Testing for Negligible Interaction: A Coherent and Robust Approach

Robert A. Cribbie, Chantal Ragoonanan, & Alyssa Counsell

Quantitative Methods Program

Department of Psychology

York University

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

Abstract

Researchers often want to demonstrate a lack of interaction between two categorical predictors on an outcome. To justify a lack of interaction, researchers typically accept the null hypothesis of no interaction from a conventional analysis of variance (ANOVA). This method is inappropriate as failure to reject the null hypothesis does not provide statistical evidence to support a lack of interaction. This study proposes a bootstrap-based intersection-union test for negligible interaction that provides coherent decisions between the omnibus test and post hoc interaction contrast tests and is robust to violations of the normality and variance homogeneity assumptions. Further, a multiple comparison strategy for testing interaction contrasts following a nonsignificant omnibus test is proposed. Our simulation study compared the Type I error control, omnibus power and per-contrast power of the proposed approach to the noncentrality-based negligible interaction test of Cheng and Shao (2007). For 2 x 2 designs, the empirical Type I error rates of the Cheng and Shao test were very close to the nominal α level when the normality and variance homogeneity assumptions were satisfied, however only our proposed bootstrapping approach was satisfactory under nonnormality and/or variance heterogeneity. In general *a* x *b* designs, although the omnibus Cheng and Shao test, as expected, is the most powerful, it is not robust to assumption violation and results in incoherent omnibus and interaction contrast decisions that are not possible with the intersection-union approach.

Testing for Negligible Interaction: A Coherent and Robust Approach

Psychological researchers are frequently interested in finding a lack of interaction between two or more variables on an outcome. A lack of interaction might be the primary hypothesis of the study or a lack of interaction may be necessary in order to justify removing an interaction term from a model or interpreting the main effects. For example**,** Paolieri, Lotto, Leoncini, Cubelli, and Job (2011) conducted a study investigating the process by which the congruency of target and distractor noun genders (congruent/incongruent) and ending identity (same distractor word ending/different distractor word ending) affect picture naming latencies. It was hypothesized and established that no interaction would occur between gender congruence and ending identity because the processing of the gender congruency and ending identity were expected to be independent.

It is essential to recognize that the example above utilized a factorial analysis of variance test (ANOVA) in order to assess whether there was a lack of interaction. As has been extensively discussed in the quantitative methodology literature**,** conventional tests of mean differences or associations (e.g., factorial ANOVA *F* test) are not suitable for assessing a lack of relationship (e.g., Rogers, Howard, & Vessey, 1993; Wellek, 2010). More specifically, in the example described above, it is not acceptable to use non-significance of the interaction term from an ANOVA as evidence of no interaction. Our description of the study was not to criticize the authors, but to highlight the issue that applied researchers need an appropriate statistical tool to test for a lack of interaction. The current study will examine the test of negligible interaction due to Cheng and Shao (2007) and propose an alternative bootstrapping approach that extends this work in important directions. However, we must first describe the methodological difficulties with testing a lack of interactions before providing details on the tests.

Traditional Test for the Presence of an Interaction

The standard test for an interaction in a two-way ANOVA has the null hypothesis (H_0) of no interaction and alternative hypothesis (H_1) that there is an interaction. More formally, for discrete variables *J*, with levels *j* = 1, ..., *a*, and *K*, with levels *k* = 1, ..., *b*, the interaction null hypothesis is:

$$H_o\text{: } \beta_{jj'kk'} = 0 \text{ for all } j,j' \text{ and } k,k', j\neq j', k\neq k',$$

where $\beta_{jj'kk'} = \mu_{jk} - \mu_{jk'} - \mu_{j'k} + \mu_{j'k'}$ and represents all 2 x 2 interactions subsumed within the omnibus *a x b* interaction.

When the research hypothesis relates to detecting an interaction, the objective is to reject the null hypothesis of no interaction. However, when attempting to establish no interaction, the goal is to retain the null hypothesis of no interaction. This goal is problematic because there is a lack of statistical evidence to support any conclusions when the null hypothesis of zero interaction is not rejected (Cheng & Shao, 2007), as "absence of evidence is not evidence of absence" (Altman & Bland, 1995). Further, power (to detect no interaction) would behave in the opposite direction expected, because in order to maximize power for not rejecting the null hypothesis, one would need to minimize, rather than maximize, the sample size. Additionally, it would be unlikely to find population interaction effects exactly equal to zero. Although the interaction effect may be small with no practical significance, it may not meet the strict requirement outlined in the null hypothesis (i.e., nil interaction). Thus, this calls for factorial independent group models to account for the methodological difficulties when attempting to detect a lack of association. More specifically, when the goal is to detect a lack of interaction, the null and alternate hypotheses of the traditional ANOVA must be reversed. An appropriate approach for this scenario is equivalence testing.

Introduction to Equivalence Testing

Equivalence testing is frequently used by biostatisticians who are interested in whether two drugs have an equivalent impact. However, equivalence testing is not restricted to the pharmacological field; it can be a beneficial statistical tool for the evaluation of many psychological hypotheses. For example, researchers may be interested in determining whether two means are equal (e.g., Cribbie, Gruman & Arpin-Cribbie, 2004; Rogers et al., 1993) or whether there is lack of association between variables (e.g., Goertzen & Cribbie, 2010; Mascha & Sessler, 2011; Robinson, Duursma & Marshall, 2005). Establishing a complete lack of association, however, is an unrealistic statistical goal because variables are rarely completely unrelated. With equivalence testing, the goal is not to demonstrate no relationship among variables, but instead that any relationship among the variables is too small to be of any practical value. Thus, if a researcher aims to deem two population means "equivalent", an a priori decision must be made regarding the smallest mean difference that is practically meaningful. This difference (ε) can be expressed as an equivalence interval (i.e., EI = $\{- \varepsilon, \varepsilon \}$). Specifically, equivalence tests measure whether the effect under study falls within the pre-specified interval (e.g., $-\varepsilon < \mu_1 - \mu_2 < \varepsilon$).

