CORRELATIONAL EQUIVALENCE TESTING

By

Miriam Kraatz

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Psychology

May, 2007

Nashville, Tennessee

Approved:

Professor James H. Steiger

Professor David Cole

Professor David Lubinski

Professor Andrew J. Tomarken

ACKNOWLEDGEMENTS

I would like to thank my committee members Dr. David Lubinski, Dr. David Cole, Dr. Andrew Tomarken and Dr. James Steiger for their much valued input and advice during the past three years and during the completion of this project. Special thanks go to Dr. James Steiger, who has been a very supportive and patient advisor and who's competence and exactness can only be a distant goal for me and from whom to learn is a great pleasure. Further I would like to thank David Bass for his help and advice in computer related concerns.

Thanks to my friends, especially David, who is my admin, best friend, travel partner and German teacher all in one; to Amanda, Kim, Jon, Mohammad, Dschin-u, and Tobias; the other graduate students, Pong Lai, Rocketown, the German Stammtisch, and last but not least, my family, for making my experience in a country far away from Germany so unique.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Hypothesis testing is one of the main elements of scientific inquiry. While many traditional hypothesis tests are aimed at showing that two parameters from two populations are different from each other, showing that two parameters are the same has increasingly received attention. For example, a researcher might want to show that IQ in two clinical populations is the same or that the correlations between the outcomes of two achievement measures are the same for male and female 8th graders. Whether we want to test that two means, variances or covariances are practically equal, the development of adequate methods to do so has only gathered momentum in the past 30 years.

Accept-Support and Reject-Support Testing

Scientific progress in the social sciences as well as many other areas of science is measured by the quality of its theories. Quality of a theory is based on whether it can be falsified, how much it explains compared to other theories and how much it explains overall.

Often a researcher sets one theory against another by comparing "null and alternative" hypotheses that are implied by each theory. The alternative hypothesis is the one implied by the

researcher's preferred theory, and data that speak against the null provide evidence in favor of the preferred theory. The term *Reject-Support* (RS) testing highlights the fact

that the researcher's hypothesis will be supported when he/she is able to reject the null hypothesis. Research hypotheses are often expressed as hypotheses about the value of a parameter (or parameters), the classical example being

$$H_0: \theta = a \text{ and } H_a: \theta \neq a \tag{1}$$

for some parameter θ of the distribution of interest. An equally well known but slightly more complicated example, which is our focus in this study, compares two parameters. We can ask ourselves by how much two parameters differ, i.e. how big is the difference between θ_1 and θ_2. The null and alternative hypotheses can then be stated as follows:

$$H_0 : \theta_1 - \theta_2 = a \tag{2}$$

and

$$H_1 : \theta_1 - \theta_2 \neq a . \tag{3}$$

If we set $a = 0$, we are testing the hypothesis whether θ_1 and θ_2 are equal.

In many areas of application, a statistical null hypothesis that consists of a point value (in this case 0), such as in Equation (2), is virtually certain to be false, if only by a small amount. For example, in the behavioral sciences, most psychological manipulations have *some* effect. Consequently, whenever the null hypothesis consists of a point value, it can always be rejected if the sample size is large and therefore the degree of precision high enough. On the other hand, this means that a point value hypothesis can never be shown to be precisely true. Sometimes, however, the researcher wishes to show that a certain point hypothesis is *effectively true,* in the sense that the true parameter in the population is so close to some point value that the amount by which the null hypothesis is

false has no practical significance. In our case, he or she might want to show that two means are equal, i.e. the difference between them is zero. The hypothesis that the two means are equal is almost certainly false in the strict sense. However, it may well be true in the practical sense that the difference between the two means is *trivial*.

