British Journal of Mathematical and Statistical Psychology (2013), 66, 426–434

The British Psychological Society

www.wileyonlinelibrary.com

Robust tests of equivalence for *k* independent groups

Andy Koh and Robert Cribbie*

Quantitative Methods Program, Department of Psychology, York University, Toronto, Canada

A common question of interest to researchers in psychology is the equivalence of two or more groups. Failure to reject the null hypothesis of traditional hypothesis tests such as the ANOVA *F*-test (i.e., H_0: $\mu_1 = \ldots = \mu_k$) does not imply the equivalence of the population means. Researchers interested in determining the equivalence of *k* independent groups should apply a one-way test of equivalence (e.g., Wellek, 2003). The goals of this study were to investigate the robustness of the one-way Wellek test of equivalence to violations of homogeneity of variance assumption, and compare the Type I error rates and power of the Wellek test with a heteroscedastic version which was based on the logic of the one-way Welch (1951) *F*-test. The results indicate that the proposed Wellek–Welch test was insensitive to violations of the homogeneity of variance assumption, whereas the original Wellek test was not appropriate when the population variances were not equal.

1. Introduction

Researchers often want to evaluate whether two or more independent groups are equivalent. For example, Mead and Drasgow (1993) wanted to find if writing the paper-and-pencil version of the Graduate Record Exam was equivalent to writing the electronic test. Barker, Luman, McCauley, and Chu (2002) were interested in evaluating whether or not immunization coverage is equivalent across different cultural groups. In another example, Mueller, Liebig, and Hattrup (2007) did a study to investigate the psychometric equivalence of computerized and paper-and-pencil job satisfaction measures.

Researchers often mistake non-rejection of the null hypothesis with traditional tests (i.e., H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$, $i = 1,\ldots, k$) for equivalence; in other words, a lack of evidence to declare the groups different does not imply that they are equivalent. The rapidly expanding field of equivalence testing is starting to make this point more salient with researchers in psychology (e.g., Cribbie & Arpin-Cribbie, 2009; Rogers, Howard, & Vessey, 1993; Seaman & Serlin, 1998). Although traditional tests and tests of equivalence sometimes agree (e.g., the treatment means are deemed not statistically different using

*Correspondence should be addressed to Robert Cribbie, Quantitative Methods Program, Department of Psychology, York University, Toronto, ON, Canada M3J 1P3. (e-mail: cribbie@yorku.ca).

DOI:10.1111/j.2044-8317.2012.02056.x

traditional null hypothesis testing, and are deemed equivalent with a test of equivalence), since difference- and equivalence-based tests are examining different hypotheses, it also common that the treatment means are deemed not statistically different using traditional null hypothesis testing, but are not deemed equivalent with a test of equivalence, or that the treatment means are deemed statistically different using traditional null hypothesis testing, but are deemed equivalent using an equivalence test. More specifically, a traditional test of differences would be inappropriate for assessing equivalence for two reasons. First, as sample sizes increase, the likelihood of finding significant differences with a traditional test increases, whereas the probability of finding no significant difference decreases. Thus, if a researcher was interested in finding the groups equivalent using a traditional test of differences, then power would be maximized by using the minimum number of subjects. In other words, power and sample size, which in a proper null hypothesis testing environment should be directly related, are instead inversely related. Second, non-rejection of the null hypothesis associated with a traditional test of differences does not prove that the null hypothesis is true (only that it cannot be rejected). To summarize, a traditional test of differences cannot be used to answer questions relating to the equivalence of groups, and instead, as discussed below, researchers should adopt tests of equivalence.

1.1. Equivalence of multiple groups

Often researchers are interested in evaluating the equivalence of multiple independent groups. For example, a researcher may want to know if three or more different treatments for depression are equally effective. For example, in Barker *et al.*'s (2002) study discussed above, researchers set out to assess the equivalence of early childhood immunization coverage by different ethnic groups (Whites, Blacks, Hispanics, and Asians). Two popular options for evaluating the equivalence of multiple independent groups are a simultaneous test of equivalence and multiple pairwise tests of equivalence. Cribbie, Arpin-Cribbie, and Gruman (2010) recommend the use of Wellek's (2003) one-way test of equivalence for assessing the equivalence of multiple independent groups, instead of using multiple pairwise tests of equivalence, since the pairwise tests are overly conservative for assessing the simultaneous equivalence of all groups. In other words, Type I error rates are more accurate and power is higher when using a simultaneous test of the equivalence of k groups, such as that proposed by Wellek (2003), than if a researcher were to conduct all pairwise tests of equivalence.