One challenge of equivalence testing in psychology is determining an EI due to the paucity of established guidelines for selecting an appropriate interval. EIs are determined based on the researchers' knowledge of their field, expertise with the constructs and samples under study, and understanding of how "meaningful" might be quantified for their research question. For example, one researcher might decide that that one standard deviation quantifies a meaningful difference, whereas another might choose 10% of the standard deviation as a meaningful difference. The key is that researchers must be able to justify their choice of an EI within the specific study and research field.

Cheng and Shao Test for Negligible Interaction

Cheng and Shao's (2007) negligible interaction test (CS) was stimulated by testing treatment-by-centre interactions in multicentre clinical trials using factorial ANOVAs. In this context, treatment-by-centre interaction represents heterogeneous treatment effects across centres. When the treatment-by-centre interaction is close to zero, the overall treatment effect can be analyzed using the average of the treatment effects across all the centres. If the treatment-by-centre interaction is large, researchers are not able to combine the centres' effects. Therefore, Cheng and Shao's (2007) objective was to develop tests for negligible interaction in two-way models.

CS tests the null hypothesis H_0: $\delta \geq \delta_0$ against the alternate hypothesis $H_{1:}$ $\delta < \delta_0$, where δ is the quantitative measure of interaction and δ_o represents a tolerance margin for the interaction. More specifically, an interaction effect would be considered negligible when $\delta < \delta_o$ (see Cheng and Shao, 2007, for the specification of δ in balanced and unbalanced models). The CS approach compares null and alternative hypotheses, following the two competing hypothesis approach of Neyman and Pearson. Hence, a priori statistical power can be computed similarly to traditional null hypothesis based tests. As stated above, the use of equivalence-based tests relies on an appropriate EI. δ_o is used to represent the largest effect for the interaction considered inconsequential. Note that δ_o represents the tolerance margin for interaction in standardized units; therefore it makes sense to select a value for δ_o that represents a small number of average standard deviations. Cohen (1988) has suggested that $f^2 = \delta_o = .32$ as a small effect and Cheng and Shao use $\delta_o = .25$ and .49 in their examples. However, it is always best to select an appropriate δ_o based on the nature of each individual study.

We reject H_0: $\delta \geq \delta_o$ if $F_{ab} < F_{crit}$, where F_{crit} is the critical value for the chosen α in the noncentral F distribution $F_{\propto,\, df_{ab},\, df_e, N\delta_0}$, and $N\delta_0$is the noncentrality parameter.

Bootstrap Approach

The CS approach fills an important void in the literature, albeit with shortcomings. This paper will extend the CS approach by proposing an approach that: 1) demonstrates coherence between the omnibus test and the encompassed interaction contrasts; 2) utilizes an unstandardized EI; and 3) is insensitive to violations of the normality and/or homogeneity of variance assumptions.

The first extension is toward an approach that results in a coherent decision between the omnibus lack of interaction test and the $S = [a(a-1)/2][b(b-1)/2]$ interaction contrasts. The CS approach would be useful, as highlighted by an anonymous reviewer, if a researcher was interested in demonstrating a ‘pattern’ of lack of interaction in a larger factorial design; in other words, the omnibus interaction might indicate a lack of interaction, even if there are interaction contrasts that indicate the presence of an interaction. However, caution must be taken in this situation since concluding a lack of interaction when large interaction contrasts are present may be misleading. Take, for example, the 2 x 4 population mean pattern $\mu_{11} = 11$; $\mu_{12} = 0$; $\mu_{13} = 0$; $\mu_{14} = 0$; $\mu_{21} = 0$; $\mu_{22} = 0$; $\mu_{23} = 0$; $\mu_{24} = 11$ with $\delta_o = .25$ and all $\sigma_{jk} = 10$. This is a power condition for the CS because the population interaction effect δ (.125) < δ_o (.25). The coherence issue is that, for example, the 2 x 4 design contains an interaction contrast with a population mean pattern of $\{\mu_{11} = 11; \mu_{14} = 0; \mu_{21} = 0; \mu_{24} = 11\}$ with $\delta = .303$ (a Type I error condition since $\delta > \delta_o$). For example, if we simulate data from these population conditions with $n_{jk} = 40$ and $\alpha = .05$, we find that the omnibus CS test rejects the null hypothesis approximately 60% of the time, whereas all interaction contrasts reject the null about 1% of the time (the interaction contrast discussed above should only reject the null less than 5% of the time). A coherent, but consequently more conservative, approach uses an intersection-union test (Berger, 1982), where the significance of the omnibus test is implied if all lower order effects (in this case the interaction contrasts) are significant. Berger demonstrated that Type I error rates for the omnibus test cannot exceed α if the significance of the omnibus test is implied by the significance of each of the α-level lower-order effects.

The second extension of the CS approach is utilizing an unstandardized EI. In defining an EI, the goal is to establish the smallest effect of practical importance. While using standardized effect sizes can simplify defining the EI (e.g., Cheng & Shao, 2007, utilize liberal and conservative cutoffs for the CS of $\delta_o = .49$ and $\delta_o = .25$), it is important for researchers to consider, precisely, the smallest meaningful difference for their study. Further, an unstandardized effect size is preferred when the units are meaningful (Wilkinson et al, 1999). In factorial ANOVA designs, the units are the differences in the row means across the columns (or vice versa). These units are interpretable as long as the researcher can quantify a meaningful difference in the cell means. In other words, if a researcher can quantify a meaningful difference in raw scale means then they almost certainly they can quantify a minimally important interaction effect. Imagine, for example, that a researcher proposes the following population mean configuration, $\mu_{11} = 5$; $\mu_{12} = 0$; $\mu_{21} = 0$; $\mu_{22} = 0$, as the smallest interaction effect that would still be meaningful. The half-width of the equivalence interval for the interaction effect is thus set at 5 [$(\mu_{11} - \mu_{12}) - (\mu_{21} - \mu_{22})$]. Further, the meaningful difference is always expressed for the simplest interaction effect, the 2 x 2, and thus the task is made even simpler.