This raises a key question. How can you show that the difference between two means is trivially small? One early attempt at accomplishing this was *Accept-Support* (AS) testing. AS testing keeps the same null and alternative hypothesis as above (Equations (2) and (3)) but reverses the researcher's intentions. Conducting the exact same statistical procedures, the researcher now wishes *not* to reject the null hypothesis. As mentioned before, the probability of not rejecting the null hypothesis and deciding that the two means are equal will decrease and go towards zero when precision increases, no matter how small the difference between the two means. The higher the precision of the test – and with that power – the more likely is the rejection of H_0 and the researcher who does not want to reject will be punished for the high precision gained in the experiment.

Therefore, the best way to guarantee keeping a favored null hypothesis when employing Accept-Support testing is to maintain low precision in your study, either by having small sample sizes or a lot of variation due to error, i.e. a sloppy experiment.

In order to guard against researchers (either intentionally or inadvertently) affirming a null hypothesis by conducting low-power experiments, alert journal editors and reviewers generally adopted what might be called the *power approach* (e.g., Schuirmann, 1987). The power approach still follows accept-support testing logic, however, it requires the researcher to establish a minimum power of 0.8 for detecting an

effect that would be considered non-negligible. A value Δ is chosen that reflects the minimum departure from zero which will no longer be accepted as trivial. When executing the test, the probability of detecting a difference that is equal to or larger than Δ should be at least 80%. The Power Approach will ensure that experiments are carried out with satisfactory care, however, the logic of the testing process itself stays flawed. If we only require power to be secured at 0.8, any further precision is still acting against the researcher's interests because it will increase the probability of rejection. Schuirmann (1987) gives an excellent discussion of the above-mentioned logical flaws. A solution to this dilemma is suggested by Equivalence Testing.

Equivalence Reject-Support testing

Equivalence testing, introduced some 30 years ago in the field of pharmaceutical research (Westlake, 1972), is a feasible alternative to *accept-support* testing that has received increasing attention in the social sciences in the past decade (Cribbie, Gruman, & Arpin-Cribbie, 2004; Rogers, Howard, & Vessey, 1993; Seaman & Serlin, 1998; Stegner, Bostrom, & Greenfield, 1996). Equivalence testing chooses new null and alternative hypotheses:

$$H_{0_a}: \ \theta_1 - \theta_2 \leq -\Delta \text{ and } H_{0_b}: \ \theta_1 - \theta_2 \geq \Delta \tag{4}$$

$$H_1: \ -\Delta < \theta_1 - \theta_2 < \Delta \tag{5}$$

The new null hypotheses state that there is a minimum difference of at least some amount Δ between the two parameters whereas the alternative hypothesis states that the actual difference is smaller than Δ. We are interested in deciding whether the difference between the two parameters, $\theta_1 - \theta_2$ lies inside the interval $[-\Delta, \Delta]$ represented by the statistical alternative hypothesis, or in the null hypothesis region $(-\infty, -\Delta] \cup [\Delta, \infty)$. Notice that the null hypothesis in this case is composed of two parts *a* and *b*. This change in hypotheses allows us to test for practical equivalence using *reject-support* testing logic, which will reward high precision and has a much higher logical congruence with the research intention of establishing equivalence. A more detailed discussion of the test statistics and procedures for equivalence testing will be given in subsequent sections.

Previous applications of equivalence testing in the social sciences have focused mainly on the difference between two means; an extension of this technique to the difference between two correlations is the topic of this study. My goal is to portray as completely and accurately as possible the different aspects of conducting an equivalence test of the difference between two correlations utilizing confidence intervals. As in traditional hypothesis testing, the test between null and alternative hypothesis in equivalence testing can be substituted by constructing a confidence interval and following certain decision rules. This will be described in more detail in subsequent sections. We chose to investigate confidence intervals because, as opposed to a significance test by itself that only tells whether the null hypothesis can be rejected or not, they provide information on the precision with which the parameter has been estimated. However, confidence intervals have several important properties, and there has been some controversy about which particular confidence interval provides the best and

most balanced optimization of these properties when used for equivalence testing. This point of investigation has already spurred discussion some 30 years ago.

