section presents a demonstration of how to use the Levene-Wellek Welch test, the best-performing equivalence-based homogeneity of variance test in terms of

power and Type I error control, using a substantive example from psychological research. We also contrast the results of this test with that of the original Levene median-based test using the same data. This example achieves two goals: 1) demonstrate the use of the new equivalence-based homogeneity of variance procedure; and 2) highlight the fundamental flaws of the original Levene-type difference-based tests for homogeneity of variances.

Data were taken from Arpin-Cribbie, Irvine, and Ritvo (2011). Participants scoring very high on maladaptive perfectionism 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 pretest 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 pretest measures, but the current example just tests for the equivalence of variances on the baseline measure of the Perfectionism Cognitions Inventory (PCI; Flett, Hewitt, Blankstein, & Gray, 1998) for the purpose of demonstration. The variances for the stress management group ($s^2 = 110.79$) and the no treatment group ($s^2 = 156.28$) were similar, but the CBT group variance ($s^2 = 241.79$) was more than two times larger than the Stress Management group.

The original Levene test indicated that there were no statistically significant differences among the group variances when using the popular $\alpha = .05$, $F = 2.50$, $p = .09$. 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_{Med} = 2.10, p = .13. Next, the newly developed median-based Levene-Wellek-Welch equivalence test was used, setting an equivalence interval to ε = .50. This equivalence test found that the variances were not significantly equivalent (LWW_{med} = 6.42 > $F_{\alpha, J-1, N-J, \bar{n}\varepsilon^2}$ = 4.10). 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 primary 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 by the traditional tests 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. 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.

Future Directions

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 the educational and behavioral sciences as in other disciplines. For example, Sagrestano, Heavey, and Christensen (1998) argue that different perspectives 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.

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 support their research hypotheses by failing to reject the null hypothesis. 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 research hypotheses of equivalent population 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.

Borkenau, P., Hrebícková, M., Kuppens, P., Realo, A., & Allik, J. (2013). Sex differences in variability in personality: A study in four samples. *Journal of Personality, 81,* 49-60.

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.

Flett, G. L., Hewitt, P. L., Blankstein, K. R., & Gray, L. (1998). Psychological distress and the frequency of perfectionistic thinking. Journal of Personality and Social Psychology, 75, 1363-1381.

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. *British Journal of Mathematical and Statistical Psychology, 61,* 235-253.

Koh, A. & Cribbie, R. A. (2013). Robust tests of equivalence for *k* independent groups. *British Journal of Mathematical and Statistical Psychology, 66,* 426–434.

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.

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

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://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.

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

Wellek, S. (2010). *Testing statistical hypotheses of equivalence, 2nd edition.* Boca Raton, FL: 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.

Equivalence intervals (ε), population variances (σ^2), distribution shapes (λ), and sample sizes (n) for the simulation study.

Condition	σ^2 (λ = Normal)		σ^2 (λ = χ^2, 3 *df*)	
Type	ε = .25	ε = .50	ε = .25	ε = .50
Type I Error	1, 1.22, 1.45, 1.67	1, 1.64, 2.28, 2.92	1, 1.28, 1.56, 1.84	1, 1.85, 2.70, 3.55
Power	1, 1, 1, 1	1, 1, 1, 1	1, 1, 1, 1	1, 1, 1, 1
Type I Error	1, 3, 4, 6	1, 3, 4, 6	1, 3, 4, 6	1, 3, 4, 6
Power	1, 1.1, 1.2, 1.3	1, 1.33, 1.66, 2	1, 1.1, 1.2, 1.3	1, 1.33, 1.66, 2
	Sample Size Conditions			
	$\bar{n}$ = 10	$\bar{n}$ = 25	$\bar{n}$ = 50	$\bar{n}$ = 100
equal samples sizes	10, 10, 10, 10	25, 25, 25, 25	50, 50, 50, 50	100, 100, 100, 100
positive pairings	5, 8, 12, 15	18, 22, 28, 32	25, 40, 60 75	50, 80, 120, 150
negative pairings	15, 12, 8, 5	32, 28, 22, 18	75, 60, 40, 25	150, 120, 80, 50

Table 2.

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 24 null conditions (only where $\Psi^2 = \varepsilon^2$).*

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	.0228	.1113	4
Levene-Wellek median	.0223	.0869	3
Levene-Wellek-Welch mean	.0196	.0850	3
Levene-Wellek-Welch median	.0356	.1014	1

**Note:* This table does not include conditions where $\Psi^2 > \varepsilon^2$.

Table 3.

Power summary: Number of conditions (out of 48 conditions) in which a specific equivalence procedure had the highest power (i.e., conditions where the null hypothesis was false).