The final proposed extension of the CS test is toward a method that is robust to nonnormality and/or heteroscedasticity. It is well known that distributions in psychology rarely match the characteristics of the normal distribution with equal population variances (e.g., Golinski & Cribbie, 2009; Keselman et al., 1998; Micceri, 1989; Wilcox, 2012). When these assumptions are violated, the Type I error rates and power of procedures that rely on these assumptions (such as the CS test) can be severely affected (e.g., Cribbie, Fiksenbaum, Wilcox & Keselman, 2012; Cribbie, Wilcox, Bewell & Keselman, 2007; Mills, Cribbie, & Luh, 2009). The approach adopted here is bootstrapping (case resampling) which has been applied to many problems where the underlying sampling distribution of a statistic is unknown, which applies when the normality and/or variance homogeneity assumptions are violated.

The next two sections describe the unstandardized intersection-union and bootstrap-based test of negligible interaction (IU-B) for 2 x 2 and general *a x b* designs.

2 x 2 Design. The simplest case for the IU-B is a 2 x 2 independent groups factorial design,. Since only one interaction contrast is contained within the omnibus interaction, the null hypothesis for the omnibus test is rejected if the only interaction contrast null hypothesis is rejected. Following the logic of the interval approach (Westlake, 1976) and the two one-sided approach (Schuirmann, 1987), the null hypotheses are:

$$H_{o1}: \beta \geq \gamma$$

$$H_{o2}: \beta \leq -\gamma$$

where β represents the unstandardized population interaction effect (μ_{11} - μ_{12} - μ_{21} + μ_{22}) and γ represents the smallest meaningful unstandardized interaction effect [i.e., the EI is ($-\gamma$, γ)]. The alternate hypothesis can be more simply represented by:

$$H_{a:} -\gamma < \beta < \gamma$$

Note that the unstandardized interaction effect may be equivalently represented in other ways, e.g., (μ_{11} - μ_{21}) - (μ_{12} - μ_{22}) or (μ_{11} - μ_{12}) - (μ_{21} - μ_{22}). To avoid the usual ANOVA assumptions for our proposed test, we bootstrap from the observed observations in order to generate an empirical distribution function instead of using a theoretical sampling distribution (e.g., *F* distribution). The empirical distribution function for the statistic of interest, in this case the interaction effect, considers the nonnormality and/or heteroscedasticity of the data.

Consider an *N* x 3 dataset with one column for the first predictor, second predictor and outcome, respectively. Resample with replacement *N* rows of the dataset *b* = 1, ..., *B* times, and for each bootstrap sample compute:

$$I_b = M_{11} - M_{12} - M_{21} + M_{22}.$$

The distribution of the *B* I_b statistics is I. Then, reject H_{01} and H_{02} if the interval (I_α, $I_{1-\alpha}$) falls completely within the unstandardized equivalence interval (-γ, γ), where I_α and $I_{1-\alpha}$ represent the α and 1-α quantiles from distribution I.

***a x b* Design**. Following the logic of intersection-union tests, the null hypothesis for the omnibus *a* x *b* interaction effect is rejected only if each of the null hypotheses for each of the interaction contrasts subsumed within the omnibus test are rejected. First, the composite omnibus null hypothesis is:

$$H_o: \beta_{jj'kk'} \leq -\gamma \mid \beta_{jj'kk'} \geq \gamma \text{ for any } j,j' \text{ and } k,k', j\neq j', k\neq k'$$

where $\beta_{jj'kk'} = \mu_{jk} - \mu_{jk'} - \mu_{j'k} + \mu_{j'k'}$ and represents all possible interaction contrasts encompassed within the omnibus interaction effect. The alternate hypothesis is:

$$H_a: -\gamma < \beta_{jj'kk'} < \gamma \text{ for all } j,j' \text{ and } k,k', j\neq j', k\neq k'.$$

In words, if the null hypothesis associated with *any* of the 2 x 2 interaction contrasts is not rejected then the omnibus null hypothesis of a meaningful interaction is also not rejected. If the null hypothesis associated with *all* of the 2 x 2 interaction contrasts is rejected then the omnibus H_o is rejected, so one concludes a negligible interaction is present. The interaction contrasts are each tested using the unstandardized bootstrap-based procedure described above for the 2 x 2 design at level α.

A multiplicity issue arises when selecting a significant effect out of a pool of effects that may also include null effects; more specifically, the probability of a Type I error increases when multiple tests of significance are conducted and more than one test may include null effects. This occurs when the omnibus null hypothesis is not rejected for the negligible interaction test, but the researcher tests whether any of the S contained 2 x 2 interactions are negligible ($S = [a(a-1)/2][b(b-1)/2]$). The naive approach would be to use an adjusted α level of α / S, however, as in the multiplicity problem for one-way independent groups equivalence tests (see Lauzon, Caffo, & Rohmel, 2013), only a subset of the S interaction contrasts can be detrimental to the familywise Type I error rate. Following the logic of Rohmel (2011) and Lauzon & Caffo (2009), an effect is detrimental to the familywise Type I error rate if $\gamma \leq |\beta_{jj'kk'}| < 2\gamma$. Here, if the population interaction effect falls at or slightly greater than γ, it could result in a Type I error but $|\beta_{jj'kk'}| \geq 2\gamma$ are unlikely to contribute to inflation of the Type I error rate given their magnitude. In order to control the familywise error rate at α, one must determine the maximum number of detrimental 2 x 2 null interaction effects (ζ) of the S 2 x 2 interactions encompassed by the larger a x b interaction.