This study is divided into two main parts. First I describe properties such as confidence coefficient, bias, and width of three interval procedures that could be used to replace the classical equivalence test. Secondly, I apply the confidence interval procedure of my choice to correlational equivalence testing and examine its performance using Monte Carlo analyses.

CHAPTER II

EQUIVALENCE TESTING WITH CONFIDENCE INTERVALS

Properties and Philosophy of Equivalence Testing

Although equivalence testing is RS testing, it features some properties that are new to most readers and deserve proper explanation in order to be fully understood. Since others have used other symbols and terminology for the interval $[-\Delta, \Delta]$ that constitutes the alternative hypothesis, I would like to make some references to other publications and clarify in order to prevent possible confusion. The interval $[-\Delta, \Delta]$ has been referred to as an *equivalence interval* (e.g., Cribbie et al., 2004; Schuirmann, 1987). However, I find this terminology confusing because the usage of the word "interval" in psychological methods suggests to many readers that the region has been generated by a statistical procedure. Therefore, in what follows, I refer to $[-\Delta, \Delta]$ as the *equivalence region.*

For both the lower and the upper bound of the equivalence region, I use the same symbol Δ, while others have named the bounds of the equivalence region δ_1 and δ_2 (or C_1 and C_2), thereby allowing for the possibility that the lower and upper end of this interval are not equal in absolute value, i.e. $|\delta_1| \neq |\delta_2|$ (Rogers et al., 1993; Schuirmann, 1987; Westlake, 1976). For the purpose of this study and for the sake of simplicity and comprehensibility, we assume that the equivalence region is symmetric about zero, i.e. $|-\Delta| = |\Delta|$. This does not take away much from the generality of the study, since the reader will soon learn that an equivalence region symmetric about zero is the most

common and useful option. Therefore, in this discussion, I use only one symbol Δ for a value greater than zero. The lower bound for the equivalence region will be designated by $-\Delta$.

How to choose the Δ

When we would like to show that two parameters are sufficiently close together, we need to decide what "sufficiently close" means. This is obviously a question that can only be answered based on the meaning the scale has in a specific area or on previous research. For two groups of children a difference of 3 IQ points might be close enough to say that the groups are equally smart. In another setting, the difference between 5 IQ points might be important enough that we will not declare practical equivalence. Originally, equivalence testing was developed for comparing two drugs or two formulations of the same drug in pharmaceutical research. The question of interest was often how fast two drugs dissolve into the blood system and when the difference between the two drugs was equal to or less than 20%, they were often considered to practically dissolve equally fast.

In several papers (e.g. Phillips, 1990), possibly in an attempt to replicate the formulation of equivalence bounds from pharmaceutics, equivalence bounds were established using percentages of the mean, resulting in statements like "any value within 20% of the value of mean will be sufficient for practical equivalence." Obviously, generating equivalence bounds this way only has a meaning if the scale we are measuring on has an absolute zero, i.e. is a ratio scale. Most measurements in Psychology are made on pseudo interval scales with mainly arbitrary mean and variance. A statement like

"20% within the mean" then is only meaningful if the size of the variance/standard deviation is taken into account. As an example, assume that we are measuring IQ in two different schools, and would like to assess whether IQ is practically the same in those schools. IQ measures are well studied, and so we can generally assume that the population has a mean of 100 and a standard deviation of 15 before actually looking at the sample data. To state that the two populations have practically equal means if $\bar{x}_2$ lies within 20% of $\bar{x}_1$ would render a very wide equivalence region of approximately 40 points, which on average will contain almost 82% of the general population.

The problems outlined above might suggest using standardized effect sizes to establish equivalency bounds when the scale on which we are measuring the variable or construct of interest does not have an inherent meaning (in fact, Phillips indirectly constructs power curves with respect to standardized effect sizes). A number of authors have discussed typical values for "small," "medium," and "large" effect sizes (Cohen, 1962; Sedlmeier & Gigerenzer, 1989), and the researcher might set up a Δ according to these suggestions. Nevertheless one should keep in mind that it depends on the field of research what constitutes a small or a medium effect size or what effect size seems meaningful in a given study.

