The Journal of Experimental Education, 2010, 78, 1–13

ISSN: 0022-0973 print
DOI: 10.1080/00220970903224552

MEASUREMENT, STATISTICS, AND RESEARCH DESIGN

Tests of Equivalence for One-Way Independent Groups Designs

Robert A. Cribbie
York University

Chantal A. Arpin-Cribbie
Laurentian University

Jamie A. Gruman
University of Guelph

Researchers in education are often interested in determining whether independent groups are equivalent on a specific outcome. Equivalence tests for 2 independent populations have been widely discussed, whereas testing for equivalence with more than 2 independent groups has received little attention. The authors discuss alternatives for testing the equivalence of more than 2 independent populations, and they use a Monte Carlo study to demonstrate and compare the performance of these alternatives under several conditions. The results indicate that a 1-way test (e.g., Wellek's *F* test) is recommended for assessing the equivalence of more than 2 independent groups because approaches based on conducting pairwise tests of equivalence are overly conservative.

Keywords: equivalence testing, independent groups designs, Monte Carlo research

MANY EMPIRICAL QUESTIONS IN EDUCATIONAL RESEARCH involve assessing the differences among independent groups on a specific dependent

Address correspondence to Robert A. Cribbie, York University, Department of Psychology, 4700 Keele Street, Toronto, Onatario, Canada. E-mail: cribbie@yorku.ca

variable. For example, a researcher may be interested in demonstrating that mean scores differ for participants taking a paper-and-pencil test and for those taking a computer-based test. The null hypothesis in this case would be that the population group means are equal, and the researcher would typically use a two-independent-samples *t* test (or a one-way-independent-samples analysis of variance [ANOVA] if there were more than two groups) to evaluate this hypothesis. In fact, evaluating the null hypothesis that independent population means are equal accounts for almost all hypothesis testing involving independent groups, despite the fact that in many cases the researcher's primary interest is in whether or not the population means are equivalent. Testing the null hypothesis of equal population means is inappropriate for studies in which the primary objective is to demonstrate that groups are equivalent, rather than different, on a particular measure. In this case, equivalence tests are available for demonstrating that population means are equivalent—in other words, that any differences between the means of the populations can be considered trivial.

When using tests of equivalence, the goal is not to show that treatment conditions are perfectly identical, but that the differences between the treatments are too small to be considered meaningful. One example is an investigation in which an attempt is made to demonstrate that scores from a computer-based test are equivalent to those from a paper-and-pencil test (e.g., Epstein, Klinkenberg, Wiley, & McKinley, 2001). In this example, the researchers may not need to show that the test scores are exactly equivalent, as with the traditional null hypothesis (H_o: $\mu_1 = \mu_2$), but that any differences in test scores are inconsequential (i.e., $|\mu_1 - \mu_2| < D$, where D represents an a priori critical difference for determining equivalence). As Cribbie, Gruman, and Arpin-Cribbie (2004) and Rogers, Howard, and Vessey (1993) noted, the rejection or nonrejection of the null hypothesis of traditional tests tells us very little about the potential equivalence of the groups in question. More specifically, traditional tests of H_o: $\mu_1 = \mu_2$ (e.g., two independent-samples *t* test) evaluate whether the means are exactly identical, and larger sample sizes result in greater power for detecting any differences between the means. Hence, even minute differences between the means of the populations may be statistically significant with traditional tests; however, this result provides no valuable information to a researcher who would like to know whether the population means are equivalent.

Several tests have been designed to evaluate the equivalence of two population means, the test designed by Schuirmann (1987) being one of the most popular. Schuirmann's test of equivalence has been introduced to the behavioral sciences through influential articles by Rogers et al. (1993), Seaman and Serlin (1998), and others. The first step in applying Schuirmann's test of equivalence is to establish a critical mean difference for declaring two population means equivalent (D). Any mean difference smaller than D would be considered meaningless within the framework of the experiment. It is assumed that the two samples are randomly

and independently selected from normally distributed populations with equal variances. Two one-sided hypothesis tests can be used to establish equivalence, where the null hypothesis relates to the nonequivalence of the population means and can be expressed as two separate composite hypotheses, thus: $H_{o1} : \mu_1 - \mu_2 \geq D$; $H_{o2} : \mu_1 - \mu_2 \leq -D$.