It is obvious for an omnibus 2 x 2 interaction that $S = \zeta = 1$ (e.g., $\mu_{11} =1$; $\mu_{12} = 0$; $\mu_{21} = 0$; $\mu_{22} = 0$ with $\gamma = 1$) and thus no multiplicity control is necessary. Following Rohmel's (2011) logic, the worst-case scenario in greater than 2 x 2 designs does not occur when all effects are null, but when the greatest number of interaction effects fall between γ and 2γ. Take, for example, a 4 x 2 design ($S = 6$). Although it might be tempting to consider [$\mu_{11} = 0$; $\mu_{12} = 1$; $\mu_{13} = 2$; $\mu_{14} =3$; $\mu_{21} = 0$; $\mu_{22} = 0$; $\mu_{23} = 0$; $\mu_{24} = 0$ with $\gamma = 1$] a worst case scenario since all interaction effects are null, ζ is only equal to three (comparing columns one and two, two and three, and three and four). In contrast, the patterns [$\mu_{11} = 1$; $\mu_{12} = 0$; $\mu_{13} = 0$; $\mu_{14} = 0$; $\mu_{21} = 0$; $\mu_{22} = 0$; $\mu_{23} = 0$; $\mu_{24} = 1$ with $\gamma = 1$], [$\mu_{11} = .5$; $\mu_{12} = 0$; $\mu_{13} = .5$; $\mu_{14} = 0$; $\mu_{21} = 0$; $\mu_{22} = .5$; $\mu_{23} = 0$; $\mu_{24} = .5$ with $\gamma = 1$], each have $\zeta = 4$ (in the first set comparing columns one and two, one and three, two and four, and three and four; in the second set comparing columns one and two, one and three, two and three, and three and four).

It can be shown that $\zeta = 2$ for 2 x 3 designs ($S = 3$), $\zeta = 4$ for 2 x 4 designs ($S = 6$), $\zeta = 6$ for 3 x 3 designs ($S = 9$), and $\zeta = 12$ for 3 x 4 designs ($S = 18$), etc. Generally, two-thirds of the total number of interaction contrasts can be worst-case scenarios (i.e., $\zeta = 2S / 3$). We verified this result through simulation by randomly sampling population cell means and computing the maximum number of detrimental Type I error cases for each design. Dividing the familywise Type I error rate (α) by $2S / 3$ (i.e., α / ζ) should provide strong familywise Type I error control.

Our primary goal is to demonstrate the analytical differences between the CS test and our proposed bootstrap intersection-union test of negligible interaction (IU-B). We also compare both omnibus and interaction contrast Type I error control and power of an intersection-union version of the Cheng and Shao test (CS-IU) to the IU-B test.

Method

A simulation study was conducted using *R* (*R* Core Team, 2014) to investigate the omnibus and familywise Type I error control and omnibus and per-comparison power rates of the CS and the proposed IU-B procedures. An intersection-union version of the CS test (CS-IU) was also evaluated by using the CS test on all *S* 2 x 2 interaction contrasts and rejecting the omnibus test if all tests are statistically significant.

A two independent-groups factorial ANOVA design was used in this study. Several variables were manipulated including: a) design (2 x 2; 4 x 2); b) distribution shape (normal, positively skewed); c) balanced or unbalanced cell size (average cell size was set at $n = 40$); d) equal or unequal population cell standard deviations (average $\sigma = 10$); e) combination of unequal cell sizes and unequal populations standard deviations (direct; inverse); and f) population cell mean pattern.

For the positively skewed distribution the outcome was χ^2distributed with three degrees of freedom (skewness = 1.59, kurtosis = 7). Several Type I error and power population mean configurations were used. For the Type I error condition, we deemed $\pm .2\alpha$ an acceptable range of deviation from the nominal rate (these bounds fall between the liberal and conservative bounds proposed by Bradley, 1978). For all Type I error conditions, the omnibus test interaction effect for the CS test (δ) was equal to δ_o. For the IU-B test in the 2 x 2 design, the interaction effect ($\beta_{jj'kk'}$) was equal to γ. The omnibus IU-B and CS-IU tests in the 4 x 2 design are expected to be conservative since multiple interaction contrasts can be null and/or power can be less than perfect (with the degree of conservativeness depending on how many of the *S* interaction contrasts are Type I error conditions and how far the means are from the border of the EI). For example, with the mean pattern μ_{11}, μ_{12}, ..., μ_{42} = 23.094, 0, ..., 0, there are 4 interaction contrasts that fall at the edge of the EI, with the remaining contrasts being power conditions. Thus, even if power for detecting the non-null interaction contrasts was at 100%, the probability that all interaction contrasts are rejected would be low since four of them would have approximately a nominal α-level rejection rate. When both cell sizes and population standard deviations were unequal, they were either directly paired (largest size with largest standard deviation) or inversely paired (largest size with smallest standard deviation). The specific conditions can be found in Tables 1, 2, 4 and 5.

Five thousand simulations were conducted for each condition using a nominal significance level of .05 and a tolerance margin for interaction of δ_o= .25. With α = .05, the acceptable range of deviation from the nominal Type I error rate was set at .04 - .06 [.05 +/- .2(.05)].With 5000 simulations the Type I error results have a standard error of approximately .003. For the IU-B, γ was set at the largest 2 x 2 population interaction contrast effect. For example, with the mean configuration μ_{11}, μ_{12}, ..., μ_{42} = 14.142,0,...,0,14.142, γ was set at 28.184 since the largest 2 x 2 interaction contrast effect occurs when the mean pattern is μ_{11}, μ_{12}, μ_{41}, μ_{42} = 14.142,0,0,14.142. For the power conditions, five points was subtracted from the unstandardized interaction effect with the population mean pattern left unchanged.