The discussion of what constitutes a meaningful Δ is complex and not the focus of this study and for now I would like to assume that we were able to come up with a number Δ as a limit for "sufficiently close". Once such a Δ has been chosen, we will be able to say: If the parameter for our second group (θ_2) lies within Δ of the parameter for our first group (θ_1), the parameters for the two groups will be sufficiently close to

consider them practically equal. This the same as saying that the two parameters can be considered practically equal if their difference lies within the interval $[-\Delta, \Delta]$.

α Level in Equivalence Testing

The null hypothesis in equation (4) consists of two parts, namely H_{0_a} and H_{0_b}, that both need to be simultaneously rejected if we want to decide in favor of the alternative hypothesis. When the null hypothesis consists of several partial null hypotheses that simultaneously need to be rejected in order to reject the overall null hypothesis, this is called an *intersection–union test*.

Calculating t–statistics, we will reject H_{0_a} when

$$\frac{\Delta + \left(\hat{\theta}_1 - \hat{\theta}_2\right)}{s_{\hat{\theta}_1 - \hat{\theta}_2}} \geq t_{\alpha,\nu}. \tag{6}$$

Equivalently, we will reject H_{0_b} when

$$\frac{\Delta - \left(\hat{\theta}_1 - \hat{\theta}_2\right)}{s_{\hat{\theta}_1 - \hat{\theta}_2}} \geq t_{\alpha,\nu}, \tag{7}$$

where ν are the degrees of freedom.

If we are able to reject *both* H_{0_a} and H_{0_b}, we can assume that the true difference $\theta_1 - \theta_2$ lies within the interval $[-\Delta, \Delta]$, i.e. that the two parameters θ_1 and θ_2 are equivalent for all practical purposes. This procedure is called the *Two One-Sided Tests* procedure (TOST) (Schuirmann, 1987). Note that each individual hypothesis test is

executed at the α-level. This does not result in an overall 2α-level test, since for intersection–union tests the following holds:

Theorem 1: An intersection–union test where the i^{th} null hypothesis H_{0_i} is tested at the α_i-level and has rejection region R_i such that the overall rejection region is $R = \bigcap_i R_i$ has overall α-level equal to $\sup_i \alpha_i$ (see, e.g. Berger & Hsu, 1996; Casella & Berger, 2002, page 395).

For example, if we test each hypothesis separately at the .05 level, the test of the overall null hypothesis that the difference between θ_1 and θ_2 is at least Δ is a level .05 test. To see more clearly why the overall α-level is no greater than the greatest individual α-level in the equivalence testing situation, we need to realize that the true parameter does only have one value and can lie in only one of the two null hypothesis regions. Hence, when we commit a Type I error, we can only commit it at the α-level of the null hypothesis region the true value lies in. If we wrongly reject H_{0_a} (and $\theta_1 - \theta_2$ lies in $(-\infty, -\Delta]$), we cannot at the same time wrongly reject H_{0_b}, since $\theta_1 - \theta_2$ does not lie in $[\Delta, \infty)$ and vice versa.

Replacing Hypothesis Tests by Confidence Intervals

In the general introduction I have mentioned that the simultaneous test of H_{0_a} and H_{0_b} can be replaced by a confidence interval procedure. The correspondence between confidence intervals and hypothesis tests is not new. It is standard in introductory statistics courses to note that in the traditional hypothesis testing situation,

when $H_0 : \theta = a$, constructing a 95% confidence interval around the parameter estimate allows the following conclusions: if the confidence interval includes the null hypothesis value *a*, H_0 cannot be rejected, if it does not, it can.