Rejection of H_{o1} implies that $\mu_1 - \mu_2 < D$, and rejection of H_{o2} implies that $\mu_1 - \mu_2 > -D$. Further, rejection of both hypotheses implies that $\mu_1 - \mu_2$ falls within the bounds of (−D, D), and the means are deemed equivalent.

H_{o1} is rejected if $t_1 \leq -t_{\alpha,df}$ where

$$t_1 = \frac{(\bar{X}_1 - \bar{X}_2) - D}{\sqrt{\frac{(n_1+n_2)[(n_1-1)s_1^2+(n_2-1)s_2^2]}{n_1 n_2 (n_1+n_2-2)}}}$$

and H_{o2} is rejected if $t_2 \geq t_{\alpha,df}$ where

$$t_2 = \frac{(\bar{X}_1 - \bar{X}_2) - (-D)}{\sqrt{\frac{(n_1+n_2)[(n_1-1)s_1^2+(n_2-1)s_2^2]}{n_1 n_2 (n_1+n_2-2)}}}$$

$\bar{X}_1$ and $\bar{X}_2$ are the group means, n_1 and n_2 are the group sample sizes, s_1 and s_2 are the group standard deviations, and $t_{\alpha,df}$ is the upper-tailed α-level t critical value with $n_1 + n_2 - 2$ degrees of freedom. Several articles have discussed the use of equivalence tests in two-independent-group designs (e.g., Cribbie et al., 2004; Rogers et al., 1993; Seaman & Serlin, 1998; Tryon, 2001); more recently, Gruman, Cribbie, and Arpin-Cribbie (2007) have discussed the use of a heteroscedastic version of the Schuirmann test statistic, the Schuirmann-Welch test, where the denominator in both t_1 and t_2 is replaced with

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

and the degrees of freedom are replaced by

$$df_w = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$$

Although the Schuirmann-Welch test is designed for situations in which the population variances are unequal (which is not the case in the present study), the Schuirmann-Welch test performs well both when variances are equal and when

they are unequal; therefore, it is included as an alternative test in the present research and recommended as a generic test for assessing the equivalence of two independent groups (whether the variances are equal or unequal).

Although equivalence tests for two-independent groups have been discussed frequently, the case of general one-way-independent-groups designs has received little attention even though hypotheses concerning equivalence often deal with more than two groups. For example, one researcher might be interested in demonstrating that satisfaction with university life is not a function of type of living accommodation, where students who live on-campus, at home with their parents, off-campus alone, or off-campus with others, all score similarly on satisfaction with university life. Another researcher might be interested in demonstrating that students from many cultural backgrounds (e.g., North American, South American, Asian, European) score equivalently on standardized tests (e.g., Graduate Record Exam). Important methodological questions arise when more than two groups are being considered, including what form of test should be applied.

Researchers familiar with the popular two-independent-samples tests of equivalence (e.g., Schuirmann test) would be likely to evaluate the equivalence of J population means by demonstrating in a pairwise manner that each group was equivalent to every other group. For example, if a researcher wanted to demonstrate that three population means are equivalent, he or she might demonstrate that the first population is equivalent to the second population, that the first population is equivalent to the third population, and that the second population is equivalent to the third population. Another option is to evaluate the equivalence of J population means by demonstrating the equivalence of the two means with the largest mean difference.

An alternative approach, outlined by Wellek (2003), is to use a one-way test of equivalence, where the equivalence of all J population means is simultaneously evaluated.