Results

Type I Error Rates

2 x 2 Design. Type I error rates are presented in Table 1. When all of the assumptions were met, both the CS and IU-B procedures maintained rates within the acceptable bounds. However, when the normality and/or variance homogeneity assumptions were not met, only the IU-B was able to maintain the rates within the acceptable range over all conditions. Type I error rates for the CS were severely inflated in some conditions, with a maximum rate of 12.6% with a skewed distribution and directly paired cell sizes and variances.

4 x 2 Design. Omnibus Type I error rates and familywise Type I error rates are presented in Tables 2 and 3, respectively. Like in the 2 x 2 design, when all of the assumptions were met the CS procedure maintained rates within the acceptable bounds, however when the assumptions were violated, the rates regularly fell outside of the acceptable bounds. As expected, the omnibus rates for the IU-B depended on the population mean configuration with rates falling at the nominal rate when only one interaction contrast was at the bound of the equivalence interval (e.g., μ_{11}, μ_{12}, ..., μ_{42} = 14.142,0,...,0,14.142) and were conservative otherwise. Rates for the CS-IU were very conservative. This is expected because, as discussed earlier, δ_o is set for the full 4 x 2 mean configuration, but the effect size for one or more of the contained 2 x 2 interaction contrasts could exceed δ_o. For example, the effect size for the μ_{11}, μ_{12}, μ_{41} μ_{42} = 14.142, 0, 0, 14.142 interaction contrast is much larger than δ_o for the full 4 x 2 mean configuration so this contrast would rarely, if ever, be rejected.

Familywise error rates for both the CS and IU-B procedures were determined using an interaction contrast Type I error rate of $\alpha/\ \zeta$. Familywise error rates for the CS procedure, in addition to being overly conservative in many conditions, were again affected nonnormality and/or variance heterogeneity, with rates exceeding the upper bounds in some conditions. Rates for the IU-B never exceeded the upper bound (.06), with rates again depending on the configuration of the means. More specifically, for mean configurations with a high number of detrimental Type I error conditions (e.g., μ_{11}, μ_{12}, ..., μ_{42} = 10, 0, 10, 0, 0, 10, 0, 10 with $\zeta = 4$) the familywise rates were closer to α, whereas for mean configurations with a low number of detrimental Type I error conditions (e.g., μ_{11}, μ_{12},..., μ_{42} = 14.142,0 , ..., 0, 14.142 with $\zeta = 1$) the familywise rates were conservative (as the interaction contrast Type I error rate was α divided by the maximum ζ for that design).

Power

2 x 2 Design. Power rates are presented in Table 4. In conditions in which the Type I error rates were controlled for the CS procedure, the power rates for the IU-B were always higher than those for the CS. When the Type I error rates were not maintained within the acceptable bounds for the CS test one should not conduct power comparisons.

4 x 2 Design. Omnibus and per-contrast power rates are presented in Tables 5 and 6, respectively. In general *a x b* designs, omnibus power comparisons between the CS procedure and IU types tests are difficult to make given the theoretical differences between the procedures. In other words, although the power rates for the CS procedure are generally higher than those for the IU-based procedures, the CS procedure, even in power conditions, can contain interaction contrasts with effects much greater than δ_o.

When an IU-based version of the CS is compared to the IU-B, the IU-B is generally more powerful since the presence of interaction contrasts greater than δ_o (which is, in fact, not a power condition) reduces the power of the CS-IU procedure. On the other hand, the interaction contrasts for the IU-B cannot exceed γ and thus they can have relatively greater power. The degree of conservativeness of the IU-B relative to the CS depends on how far each of the interaction contrast effects are from γ. For example, with μ_{11}, μ_{12}, ..., μ_{42} = 11.642, 0, ..., 0, 11.642, only one $\beta_{jj'kk'}$ is close to γ and the rates for the IU-B are more powerful than the CS.

When familywise error control is imposed for post hoc testing of the *S* interaction contrasts, the average per-pair power rates of the CS and IU-B tests are of interest. The IU-B was always more powerful than the CS, except for cases in which the CS procedure had inflated Type I error rates due to its lack of robustness to variance heterogeneity (more specifically, due to inflated Type I error rates for the CS procedure when unequal samples sizes and standard deviations are directly paired).

Discussion

We compared two methods for detecting negligible interactions, Cheng and Shao’s noncentrality-based procedure and our proposed bootstrap-based intersection-union procedure. There are numerous instances where researchers are interested in detecting negligible interaction and thus the availability of an appropriate test is extremely valuable. Assessing a lack of interaction using non-rejection of the null hypothesis of no interaction in a traditional ANOVA is not acceptable

The IU-B procedure was proposed to improve on some of the characteristics of the CS procedure. More specifically, the IU-B procedure provides coherent decisions between the omnibus test and the contained interaction contrasts, utilizes an unstandardized equivalence interval, and is insensitive to assumption violations. The first two potential advantages are theoretical, whereas the third was verified through a simulation study.

In 2 x 2 independent groups factorial designs, the procedures are most comparable because only a single test is conducted with each method (although the CS uses a standardized EI whereas the IU-B uses an unstandardized EI). The proposed IU-B procedure provided excellent Type I error control (even using the conservative $\alpha \pm .2\alpha$ criteria), whereas the CS procedure was not able to control the Type I error rates within these bounds when the normality and/or homogeneity of variance assumptions were violated. Further, even when the assumptions were satisfied, the IU-B was always more powerful than the CS.