1.2. Wellek's one-way test of equivalence

Wellek's (2003) one-way test of equivalence tests the null hypothesis H_0: $\varphi^2 \geq \varepsilon^2$ against the alternative H_a: $\varphi^2 < \varepsilon^2$, with ε representing the equivalence interval and

$$\varphi^2 = \frac{\sum_{i=1}^{k} \left(\frac{n_i}{\bar{n}}\right)\left(\bar{x}_i - \bar{x}_.\right)^2}{\sigma^2},$$

in which $\bar{n}$ stands for the mean sample size of the groups, $\bar{x}_i$ is the sample mean of the ith group, $\bar{x}_.$ is the average of the sample means for the k groups, and σ^2 is the average within group variability. If the combined difference between all the groups falls within the equivalence interval, then the researcher rejects the null hypothesis. The estimator of φ^2, $\hat{\varphi}^2$, incorporates the standard error from the fixed effects ANOVA F-test:

$$\hat{\varphi}^2 = \frac{\sum_{i=1}^{k} \left(\frac{n_i}{\bar{n}}\right)(\bar{x}_i - \bar{x}_{..})^2}{(N-k)^{-1} \sum_{i=1}^{k} \sum_{v=1}^{n_i} (x_{iv} - \bar{x}_i)^2}.$$

H_0 is rejected if $\hat{\varphi}^2 < \varphi_{crit}$, where

$$\varphi_{crit} = \left(\frac{k-1}{\bar{n}}\right) F_{k-1,N-k;\alpha}(\bar{n}\varepsilon^2).$$

$F_{k-1,N-k;\alpha}(\bar{n}\varepsilon^2)$ is the lower 100α percentage point of a non-central F with $k-1$ and $N-k$ degrees of freedom, where N is the total sample size and the non-centrality parameter is $\bar{n}\varepsilon^2$.

1.3. Wellek–Welch: A heteroscedastic Wellek procedure

A common problem that researchers encounter when conducting the Wellek one-way test of equivalence is that the assumption of variance homogeneity is violated. In fact, numerous studies have found that variances are often extremely different across independent groups (e.g., Golinski & Cribbie, 2009; Keselman *et al.*, 1998).

The Welch (1951) procedure, an alternative to the traditional one-way ANOVA F-test, does not require that the variances of the populations are equal. The Welch test can be represented by:

$$F' = \frac{\frac{\sum w_k (\bar{x}_k - \bar{x}'_.)^2}{k-1}}{1 + \frac{2(k-2)}{k^2-1} \sum \left(\frac{1}{n_k - 1}\right) \left(1 - \frac{w_k}{\sum w_k}\right)^2},$$

where $w_k = n_k / s_k^2$, $\bar{x}'_. = \sum w_k \bar{x}_k / \sum w_k$, n_k is the size of the kth group, s_k^2 is the variance of the kth group, and $\bar{x}_k$ is the mean of the kth group.

The traditional F statistic can be computed from the $\hat{\varphi}^2$ statistic as $F = \hat{\varphi}^2(\bar{n}/(k-1))$. Following that logic, the estimator of the new proposed statistic $\hat{\varphi}^{2\prime}$ (Wellek–Welch) can be defined as:

$$\hat{\varphi}^{2\prime} = F' \left(\frac{k-1}{\bar{n}}\right).$$

The F' statistic is approximately distributed as F with $k-1$ numerator degrees of freedom and df' denominator degrees of freedom, where

$$df' = \frac{k^2 - 1}{3 \sum \left(\frac{1}{n_k - 1}\right) \left(1 - \frac{w_k}{\sum w_k}\right)^2}.$$