The null hypothesis for a one-way equivalence test would be that the combined difference between multiple groups falls within an equivalence interval. Wellek suggested the following hypotheses: H_o: $\Psi^2 \geq \varepsilon^2$, H_a: $\Psi^2 < \varepsilon^2$, where ε is the equivalence interval and

$$\Psi^2 = \frac{\sum_{i=1}^{j} \left(\frac{n_i}{\bar{n}}\right) (\bar{X}_i - \bar{X}.)^2}{\sigma^2}$$

$\bar{n}$ represents the mean sample size of the groups, $\bar{X}_i$ represents the sample mean of the ith population, $\bar{X}.$ represents the average of the sample means for the J populations, and σ^2 represents the average within group variability (assumed to

be equal across populations). H_o: $\Psi^2 \geq \varepsilon^2$ is rejected if $\Psi^2 < \Psi_{crit}$, where

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

and $\bar{n}\varepsilon^2$ represents the noncentrality parameter.

The purpose of the present study was to evaluate the available approaches to assessing the equivalence of J independent groups and to be able to make recommendations regarding the most appropriate test.

METHOD

A simulation study was used to compare the performance of three approaches to assessing the equivalence of J independent groups: (a) Wellek's one-way equivalence test (Wellek, 2003), (b) multiple pairwise Schuirmann tests (where the J populations are considered equivalent if each group is equivalent to each other group) using either the original Schuirmann statistic (S_P) or the heteroscedastic Schuirmann-Welch statistic (SW_P), and (c) largest mean difference Schuirmann test (where the J populations are considered equivalent if the two groups with the largest mean difference are declared equivalent) using either the original Schuirmann statistic (S_L) or the heteroscedastic Schuirmann-Welch statistic (SW_L). For the Wellek procedure, ε^2 was set equal to Ψ^2 for the Type I error conditions. For the Schuirmann test, D was set equal to 1 for all tests. Several variables were manipulated in this study including (a) number of groups ($J = 3$ and 5), (b) average sample size (average $n = 20$ and average $n = 75$), (c) degree of sample size heterogeneity (equal n, moderately unequal n, extremely unequal n), and (d) population mean configuration (see Table 1).

Ten thousand simulations were performed using the SAS software package, specifically SAS's Interactive Matrix Language package (SAS Institute, 1999). Normally distributed random variables were generated using the RANNOR random number generator. A nominal α level of .05 was used for all analyses.

RESULTS

$J = 3$

The Type I error and power rates for the Wellek, S_P, SW_P, S_L, and SW_L, with three independent groups are presented in Tables 2 and 3. When there were three independent groups, the Type I error rates for the Wellek one-way test of

TABLE 1
Population Mean (μ_j) Configurations Used in the Monte Carlo Study

Configuration Type	*Type I Error Condition*	*Power Conditions*
J = 3		
Not Equally Spaced	$\mu_j = 0,0, 1$	$\mu_j = 0,0, .8$
		$\mu_j = 0,0, .6$
Equally Spaced	$\mu_j = 0,.5, 1$	$\mu_j = 0,.4, .8$
		$\mu_j = 0,.3, .6$
J = 5		
Not Equally Spaced	$\mu_j = 0,0, 0, 1, 1$	$\mu_j = 0,0, 0, .8, .8$
		$\mu_j = 0,0, 0, .6, .6$
Equally Spaced	$\mu_j = 0,.25, .5, .75, 1$	$\mu_j = 0,.2, .4, .6, .8$
		$\mu_j = 0,.15, .3, .45, .6$

equivalence were maintained at approximately α under all conditions. The empirical Type I error rates for the S_P, SW_P, S_L, and SW_L were very conservative with the population mean configuration $\mu_1 = 0$, $\mu_2 = 0$, $\mu_3 = 1$, ranging between .007 and .014. The empirical Type I error rates for the S_P, SW_P, S_L, and SW_L were less conservative with the population mean configuration $\mu_1 = 0$, $\mu_2 = .5$, $\mu_3 = 1$, and $n = 20$, ranging between .032 and .037, and accurate with the population mean configuration $\mu_1 = 0$, $\mu_2 = .5$, $\mu_3 = 1$, and $n = 75$, ranging between .048 and .049.