In larger *a x b* designs, the procedures differ theoretically and therefore direct comparisons are difficult. However, it remained clear that the omnibus CS test is greatly affected by nonnormality and variance heterogeneity as the probability of a Type I error approached 4α under these conditions. Since the IU-B procedure is an intersection-union test, the omnibus Type I error rates depend on the population mean configuration. Specifically, the overall Type I error rates approach the nominal rate when there is only a single detrimental Type I error condition, whereas the test is more conservative for $\zeta > 1$. Further, as expected, the omnibus Type I error rate never exceeds the nominal rate. When interest is in the *S* interaction contrasts, again the CS test is affected by violations of the assumptions and rates exceeded the upper acceptable limits in a few conditions. For the IU-B, the familywise error rates depended on the population mean configuration; however, in this situation, the rates approach the nominal rate when ζ is large (since each interaction contrast is being evaluated against a nominal significance level of α/ ζ) and are more conservative as ζ decreases. Comparisons of the CS-IU to the IU-B, either for omnibus testing or for testing all interaction contrasts, revealed that the IU-B was more powerful, although again we must use caution when interpreting these results since the nature of the equivalence intervals differ.

For this simulation study, we used two factorial independent groups models (2 x 2; 4 x 2) and manipulated specific variables (e.g., sample size equality/inequality, population standard deviation equality/inequality, population mean configuration, population distribution shape). Therefore, a potential limitation of the study is that the results obtained are restricted to the conditions investigated. For example, only one type of nonnormal distribution was investigated and thus the pattern of results might vary over different distribution shapes. We also did not explore within-subject designs, hierarchical models or more complex factorial models, which represent important extensions of this work.

To summarize, this study proposed a novel intersection-union and bootstrap-based test of negligible interaction. It extends the negligible interaction test proposed by Cheng and Shao (2007) by devising a method that results in coherent decisions between the omnibus test and the contained interaction contrasts, utilizes an unstandardized equivalence interval, and is insensitive to violations of the normality and variance homogeneity assumptions. If a standardized equivalence interval is preferred, all assumptions are met and incoherent decisions between the omnibus test and interaction contrasts are tolerated, the Cheng and Shao test can be adopted. However, given the advantages of the proposed procedure discussed above, we recommend that researchers conduct tests of negligible interaction using the intersection-union and bootstrap-based procedure. To facilitate the use of both the CS and IU-B tests, researchers may obtain a function for use with the free, open source software R (R Core Team, 2014) from http://cribbie.info.yorku.ca/r-functions/.

References

Altman, D. G., & Bland, J. M. (1995). Absence of evidence is not evidence of absence. *British Medical Journal, 311*(7003), 485. DOI:10.1111/j.1751-0813.1996.tb13786.x

Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics, 24,* 295–300. DOI:10.2307/1267823

Bradley, J.V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144-152. DOI: 10.1111/j.2044-8317.1978.tb00581.x

Caffo, B., Lauzon, C., & Rohmel, J. (2013). Correction to "Easy Multiplicity Control in Equivalence Testing Using Two One-Sided Tests", *The American Statistician, 67,* 115-116. DOI: 10.1080/00031305.2012.760487

Cheng, B., & Shao, J. (2007). Exact tests for negligible interaction in two-way analysis of variance/covariance. *Statistica Sinica*, *17*, 1441-1455.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd Ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cribbie, R. A., Fiksenbaum, L., Wilcox, R. R. & Keselman, H. J. (2012). Effects of nonnormality on test statistics for one-way independent groups designs. *British Journal of Mathematical and Statistical Psychology*, *65*, 56-73. DOI:10.1111/j.2044-8317.2011.02014.x

Cribbie, R. A., Gruman, J., & Arpin-Cribbie, C. (2004). Recommendations for applying tests of equivalence. *Journal of Clinical Psychology, 60*(1), 1-10. DOI:10.1002/jclp.10217

Cribbie, R. A., Wilcox, R., Bewell, C. & Keselman, H. J. (2007). Tests for treatment group equality when data are nonnormal and heteroscedastic. *Journal of Modern Applied Statistical Methods, 6*, 117-132.

Goertzen, J. R., & Cribbie, R. A. (2010). Detecting a lack of association: An equivalence testing approach. *British Journal of Mathematical and Statistical Psychology, 63*(3), 527-537. DOI:10.1348/000711009X475853

Golinski, C. & Cribbie, R. A. (2009). The expanding role of quantitative methodologists in advancing psychology. *Canadian Psychology, 50,* 83-90. DOI:10.1037/a0015180

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology, 68,* 155-165. DOI:10.1037/0022-006X.68.1.155

Johansen, S. (1980). The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression. *Biometrika, 67,* 85-92. DOI: 10.1093/biomet/67.1.85

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. DOI: 10.3102/00346543068003350

LaRose, J. G., Gorin, A. A., & Wing, R. R. (2009). Behavioral self-regulation for weight loss in young adults: a randomized controlled trial. *International Journal of Behavioral Nutrition and Physical Activity, 6*(10). DOI:10.1186/1479-5868-6-10

Lauzon, C., & Caffo, B. (2009). Easy multiplicity control in equivalence testing using two one-sided tests. *The American Statistician, 63*, 147-154. DOI: 10.1198/ tast.2009.0029

Mascha, E. J., & Sessler, D. I. (2011). Equivalence and noninferiority testing in regression models and repeated-measures designs. *Anesthesia & Analgesia*, *112*, 678-687. DOI: 10.1213/ANE.0b013e318206f872

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156-166. DOI:10.1037/0033-2909.105.1.156

Mills, L., Cribbie, R. A., & Luh, W.-M. (2009). A heteroscedastic, rank-based approach for analyzing 2 x 2 independent groups designs. *Journal of Modern Applied Statistical Methods, 8*, 322-336.