The Wellek–Welch one-way test of equivalence, as with the original test, tests the null hypothesis H_0: $\varphi^2 \geq \varepsilon^2$ against the alternative H_a: $\varphi^2 < \varepsilon^2$ and H_0 is rejected if $\hat{\varphi}^{2\prime} < \varphi'_{crit}$, where

$$\varphi'_{crit} = \left(\frac{k-1}{\bar{n}}\right) F'_{k-1,df';\alpha}(\bar{n}\varepsilon^2).$$

2. Method

A simulation study was used to compare the performance of the Wellek one-way equivalence test (Wellek, 2003) with that of the heteroscedastic Wellek–Welch equivalence test. Wellek (2003) suggests setting the tolerance (equivalence interval) $\varepsilon = 0.36/\sqrt{2} \approx 0.25$ for a strict equivalence criterion, and $\varepsilon = 0.74/\sqrt{2} \approx 0.50$ for a liberal equivalence criterion. In this study, power conditions were investigated for both $\varepsilon = 0.25$ and $\varepsilon = 0.50$. Several variables were manipulated in the study, including: (1) number of groups; (2) sample size equality/inequality; (3) population variance equality/inequality; (4) pairings of unequal sample sizes and variances; and (5) population means.

We assessed the robustness of the Wellek and Wellek–Welch tests with four and seven independent groups. Four and seven groups were expected to span commonly encountered situations in psychological research. Group sample sizes were equal, slightly unequal, or highly unequal. The population variances of the groups were also set to be equal, slightly unequal, or highly unequal. Note that for both the four- and seven-group conditions the average group size was set to 50 and the average population variance was set to 10. Unequal sample sizes and variances were both positively (or directly) paired (smallest sample sizes paired with smallest variances and largest sample sizes paired with largest variances) and negatively (or inversely) paired (smallest sample sizes paired with largest variances and largest sample sizes paired with smallest variances). Table 1 provides a detailed summary of the sample size and variance conditions used in the Monte Carlo study.

The means of the groups were either all set to zero, or equally spaced apart. In order to evaluate Type I error rates, it was necessary to set the non-centrality parameter for the population *F* distribution equal to the non-centrality parameter for the critical distribution. Although Type I error rates for traditional tests of difference are evaluated when the population and critical *F* distributions are both central *F* distributions, for equivalence testing a null population distribution is actually a power condition (i.e., the population means are all equal), and thus in order to evaluate Type I error rates for *k* independent

Table 1. Summary of conditions investigated in the Monte Carlo study

K	n_i	σ_i^2	ε
4	50, 50, 50, 50	10, 10, 10, 10	0.25
	35, 45, 55, 65	4, 8, 12, 16	0.50
	20, 40, 60, 80	2.5, 5, 12.5, 20	
	65, 55, 45, 35		
	80, 60, 40, 20		
7	50, 50, 50, 50, 50, 50, 50	10, 10, 10, 10, 10, 10, 10	0.25
	35, 40, 45, 50, 55, 60, 65	4, 6, 8, 10, 12, 14, 16	0.50
	20, 30, 40, 50, 60, 70, 80	2.5, 4, 7, 10, 14, 17, 20	
	65, 60, 55, 50, 45, 40, 35		
	80, 70, 60, 50, 40, 30, 20		

Note. *k* represents the number of groups; n_i represents the group sample sizes; σ_i^2 represents the population variances, and ε represents the equivalence interval.