The power results closely mirrored those of the Type I error results. The empirical power rates for the S_P, SW_P, S_L, and SW_L were considerably lower with the population mean configuration $\mu_1 = 0$, $\mu_2 = 0$, $\mu_3 = 1$ than were the rates for the Wellek one-way equivalence test across all conditions. The empirical power rates for the S_P, SW_P, S_L, and SW_L were lower with the population mean configuration $\mu_1 = 0$, $\mu_2 = .5$, $\mu_3 = 1$, and $n = 20$, relative to the rates for the Wellek one-way equivalence test, although with an average $n = 75$ the empirical power of the S_P, SW_P, S_L, and SW_L was slightly larger than that of the Wellek one-way equivalence test.

$J = 5$

The Type I error and power rates for the Wellek, S_P, SW_P, S_L and SW_L with five independent groups are presented in Tables 4 and 5. When there were five independent groups, the Type I error rates for the Wellek one-way test of equivalence were maintained at approximately α under all conditions. The empirical Type I error rates for the S_P, SW_P, S_L, and SW_L were conservative across all conditions, with the rates extremely conservative for the population

TABLE 2
Probability of Declaring Three Normally Distributed Populations Equivalent, With the Population Mean Pattern for Declaring Equivalence Equal to $\mu_1 = 0$, $\mu_2 = 0$, $\mu_3 = 1$

μ_1, μ_2, μ_3	n_1, n_2, n_3	ε	*Wellek*	*D*	*S_P*	*SW_P*	*S_L*	*SW_L*
Type I Error Results								
0, 0, 1	20, 20, 20	.816	.049	1	.012	.012	.013	.013
	15, 20, 25	.866	.051	1	.010	.010	.011	.011
	10, 20, 30	.913	.054	1	.007	.007	.011	.011
	75, 75, 75	.816	.046	1	.010	.010	.014	.014
	70, 75, 80	.830	.048	1	.011	.011	.013	.013
	60, 75, 90	.856	.050	1	.011	.011	.011	.011
Power Results								
0, 0, .8	20, 20, 20	.816	.163	1	.056	.055	.057	.057
	15, 20, 25	.866	.175	1	.049	.048	.052	.052
	10, 20, 30	.913	.199	1	.036	.033	.045	.045
	75, 75, 75	.816	.391	1	.188	.187	.190	.189
	70, 75, 80	.830	.396	1	.175	.174	.179	.178
	60, 75, 90	.856	.409	1	.177	.177	.178	.178
0, 0, .6	20, 20, 20	.816	.389	1	.181	.181	.186	.185
	15, 20, 25	.866	.424	1	.168	.164	.172	.171
	10, 20, 30	.913	.445	1	.121	.110	.145	.144
	75, 75, 75	.816	.857	1	.671	.671	.672	.672
	70, 75, 80	.830	.868	1	.674	.674	.676	.676
	60, 75, 90	.856	.882	1	.666	.664	.668	.666

Note. ε = equivalence interval for the equivalence F-test; Wellek = equivalence F-test; D = equivalence interval for the Schuirmann equivalence t-test; S_P = Pairwise Schuirmann test of equivalence; SW_P = Pairwise Schuirmann-Welch test of equivalence; S_L = Schuirmann test of equivalence on the largest pairwise mean difference; SW_L = Schuirmann-Welch test of equivalence on the largest pairwise mean difference.

mean configuration $\mu_1 = 0$, $\mu_2 = 0$, $\mu_3 = 0$, $\mu_4 = 1$, $\mu_5 = 1$. The empirical Type I error rates for the S_P, SW_P, S_L, and SW_L were less conservative with the population mean configuration $\mu_1 = 0$, $\mu_2 = .25$, $\mu_3 = .5$, $\mu_4 = .75$, $\mu_5 = 1$, although the rates for $n = 20$ never exceeded .012, and the rates for $n = 75$ never exceeded .035.