Paolieri, D., Lotto, L., Leoncini, D., Cubelli, R., & Job, R. (2011). Differential effects of grammatical gender and gender inflection in bare noun production. *British Journal of Psychology*, *102*(1), 19-36. DOI: 10.1348/000712610X496536

R Core Team (2014). *R*: A language and environment for statistical computing. *R* Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

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

DOI: 10.1093/treephys/25.7.903

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.

DOI: 10.1037//0033-2909.113.3.553

Röhmel, J. (2011). On familywise type I error control for multiplicity in equivalence trials with three or more treatments. *Biometrical Journal*, *53*, 914-926. DOI: 10.1002/bimj.201100073

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.

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority*. CRC Press.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing, 3rd Ed.* San Diego, CA: Academic Press.

Wilkinson, L., & The APA Task Force on Statistical Inference (1999). Statistical methods in psychology journals guidelines and explanations. *American Psychologist, 54*, 594–604.

Table 1

Type I Error Rates for the 2 x 2 Design

$n_{11}, n_{12}, n_{21}, n_{22}$	$\sigma_{11}, \sigma_{12}, \sigma_{21}, \sigma_{22}$	$\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}$	Normal Distribution		Chi-Square Distribution	
			CS	IU-B	CS	IU-B
40,40,40,40	10,10,10,10	20,0,0,0	.048	.048	**.069**	.056
		10,0,0,10	.053	.051	.060	.049
	5,15,5,15	22.361,0,0,0	.049	.055	**.070**	.055
		11.181,0,0,11.181	.050	**.062**	**.074**	.056
30,30,50,50	10,10,10,10	20,0,0,0	**.061**	.043	**.088**	.060
		10.328,0,0,10.328	.055	.059	**.062**	.045
	5,5,15,15	23.094,0,0,0	**.087**	.047	**.115**	.056
		11.547,0,0,11.547	**.109**	.059	**.126**	.052
	15,15,5,5	23.094,0,0,0	**.022**	.050	**.038**	.052
		11.547,0,0,11.547	**.020**	.048	.042	.052

Note: CS = Cheng & Shao (2007) test; IU-B = Intersection-Union and Bootstrap-Based test; bolded value = Type I error rate did not fall within the acceptable range of .04-.06.

Table 2

Type I Error Rates for the 4 x 2 Design

n_{11},n_{12},...,n_{42}	σ_{11}, σ_{12},...,σ_{42}	μ_{11}, μ_{12},..., μ_{42}						
			Normal Distribution			Chi-Square Distribution		
			CS	CS-IU	IU-B	CS	CS-IU	IU-B
40,...,40	10,..,10	14.142,0,...,0,14.142	.048	.000	.054	.066	.000	.059
		10,0,10,0,0,10,0,10	.050	.002	.002	.068	.002	.002
		23.094,0,...,0	.046	.000	.003	.060	.000	.003
	5,5,5,5,15,15,15,15	15.812,0,...,0,15.812	.050	.000	.049	.072	.000	.049
		11.181,0,11.181,0,0,11.181,0,11.181	.056	.002	.001	.077	.004	.002
		25.819,0,...,0	.052	.000	.005	.076	.002	.003
30,30,30,30,50,50,50,50	10,...,10	14.606,0,...,0,14.606	.052	.000	.054	.060	.000	.054
		10.328,0,10.328,0,0,10.328,0,10.328	.046	.002	.002	.065	.002	.002
		23.851,0,...,0	.045	.000	.005	.068	.000	.008
	5,5,5,5,15,15,15,15	16.330,0,...,0,16.330	.157	.000	.048	.172	.000	.040
		11.547,0,11.547,0,0,11.547,0,11.547	.159	.005	.002	.186	.000	.007
		26.667,0,...,0	.172	.000	.003	.189	.006	.003
	15,15,15,15,5,5,5,5	16.330,0,...,0,16.330	.017	.000	.056	.028	.000	.052
		11.547,0,11.547,0,0,11.547,0,11.547	.013	.000	.001	.026	.002	.004
		26.667,0,...,0	.011	.000	.003	.019	.000	.006

Note: CS = Cheng & Shao (2007) test; CS-IU = ; IU-B = Intersection-Union and Bootstrap-Based test

Table 3

Familywise Error Rates for the 4 x 2 Design

n_{11},n_{12},...,n_{42}	σ_{11}, σ_{12},...,σ_{42}	μ_{11}, μ_{12},..., μ_{42}				
			Normal Distribution		Chi-Square Distribution	
			CS	IU-B	CS	IU-B
40,...,40	10,..,10	14.142,0,...,0,14.142	.000	.012	.000	.018
		10,0,10,0,0,10,0,10	.043	.044	.070	.050
		23.094,0,...,0	.003	.035	.006	.033
	5,5,5,5,15,15,15,15	15.812,0,...,0,15.812	.000	.018	.000	.014
		11.181,0,11.181,0,0,11.181,0,11.181	.044	.049	.066	.052
		25.819,0,...,0	.002	.038	.008	.042
30,30,30,30,50,50,50,50	10,...,10	14.606,0,...,0,14.606	.000	.012	.000	.022
		10.328,0,10.328,0,0,10.328,0,10.328	.044	.049	.072	.056
		23.851,0,...,0	.003	.036	.007	.040
	5,5,5,5,15,15,15,15	16.330,0,...,0,16.330	.000	.018	.000	.018
		11.547,0,11.547,0,0,11.547,0,11.547	.071	.048	.106	.053
		26.667,0,...,0	.003	.036	.009	.042
	15,15,15,15,5,5,5,5	16.330,0,...,0,16.330	.000	.018	.000	.014
		11.547,0,11.547,0,0,11.547,0,11.547	.026	.054	.051	.057
		26.667,0,...,0	.001	.033	.006	.034