groups, the non-null population distribution needs to be set equal to the non-null critical F distribution. In this study, that was established by setting the population means for $k = 4$ and $\varepsilon = 0.25$ (using equal spacing) at $\mu_{4i} = \{0, 0.354, 0.708, 1.062\}$, and for $k = 4$ and $\varepsilon = 0.5$ the population means were set at $2\mu_{4i}$. Similarly, for $k = 7$ and $\varepsilon = 0.25$, $\mu_{7i} = \{0, 0.149, 0.298, 0.447, 0.596, 0.745, 0.894\}$ and for $k = 7$ and $\varepsilon = 0.5$, the population means were set at $2\mu_{7i}$. It is important to note that the non-centrality parameter for the Welch test (and hence the Wellek–Welch) differs from that for the traditional F (and hence Wellek F), by incorporating information about the ratio of the sample sizes and variances (see Levy, 1978). Thus, slight differences between the population means used for assessing the Type I error rates of the Wellek and Wellek–Welch tests occurred in order to ensure that the non-centrality parameter of the population F distribution for the Wellek–Welch was equal to that of the non-centrality parameter of the critical F distribution. In order to investigate power, the population means were all set equal to zero, or were evenly spaced such that the non-centrality parameter for the population distributions was less than that for the critical distributions. For $k = 4$, the population means were set at $\mu_1 = 0$, $\mu_2 = 0.25$, $\mu_3 = 0.5$, $\mu_4 = 0.75$, and for $k = 7$, the population means were set at $\mu_1 = 0$, $\mu_2 = 0.10$, $\mu_3 = 0.20$, $\mu_4 = 0.30$, $\mu_5 = 0.40$, $\mu_6 = 0.50$, $\mu_7 = 0.60$.

Five thousand simulations were performed for each condition using R version 2.12.1 (R Development Core Team, 2010). A nominal α level of .05 was used for all analyses.

3. Results

The patterns of results for $k = 4$ and $k = 7$, and for $\varepsilon = 0.25$ and $\varepsilon = 0.50$, were identical and therefore only the results for $k = 4$ and $\varepsilon = 0.50$ are presented. Complete tabulated results are available from the authors. For all power conditions it is important to note that if the empirical Type I error rates are not controlled within reasonable bounds around α, then the power results are not interpretable because the test is not robust to violations of the assumptions under those conditions. For this study a test was considered to be robust if its empirical Type I error rate fell within $\alpha - 0.2\alpha$ (.04) and $\alpha + 0.2\alpha$ (.06). These bounds fall between the liberal and conservative bounds proposed by Bradley (1978).

3.1. Type I error rates

Type I error rates for the Wellek and Wellek–Welch test for $\varepsilon = 0.50$ and $k = 4$ are presented in the first row of Figure 1. The Wellek–Welch test had Type I error rates that fell within the robustness criteria across all conditions (rates ranged from .042 to .058). However, the Type I error rates of the Wellek one-way test of equivalence depended strongly on the pairings of the unequal sample sizes and variances. For example, when $k = 4$ and extremely unequal variances and sample sizes were positively paired, the Type I error rates exceeded 10%, more than double the nominal rate. Further, when unequal variances and sample sizes were negatively paired, Type I error rates were biased downwards to as low as 3%.

3.2. Power – all means equal

The power rates for the Wellek and Wellek–Welch tests for $\varepsilon = 0.50$, $k = 4$ and equal means are presented in the second row of Figure 1. The Wellek–Welch test had power rates that were consistent across all conditions (approximately .82). However, the power