The power results again mirrored those of the Type I error results. The empirical power rates for the S_P, SW_P, S_L, and SW_L were considerably lower with the population mean configuration $\mu_1 = 0$, $\mu_2 = 0$, $\mu_3 = 0$, $\mu_4 = 1$, $\mu_5 = 1$ than were the rates for the Wellek one-way equivalence test across all conditions, and they were also consistently lower for the S_P, SW_P, S_L, and SW_L with the mean configuration $\mu_1 = 0$, $\mu_2 = .25$, $\mu_3 = .5$, $\mu_4 = .75$, $\mu_5 = 1$, relative to the Wellek one-way procedure.

TABLE 3
Probability of Declaring Three Normally Distributed Populations Equivalent, With the Population Mean Pattern for Declaring Equivalence Equal to $\mu_1 = 0$, $\mu_2 = .5$, $\mu_3 = 1$

μ_1, μ_2, μ_3	n_1, n_2, n_3	ε	*Wellek*	*D*	*S_P*	*SW_P*	*S_L*	*SW_L*
Type I Error Results								
0, .5, 1	20, 20, 20	.707	.046	1	.036	.035	.036	.035
	15, 20, 25	.707	.055	1	.037	.036	.037	.036
	10, 20, 30	.707	.062	1	.032	.034	.032	.034
	75, 75, 75	.707	.048	1	.048	.048	.048	.048
	70, 75, 80	.707	.048	1	.049	.049	.049	.049
	60, 75, 90	.707	.052	1	.048	.048	.048	.048
Power Results								
0, .4, .8	20, 20, 20	.707	.141	1	.108	.108	.108	.108
	15, 20, 25	.707	.146	1	.102	.102	.102	.102
	10, 20, 30	.707	.169	1	.093	.089	.093	.089
	75, 75, 75	.707	.327	1	.339	.339	.339	.339
	70, 75, 80	.707	.314	1	.322	.322	.322	.322
	60, 75, 90	.707	.336	1	.333	.333	.333	.333
0, .3, .6	20, 20, 20	.707	.316	1	.260	.259	.260	.259
	15, 20, 25	.707	.322	1	.247	.240	.247	.240
	10, 20, 30	.707	.325	1	.195	.190	.195	.190
	75, 75, 75	.707	.762	1	.779	.779	.779	.779
	70, 75, 80	.707	.761	1	.783	.783	.783	.783
	60, 75, 90	.707	.763	1	.774	.772	.774	.772

Note. ε = equivalence interval for the equivalence F-test; Wellek = equivalence F-test; D = equivalence interval for the Schuirmann equivalence t-test; S_P = Pairwise Schuirmann test of equivalence; SW_P = Pairwise Schuirmann-Welch test of equivalence; S_L = Schuirmann test of equivalence on the largest pairwise mean difference; SW_L = Schuirmann-Welch test of equivalence on the largest pairwise mean difference.

DISCUSSION

The Wellek one-way test of equivalence performed very well across all conditions investigated in this study. The Type I error rates were very accurate, and the power was generally much larger for the Wellek procedure than it was for the Schuirmann pairwise approach or the Schuirmann approach based on the largest mean difference between groups. There was little difference between the Schuirmann pairwise approach and the Schuirmann approach based on the largest mean difference across all of the conditions in the present investigation. These results indicate that, although researchers may be more familiar with two-sample based equivalence tests for independent groups, when there are more than two groups, there is much to be gained by adopting a one-way test of equivalence

TABLE 4
Probability of Declaring Five Normally Distributed Populations Equivalent, With the Population Mean Pattern for Declaring Equivalence Equal to $\mu_1 = 0$, $\mu_2 = 0$, $\mu_3 = 0$, $\mu_4 = 1$, $\mu_5 = 1$