In equivalence testing, the situation is similar to traditional hypothesis testing, however, it is not quite the same. In order to understand the construction of the two-sided confidence interval that will reach the same conclusions as the two hypothesis tests from Equations (6) and (7), reconsider Theorem 1. The rejection region is the subset of the parameter space that will lead to a rejection of the overall null hypothesis. In equivalence testing, the null hypothesis is divided into two parts, each of which is tested at the α- level and the rejection region for $\hat{\theta}_1 - \hat{\theta}_2$ will be

$$\left[-\Delta + t_{\alpha,\nu} s,\ \Delta - t_{\alpha,\nu} s\right], \tag{8}$$

i.e. we reject the overall null hypothesis if $\hat{\theta}_1 - \hat{\theta}_2$ lies inside $\left[-\Delta + t_{\alpha,\nu} s,\ \Delta - t_{\alpha,\nu} s\right]$. This is equivalent to saying that we will reject the overall null hypothesis if and only if the interval

$$\left(\hat{\theta}_1 - \hat{\theta}_2\right) \pm t_{\alpha,\nu} s \tag{9}$$

lies in the equivalence region $\left[-\Delta, \Delta\right]$. The interval in equation (9) is a $1-2\alpha$ confidence interval with lower bound

$$L\left(\mathbf{X}\right) = \left(\hat{\theta}_1 - \hat{\theta}_2\right) - t_{\alpha,\nu} s \tag{10}$$

and upper bound

$$U(\mathbf{X}) = \left(\hat{\theta}_1 - \hat{\theta}_2\right) + t_{\alpha,\nu} s, \tag{11}$$

where $L(\mathbf{X})$ and $U(\mathbf{X})$ are more formally defined to be the lower and upper bound, respectively, of an equivariant confidence interval (or, equivalently, a confidence *set*). The term equivariant emphasizes that for each $L(\mathbf{X})$ and $U(\mathbf{X})$, the same size α is used, thus making the confidence interval symmetric about the parameter estimate when the sampling distribution is symmetric. For the remainder of the study, traditional confidence intervals are assumed to be equivariant. Note that the interval limits, $L(\mathbf{X})$ and $U(\mathbf{X})$, are random variables. Therefore, $L(\mathbf{X})$ and $U(\mathbf{X})$ have distributions on their own.

Obviously, if we were to constructed an equivariant $1-\alpha$ confidence interval around the parameter estimate, it would reject e.g. H_{0_a} at the $\alpha/2$- level. Although the hypothesis test is conducted at the α- level, the corresponding traditional two-sided confidence interval is not a $1-\alpha$ confidence interval as in the traditional case.

This non-correspondence between α- level of the hypothesis test and size of the confidence interval has inspired the construction of other confidence intervals and procedures (Anderson & Hauck 1983; Berger & Hsu, 1996; Rocke 1984; Seaman & Serlin, 1998; Tryon, 2001; Westlake,1972; Westlake, 1976) which are supposed to replace the TOST. We will see in a later section that many equivalence testing procedures tend to be conservative when the standard error of the parameter is large compared to the equivalence region. Others (e.g. Anderson & Hauck, 1983; Rocke, 1984) have been shown to be too liberal when the standard error of the estimate becomes large (Schuirmann, 1987). The attempt to find more powerful procedures has motivated much

of the research efforts on equivalence testing, however, in my opinion some aspects of the new procedures that include utilizing confidence intervals have not received sufficient attention. An in-depth discussion of the advantages and caveats of utilizing confidence intervals in equivalence testing is yet to be found in the present literature. In order to justify a preference for one of the available procedures, I try to give a well-founded argument mainly focusing on the $1-2\alpha$ confidence interval as well as two symmetric confidence intervals suggested by Westlake (1972, 1976) and Seaman and Serlin (1998), all of which control type I error at or below the nominal α-level when used to test the null hypothesis from Equation (4).

Three Suggestions for Replacing the TOST with Confidence Intervals

The following discussion of three confidence intervals that can be used to replace the TOST procedure is based on testing the equivalence of the difference between two means. Conclusions from such a discussion will then be adapted to provide an adequate technique for correlational equivalence testing.