Note: CS = Cheng & Shao (2007) test; IU-B = Intersection-Union and Bootstrap-Based test

Table 4

Power Rates for the 2 x 2 Design

n_{11},n_{12},n_{21},n_{22}	σ_{11},σ_{12},σ_{21},σ_{22}	μ_{11}, μ_{12}, μ_{21}, μ_{22}				
			Normal Distribution		Chi-Square Distribution	
			CS	IU-B	CS	IU-B
40,40,40,40	10,10,10,10	15,0,0,0	.453	.481	**.451**	.493
		7.5,0,0,7.5	.473	.488	.433	.470
	5,15,5,15	17.361,0,0,0	.404	.423	**.387**	.435
		8.681,0,0,8.681	.383	**.409**	**.380**	.424
30,30,50,50	10,10,10,10	15,0,0,0	**.502**	.455	**.502**	.470
		7.828,0,0,7.828	.418	.450	**.423**	.460
	5,5,15,15	18.094,0,0,0	**.546**	.458	**.526**	.478
		9.047,0,0,9.047	**.564**	.484	**.522**	.479
	15,15,5,5	18.094,0,0,0	**.228**	.359	**.240**	.372
		9.047,0,0,9.047	**.227**	.348	.223	.355

Note: CS = Cheng & Shao (2007) test; IU-B = Intersection-Union and Bootstrap-Based test; bolded value = Type I error rate did not fall within the acceptable range of .04-.06.

Table 5

Omnibus Power Rates for the 4 x 2 Design

$n_{11},n_{12},\ldots,n_{42}$	$\sigma_{11}, \sigma_{12},\ldots,\sigma_{42}$	$\mu_{11}, \mu_{12},\ldots, \mu_{42}$						
			Normal Distribution			Chi-Square Distribution		
			CS[1]	CS-IU	IU-B	CS[1]	CS-IU	IU-B
40,...,40	10,..,10	11.642,0,...,0,11.642	.447	.003	.483	.438	.008	.492
		7.5,0,7.5,0,0,7.5,0,7.5	.691	.138	.158	.680	.119	.150
		18.094,0,...,0	.597	.029	.238	.567	.042	.241
	5,5,5,5,15,15,15,15	13.312,0,...,0,13.312	.387	.003	.444	.376	.011	.420
		8.681,0,8.681,0,0,8.681,0,8.681	.597	.095	.124	.576	.090	.115
		20.819,0,...,0	.494	.019	.178	.471	.060	.158
30,30,30,30,50,50,50,50	10,...,10	12.106,0,...,0,12.106	.416	.005	.446	.420	.012	.478
		7.828,0,7.828,0,0,7.828,0,7.828	.674	.123	.135	.638	.111	.145
		18.851,0,0,0,0,0,0,0	.554	.021	.203	.550	.029	.231
	5,5,5,5,15,15,15,15	13.830,0,...,0,13.830	.646	.005	.452	.636	.018	.477
		9.047,0,9.047,0,0,9.047,0,9.047	.830	.208	.135	.801	.187	.143
		21.667,0,0,0,0,0,0,0	.766	.053	.220	.724	.107	.194
	15,15,15,15,5,5,5,5	13.830,0,...,0,13.830	.151	.000	.345	.156	.003	.356
		9.047,0,9.047,0,0,9.047,0,9.047	.288	.028	.076	.307	.029	.083
		21.667,0,0,0,0,0,0,0	.214	.004	.134	.229	.000	.170

Note: CS = Cheng & Shao (2007) test; CS-IU = Intersection-Union version of the Cheng & Shao (2007) test; IU-B = Intersection-Union and Bootstrap-Based test. [1] Caution should be taken when comparing these power results to the IU-based procedures since the mean configurations for the CS can contain 2x2 interactions larger than δ_o

Table 6

Per-Contrast Power Rates for the 4 x 2 Design

n_{11},n_{12},...,n_{42}	σ_{11}, σ_{12},...,σ_{42}	μ_{11}, μ_{12},..., μ_{42}				
			Normal Distribution		Chi-Square Distribution	
			CS	IU-B	CS	IU-B
40,...,40	10,..,10	11.642,0,...,0,11.642	.588	.878	.577	.878
		7.5,0,7.5,0,0,7.5,0,7.5	.499	.514	.497	.516
		18.094,0,...,0	.524	.637	.532	.641
	5,5,5,5,15,15,15,15	13.312,0,...,0,13.312	.566	.870	.545	.868
		8.681,0,8.681,0,0,8.681,0,8.681	.463	.478	.470	.488
		20.819,0,...,0	.519	.609	.523	.613
30,30,30,30,50,50,50,50	10,...,10	12.106,0,...,0,12.106	.580	.872	.574	.924
		7.828,0,7.828,0,0,7.828,0,7.828	.489	.499	.485	.508
		18.851,0,0,0,0,0,0,0	.521	.624	.527	.631
	5,5,5,5,15,15,15,15	13.830,0,...,0,13.830	.662	.875	.655	.878
		9.047,0,9.047,0,0,9.047,0,9.047	.525	.496	.524	.509
		21.667,0,0,0,0,0,0,0	.530	.628	.536	.629
	15,15,15,15,5,5,5,5	13.830,0,...,0,13.830	.455	.855	.442	.853
		9.047,0,9.047,0,0,9.047,0,9.047	.407	.446	.414	.461
		21.667,0,0,0,0,0,0,0	.510	.587	.513	.593

Note: CS = Cheng & Shao (2007) test; IU-B = Intersection-Union and Bootstrap-Based test.