$\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$	n_1, n_2, n_3, n_4, n_5	ε	*Wellek*	*D*	*S_P*	*SW_P*	*S_L*	*SW_L*
Type I Error Results								
0, 0, 0, 1, 1	20, 20, 20, 20, 20	1.095	.050	1	.000	.000	.000	.000
	14, 17, 20, 23, 26	1.136	.056	1	.000	.000	.000	.000
	10, 15, 20, 25, 30	1.162	.055	1	.000	.000	.000	.000
	75, 75, 75, 75, 75	1.095	.052	1	.001	.001	.001	.001
	65, 70, 75, 80, 85	1.114	.053	1	.000	.000	.000	.000
	55, 65, 75, 85, 95	1.131	.052	1	.000	.000	.000	.000
Power Results								
0, 0, 0, .8, .8	20, 20, 20, 20, 20	1.095	.224	1	.004	.004	.005	.005
	14, 17, 20, 23, 26	1.136	.240	1	.003	.003	.004	.004
	10, 15, 20, 25, 30	1.162	.251	1	.002	.002	.004	.004
	75, 75, 75, 75, 75	1.095	.567	1	.032	.032	.033	.033
	65, 70, 75, 80, 85	1.114	.574	1	.033	.032	.037	.037
	55, 65, 75, 85, 95	1.131	.578	1	.028	.027	.030	.029
0, 0, 0, .6, .6	20, 20, 20, 20, 20	1.095	.555	1	.033	.033	.037	.037
	14, 17, 20, 23, 26	1.136	.575	1	.024	.023	.030	.032
	10, 15, 20, 25, 30	1.162	.592	1	.019	.020	.027	.029
	75, 75, 75, 75, 75	1.095	.972	1	.402	.402	.404	.404
	65, 70, 75, 80, 85	1.114	.975	1	.406	.405	.409	.408
	55, 65, 75, 85, 95	1.131	.978	1	.393	.391	.397	.396

Note. ε = equivalence interval for the equivalence F-test; Wellek = equivalence F-test; D = equivalence interval for the Schuirmann equivalence t-test; S_P = Pairwise Schuirmann test of equivalence; SW_P = Pairwise Schuirmann-Welch test of equivalence; S_L = Schuirmann test of equivalence on the largest pairwise mean difference; SW_L = Schuirmann-Welch test of equivalence on the largest pairwise mean difference.

rather than adopting an approach that assesses equivalence using only two groups at a time.[1]

These results are interesting because they contradict recommendations for conducting traditional tests of the difference between groups (e.g., one-way ANOVA F test) in one-way designs. For example, Bernhardson (1975) and Hancock and Klockars (1996) explained that conducting pairwise multiple comparison tests of the J groups only when an omnibus test is statistically significant is not recommended unless rejection of the omnibus test is required for the multiple comparison procedure. In other words, if a researcher were interested in determining if the means of three independent groups were different, and he or she intended on using Tukey's popular honestly significant difference (HSD) multiple comparison procedure (which does not require a significant omnibus test for use) for family-wise error control, he or she should conduct the pairwise multiple

TABLE 5
Probability of Declaring Five Normally Distributed Populations Equivalent, With the Population Mean Pattern for Declaring Equivalence Equal to $\mu_1 = 0$, $\mu_2 = .25$, $\mu_3 = .5$, $\mu_4 = .75$, $\mu_5 = 1$