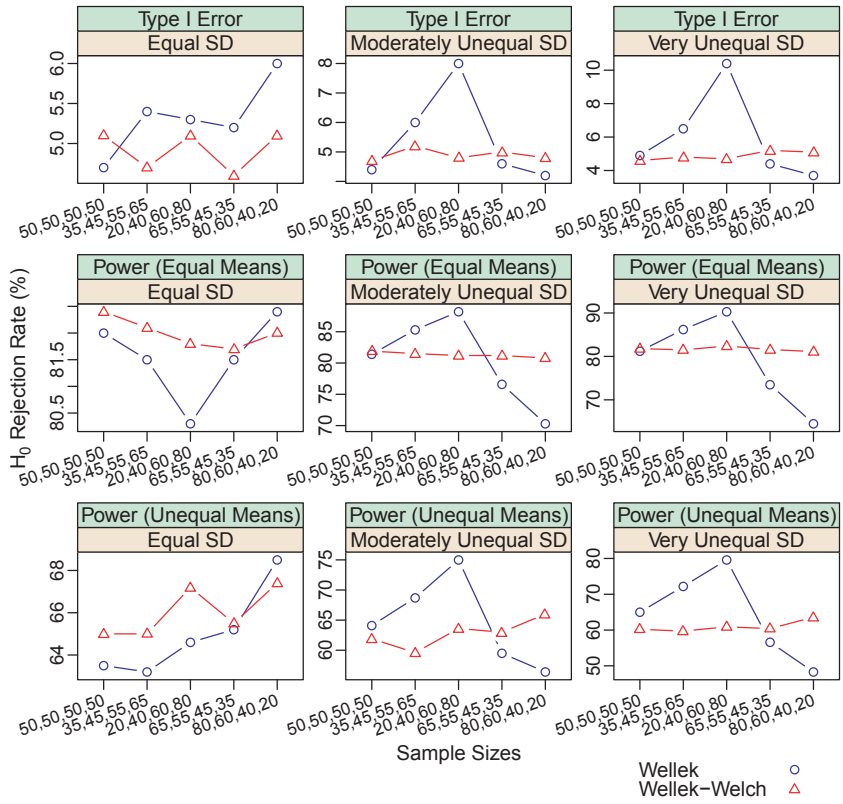

Figure 1. Type I error rates and power for evaluating equivalence with the Wellek and Wellek–Welch test statistics for $k = 4$ and $\varepsilon = 0.50$.

of the Wellek one-way test of equivalence depended strongly on the pairings of the unequal sample sizes and variances. For example, although the power of the Wellek test with equal variances or sample sizes was constant at about .82, when extremely unequal variances and sample sizes were positively paired the power rates were inflated to approximately .9, and when the unequal variances and sample sizes were negatively paired the power rates were deflated to approximately .65.

3.3. *Power – unequal means*

The power rates for the Wellek and Wellek–Welch tests for $\varepsilon = 0.50$, $k = 4$ and unequal means are presented in the third row of Figure 1. It is important to point out that although the means are unequal, the non-centrality parameter for the population mean differences does not exceed that of the critical value and therefore the null hypothesis (H_0: $\varphi^2 \geq \varepsilon^2$) is false and this is a power condition for a test of equivalence. When the means were unequal, the power rates, as expected, were reduced because the difference in the means was closer to the equivalence interval than when all the means were equal. The Wellek–Welch power rates, as with the equal means conditions, were generally consistent across all conditions. The only exception was that the power of the Wellek–Welch was slightly lower in the unequal variance conditions. This is due to the fact that the Wellek–Welch is sensitive to the pairings of the condition sample sizes, variances, and means; in other

words, differences between the sample means and the grand mean are weighted, where the weighting reflects the ratio of the sample sizes to the variances. As an example, consider a situation in which $k = 4$, the sample sizes are all 50, the variances are all 10, and the means are 0, 0.2, 0.4, 0.6 (this replicates the equal variance condition from Figure 1, but here assuming the values are sample statistics). With this data $\hat{\varphi}^{2\prime} = 0.017$. Now suppose that everything else stays the same but the sample sizes are 20, 40, 60, 80 and the variances are 2.5, 5, 12.5, 20 (this replicates the extremely unequal sample sizes and variances condition from Figure 1, but again we will assume that values are sample statistics). In this case $\hat{\varphi}^{2\prime} = 0.019$. Even though the mean differences stay the same and the average sample size and variance do not change, the value of $\hat{\varphi}^{2\prime}$ is increased slightly, which leads to the small decreases in power. However, there are two important conclusions to consider: (1) the decreases in power, even with extremely unequal sample sizes and variances, were almost always less than .05; and (2) the Wellek–Welch is not affected by the pairings of the group sample sizes and variances and provides a consistent level of power at each level of variance inequality.

On the other hand, the original Wellek test was again strongly affected by the pairings of the sample sizes and variances. For example, the Wellek test had power rates between approximately .63 and .65 with equal sample sizes, but when extremely unequal variances and sample sizes were positively paired the power rates were inflated to approximately .80, and when the unequal variances and sample sizes were negatively paired the power rates were deflated to below .50.