Test	Test had Highest Power in Equal Sample Size Conditions (out of 16)	Test had Highest Power in Positive Pairing Conditions (out of 16*)	Test had Highest Power in Negative Pairing Conditions (out of 16)
Levene-Wellek mean	0	0	0
Levene-Wellek median	15	7	6
Levene-Wellek-Welch mean	0	0	0
Levene-Wellek-Welch median	1	6	10

* Excluding conditions where there was a tie for best performing procedure

		POPULATION: Equivalent	POPULATION: Not equivalent
SAMPLE: Equivalent		*Diff* – correct decision	*Diff* – Type II error
		Equiv – Power	*Equiv* – Type I error
SAMPLE: Not Equivalent		*Diff* – Type I error	*Diff* – Power
		Equiv – Type II error	*Equiv* – correct decision

Figure 1. Description of differences in the definition of power, Type I errors, and Type II errors between difference tests and equivalence tests. For the difference based tests, *Equivalent* implies that the variances are all identical, whereas for the equivalence-based tests, *Equivalent* implies that the difference in the variances is less than the minimum meaningful difference. Further, for the difference based tests *Not Equivalent* implies any difference in the variances whereas for the equivalence-based tests *Not Equivalent* implies a difference greater than or equal to the minimum meaningful difference. *Diff = difference-based test. Equiv = equivalence test.*

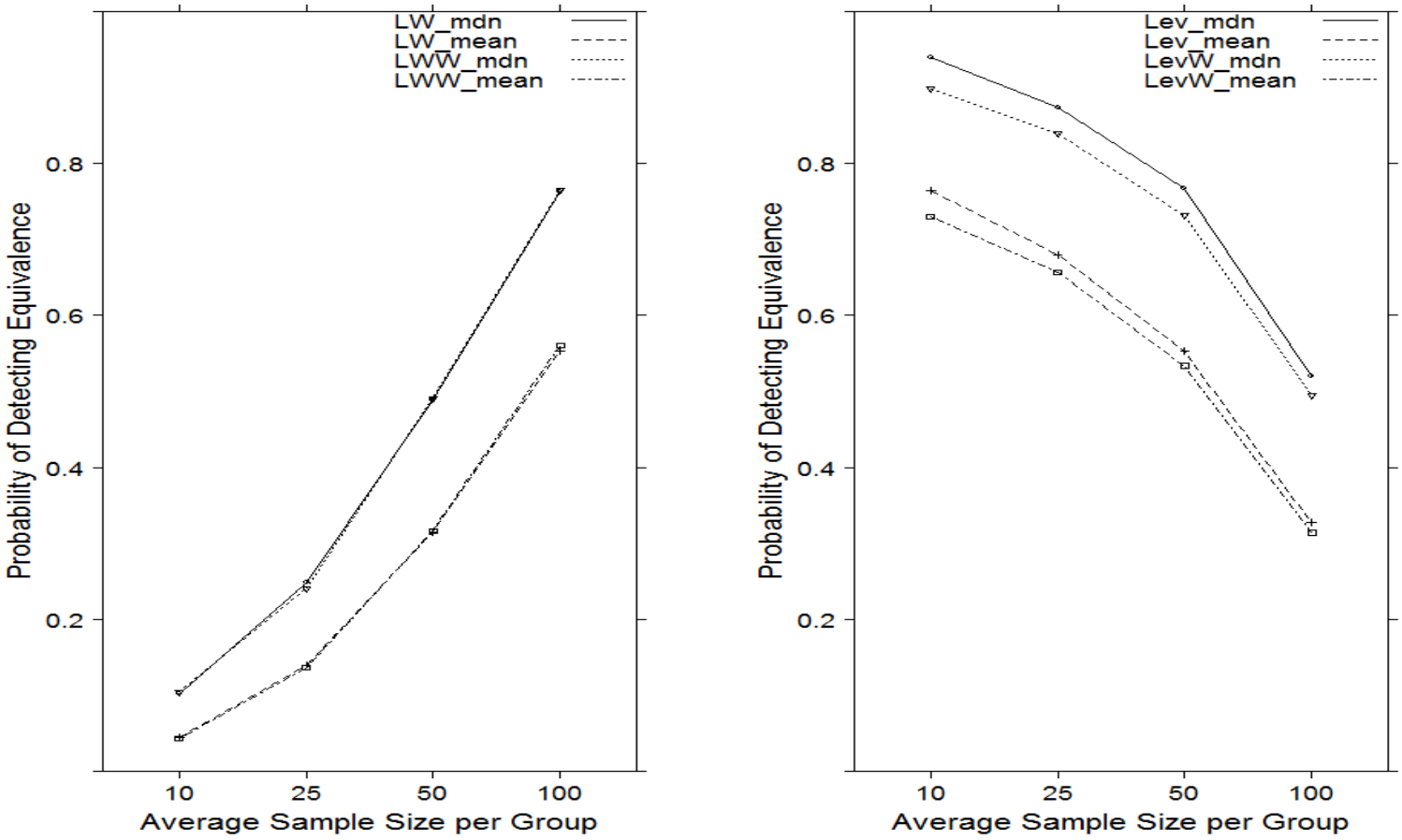

Figure 2. Average probability of declaring equivalence; ε = .50, σ^2 = 1, 1.33, 1.66, 2 ($\Psi^2 < \varepsilon^2$); χ^2 distributions; Left panel = equivalence procedures; Right panel = difference-based procedures