μ_1, μ_2, μ_3, μ_4,μ_5	n_1,n_2,n_3,n_4,n_5	ε	*Wellek*	*D*	*S_P*	*SW_P*	*S_L*	*SW_L*
Type I Error Results								
0, .25, .5, .75, 1	20, 20, 20, 20, 20	.791	.053	1	.010	.010	.012	.012
	14, 17, 20, 23, 26	.791	.056	1	.010	.009	.010	.011
	10, 15, 20, 25, 30	.791	.060	1	.006	.005	.007	.007
	75, 75, 75, 75, 75	.791	.050	1	.030	.030	.033	.033
	65, 70, 75, 80, 85	.791	.052	1	.034	.034	.035	.035
	55, 65, 75, 85, 95	.791	.051	1	.033	.033	.034	.035
Power Results								
0, .2, .4, .6, .8	20, 20, 20, 20, 20	.791	.156	1	.038	.038	.039	.039
	14, 17, 20, 23, 26	.791	.162	1	.035	.034	.043	.043
	10, 15, 20, 25, 30	.791	.168	1	.028	.027	.038	.038
	75, 75, 75, 75, 75	.791	.370	1	.290	.290	.290	.290
	65, 70, 75, 80, 85	.791	.377	1	.287	.288	.288	.288
	55, 65, 75, 85, 95	.791	.365	1	.271	.269	.272	.273
0, .15, .3, .45, .6	20, 20, 20, 20, 20	.791	.336	1	.106	.106	.115	.115
	14, 17, 20, 23, 26	.791	.336	1	.096	.096	.112	.112
	10, 15, 20, 25, 30	.791	.344	1	.069	.067	.093	.097
	75, 75, 75, 75, 75	.791	.832	1	.753	.753	.754	.754
	65, 70, 75, 80, 85	.791	.835	1	.745	.746	.747	.747
	55, 65, 75, 85, 95	.791	.831	1	.721	.723	.722	.723

Note. ε = equivalence interval for the equivalence F-test; Wellek = equivalence F-test; D = equivalence interval for the Schuirmann equivalence t-test; S_P = Pairwise Schuirmann test of equivalence; SW_P = Pairwise Schuirmann-Welch test of equivalence; S_L = Schuirmann test of equivalence on the largest pairwise mean difference; SW_L = Schuirmann-Welch test of equivalence on the largest pairwise mean difference.

comparison procedures regardless of whether or not the omnibus test is significant. If the researcher only conducted the pairwise tests if the omnibus test were statistically significant, the empirical Type I error rates and power would be biased downward. This is an important consideration that is often overlooked by applied researchers. This also contradicts the findings of the present study that conducting all pairwise tests of equivalence without conducting an omnibus test would bias the Type I error and power rates downward.

An interesting question that emerges is why the approaches based on the Schuirmann test statistic are generally conservative. The answer to this question is in the fact that declaring all J groups equivalent requires multiple statistically significant test statistics.[2] Therefore, the probability of declaring all J groups equivalent is a function of the product of the probabilities of declaring each

pairwise mean difference equivalent. For example, with the population mean configuration $\mu_1 = 0$, $\mu_2 = 0$, $\mu_3 = 1$, and the critical difference (D) set equal to 1, one knows that if all the assumptions are met, the probability of a Type I error for the pairwise null hypotheses H_o: $\mu_1 = \mu_3$ and H_o: $\mu_2 = \mu_3$ would each be approximately .05 (assuming $\alpha = .05$). The power of H_o: $\mu_1 = \mu_2$ would depend for example on the sample size and the variability. It should be clear that unless H_o: $\mu_1 = \mu_3$ and H_o: $\mu_2 = \mu_3$ were perfectly correlated and the power of H_o: $\mu_1 = \mu_2$ equalled 1.0, the Type I error rate of the pairwise Schuirmann approach would be less than .05.[3] Using this logic, one can also see that when the population mean configuration is $\mu_1 = 0$, $\mu_2 = .5$, $\mu_3 = 1$, the approach will be less conservative because now the empirical Type I error rate for declaring all J groups equivalent with the Schuirmann approach will equal .05 when the Type I error rate for H_o: $\mu_1 = \mu_3$ equals .05 and the power of H_o: $\mu_1 = \mu_2$ and H_o: $\mu_2 = \mu_3$ equals 1.