4. Discussion

When researchers want to assess the equivalence of multiple groups, they should conduct one-way tests of equivalence instead of using the non-significance of a traditional difference-based test or conducting multiple pairwise tests of equivalence. Wellek (2003) proposed a one-way test of equivalence, although this test is limited by the fact that it relies on the assumption of variance homogeneity. In this paper we propose a novel heteroscedastic one-way test of equivalence, called the Wellek–Welch test.

The results of this study demonstrate that the proposed Wellek–Welch test consistently outperformed the Wellek one-way test of equivalence when the assumption of homogeneity of variance was violated. More specifically, the Type I error rates of the Wellek–Welch were very consistent and fell between .04 and .06 (for $\alpha = .05$) under all conditions, whereas the Type I error rates of the Wellek test were biased upwards or downwards depending on the pairings of the sample sizes and variances. When larger sample sizes were paired with smaller variances (and hence smaller sample sizes were paired with larger variances), the Type I error rates for the Wellek test were deflated. On the other hand, when larger sample sizes were paired with larger variances (and hence smaller sample sizes were paired with smaller variances), the Type I error rates for the Wellek test were inflated. It is interesting to point out that these results are in the opposite direction to empirical Type I error rates when a traditional test of differences (e.g., ANOVA *F*) is applied when sample sizes and variances are unequal, because in the case of the traditional test a researcher is looking for an *F*-value larger than the critical value (whereas with equivalence testing a researcher is looking for a test statistic smaller than the critical value). Within the current equivalence testing environment (using the Wellek test), when a larger sample size is paired with a larger variance that larger variance gets weighted higher (than variances associated with smaller sized groups), which increases the

standard error and reduces the size of the test statistic (resulting in more declarations of equivalence).

With respect to power, and as expected given the Type I error results, the rates for the original Wellek test depended strongly on the combinations of the sample sizes and variances, whereas the power rates for the Wellek–Welch were generally consistent across sample size and variance conditions. In fact, since the Type I error rates of the Wellek–Welch, but not those of the original Wellek, are accurate across all combinations of equal and unequal sample sizes and variance, and the power of the Wellek–Welch is at least as high as the Wellek when variances are equal, there does not appear to be any situation in which the Wellek test outperforms the Wellek–Welch.

To conclude, one-way tests of equivalence should be used when trying to demonstrate the equivalence of the means of multiple groups. Whereas difference testing is appropriate for questions concerning disparity between group means, equivalence testing should be used to answer questions regarding the equivalence of group means. It is also possible that other strategies (e.g., Bayesian approaches – see Wellek, 2003) will be appropriate, depending on the nature of the research question and the desired outcome. However, the results of this study indicate that researchers should routinely adopt the Wellek–Welch test for conducting one-way tests of equivalence as it maintains accurate Type I error rates and consistent power when sample sizes and variances are equal or unequal.

References

Barker, L. E., Luman, E. T., McCauley, M. M., & Chu, S. Y. (2002). Assessing equivalence: An alternative to the use of difference tests for measuring disparities in vaccination coverage. *American Journal of Epidemiology*, *156*(11), 1056–1061.

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

Cribbie, R. A., & Arpin-Cribbie, C. A. (2009). Evaluating clinical significance through equivalence testing: Extending the normative comparisons approach. *Psychotherapy Research*, *19*(6), 677–686.

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

Golinski, C., & Cribbie, R. A. (2009). The expanding role of quantitative methodologists in advancing psychology. *Canadian Psychology*, *50*, 83–90.

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.

Levy, K. J. (1978). Some empirical power results associated with Welch's robust analysis of variance technique. *Journal of Statistical Computation and Simulation*, *8*, 43–48.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, *114*(3), 449–458.

Mueller, K., Liebig, C., & Hattrup, K. (2007). An investigation of the measurement equivalence of a multifaceted job satisfaction measure. *Educational and Psychological Measurement*, *67*(4), 648–678.

R Development Core Team (2010). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from: http://www.R-project.org/

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.

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. (2003). *Testing statistical hypotheses of equivalence*. Boca Raton, FL: Chapman & Hall/CRC.

Received 07 September 2011; revised version received 28 May 2012