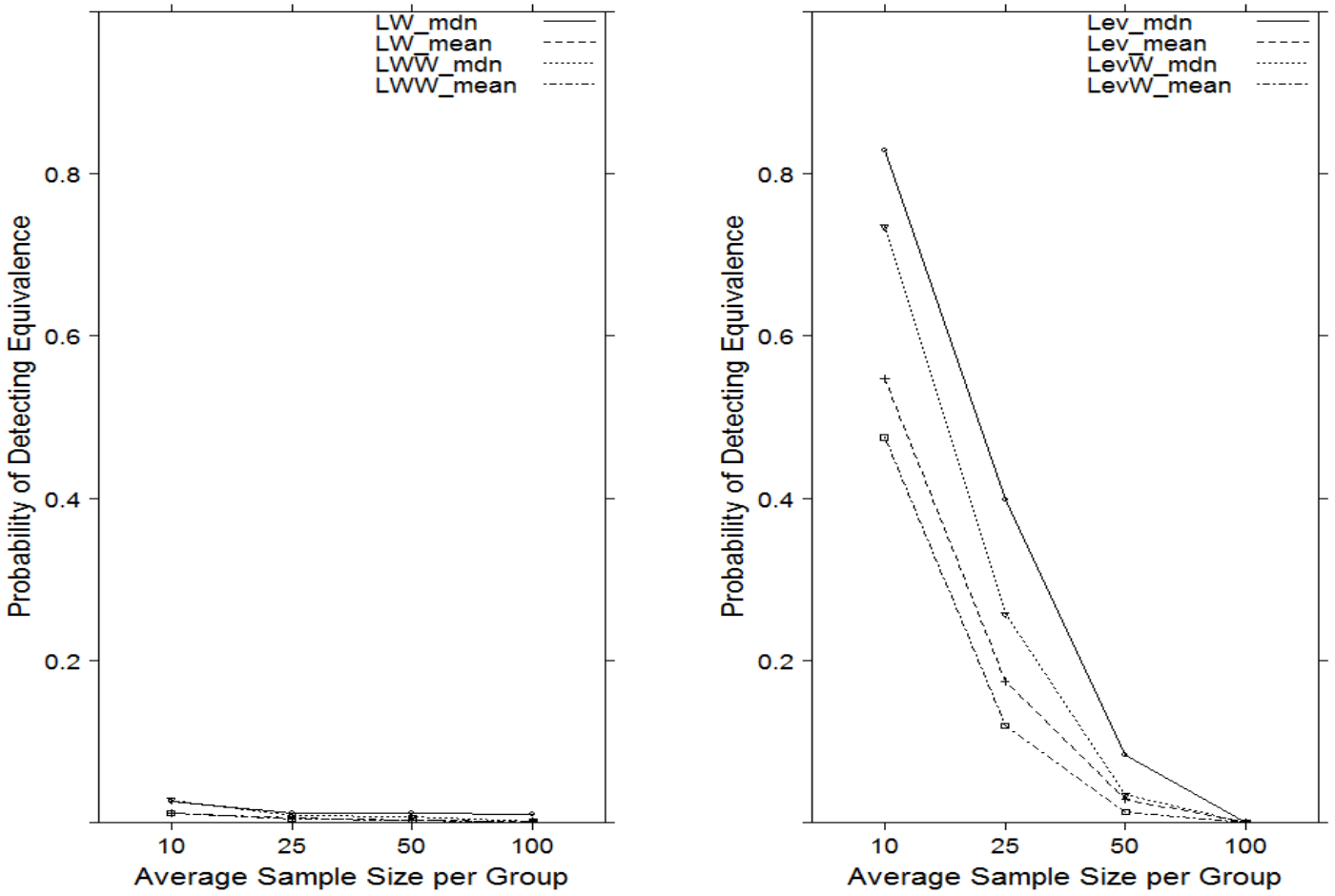

Figure 3. Average probability of declaring equivalence; ε = .50, σ^2 = 1, 3, 4, 6 ($\Psi^2 > \varepsilon^2$); χ^2 distributions; Left panel = equivalence procedures; Right panel = difference-based procedures.