An important consideration in adopting the Wellek (2003) one-way test of equivalence is what value to use for ε. Wellek recommends adopting $\varepsilon = .25$ for a strict equivalence criterion and $\varepsilon = .50$ for a liberal equivalence criterion. It is also important to note that ε^2 can be computed using the formula for Ψ^2 in the case where researchers have established population values that establish the bounds for equivalence (and the average within-group variance is known).

In summary, many empirical questions in educational research involve demonstrating the equivalence of multiple groups. For example, educational researchers may be interested in determining whether two pedagogical methods produce equivalent learning outcomes. Tests of the null hypothesis H_o: $\mu_1 = \ldots = \mu_J$, where J represents the number of groups, are inappropriate because failing to reject H_o does not imply that the groups are equal, and further, the probability of declaring the groups equivalent decreases (rather than increases) as sample size increases. Instead, tests of equivalence allow researchers to evaluate whether differences among groups are too small to be considered meaningful, where the researcher controls what difference is no longer meaningful. Although tests of equivalence are relatively new to educational researchers, we expect that as these tests become more popular, researchers will be able to use equivalence tests to address novel research questions that were previously avoided because of a lack of appropriate methodology. The results of the present study suggest that educational researchers conducting one-way tests of equivalence (i.e., assessing the equivalence of multiple independent groups) should use a one-way equivalence test (such as that proposed by Wellek, 2003), rather than a pairwise approach to assessing the equivalence of the means, to ensure that Type I error rates are accurate and power is maximized. With regard to the one-way equivalence test evaluated in the present study, pairwise approaches generally produced conservative results that are less powerful for detecting true equivalencies among means.

NOTES

1. It is also possible to conduct the Wellek one-way equivalence test with only two groups, although we found slightly inflated Type I error rates for the Wellek test with only two groups, and we therefore recommend the Schuirmann two-independent-groups equivalence test for this design.
2. Even when we are only comparing the largest difference between means, there is an expectation that smaller differences between means will be statistically significant, even though this is not always the case and hence the approach using only the largest difference between means is sometimes slightly more powerful.
3. Although it is more difficult to see, this also applies to the case of declaring all J groups equivalent if the largest mean difference is declared equivalent because generally this will only occur when all null hypotheses are rejected.

AUTHOR NOTES

Robert A. Cribbie is an associate professor in the Department of Psychology at York University in Toronto, Ontario, Canada. His current research interests include equivalence testing, robust test statistics for ANOVA designs, and multiple testing procedures. **Chantal A. Arpin-Cribbie** is an assistant professor in the Department of Psychology at Laurentian University. Her current research focuses on clinical and health psychology, including novel intervention approaches and the methodological considerations in intervention research. **Jamie A. Gruman** is an assistant professor of organizational behavior at the University of Guelph. His current research interests include employee engagement, organizational socialization, individual differences at work, and management education.

REFERENCES

Bernhardson, C. S. (1975). Type I error rates when multiple comparison procedures follow a significant F test of ANOVA. *Biometrics*, *31*, 229–232.

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

Epstein, J., Klinkenberg, W. D., Wiley, D., & McKinley, L. (2001). Insuring sample equivalenc across internet and paper-and-pencil assessments. *Computers in Human Behavior*, *17*, 339–346.

Gruman, J., Cribbie, R. A., & Arpin-Cribbie, C. A. (2007). The effects of heteroscedasticity on tests of equivalence. *Journal of Modern Applied Statistical Methods*, *6*, 133–140.

Hancock, G. R., & Klockars, A. J. (1996). The quest for a: Developments in multiple comparisons procedures in the quarter century since Games (1971). *Review of Educational Research*, *66*, 269–306.

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.

SAS Institute (1999). *SAS/IML User's Guide, Version 8, Volumes 1 and 2*. Cary, NC: SAS Institute.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing 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.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, *6*, 371–386.

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