Assume that for two experimental conditions, you have found two independent sample means $\bar{X}_1$ and $\bar{X}_2$, the estimates for the means of a standard condition μ_1 and a new condition, μ_2, with sample standard deviations s_1 and s_2. Let's further assume the sample sizes in group 1 and 2 are n_1 and n_2, respectively and that

$$s_{\bar{X}_1-\bar{X}_2} = \sqrt{\left(\frac{n_1+n_2}{n_1 n_2}\right)\left(\frac{(n_1-1)s_1^{\,2}+(n_2-1)s_2^{\,2}}{n_1+n_2-2}\right)} \qquad (12)$$

is the estimate for the standard error of the difference between the means. Our parameter estimate $\hat{\theta} = \bar{X}_1 - \bar{X}_2$ and the standard error $s_{\bar{X}_1 - \bar{X}_2}$ will be the same for all three intervals.

I have already explored the possibility of replacing the TOST with a traditional $1-2\alpha$ confidence interval, however I would like to restate Equations (10) and (11) in terms of the difference between two means. The traditional equivariant $1-2\alpha$ or $1-\alpha$ confidence interval is given by

$$L_t(\mathbf{X}) = \left(\bar{X}_1 - \bar{X}_2\right) - s_{\bar{X}_1 - \bar{X}_2} t_{\alpha,\nu} \left(\text{or } t_{\alpha/2,\nu}\right) \tag{13}$$

and

$$U_t(\mathbf{X}) = \left(\bar{X}_1 - \bar{X}_2\right) + s_{\bar{X}_1 - \bar{X}_2} t_{\alpha,\nu} \left(\text{or } t_{\alpha/2,\nu}\right). \tag{14}$$

The first to suggest an alternative to such a procedure was Westlake (1972, 1976), who proposed a $1-\alpha$ confidence interval that is symmetric around zero; this interval is predominantly running under the title *symmetric confidence interval* in the literature. The symmetric confidence interval for the difference between two means can be constructed as follows:

Since the interval is symmetric around zero, we have $L_w(\mathbf{X}) = -U_w(\mathbf{X})$. In order to find $L_w(\mathbf{X})$ and $U_w(\mathbf{X})$ with

$$L_w(\mathbf{X}) = \left(\bar{X}_1 - \bar{X}_2\right) + k_1 s_{\bar{X}_1 - \bar{X}_2} \tag{15}$$

and

$$U_w(\mathbf{X}) = \left(\bar{X}_1 - \bar{X}_2\right) + k_2 s_{\bar{X}_1 - \bar{X}_2}, \tag{16}$$

where k_1 and k_2 are values from the *t*-distribution with $2(n-1)$ degrees of freedom, k_1 and k_2 need to fulfill two conditions:

$$P_T\left(k_1 < T < k_2\right) = 0.95 \tag{17}$$

and

$$k_1 + k_2 = \frac{\sqrt{2n}\left(\bar{X}_1 - \bar{X}_2\right)}{s_{\bar{X}_1 - \bar{X}_2}}. \tag{18}$$

The values for k_1 and k_2 cannot be found analytically — they have to be either interpolated from a table or found with the help of a mathematical package. Westlake (1976) proves that the symmetric confidence interval will cover the true parameter at least $100(1-\alpha)\%$ of the time. For both the traditional $1-2\alpha$ CI and Westlake's confidence interval, we will reject the null hypothesis of non-equivalence if the confidence interval lies inside the equivalence region $[-\Delta, \Delta]$.

An interval very similar to Westlake's in many situations has been suggested by Seaman and Serlin (1998). Seaman and Serlin's interval is much easier to construct than Westlake's:

$$L_{s\&s}\left(\mathbf{X}\right) = -\left|\bar{X}_1 - \bar{X}_2\right| - s_{\bar{X}_1 - \bar{X}_2} t_{\alpha,\nu} \tag{19}$$

$$U_{s\&s}\left(\mathbf{X}\right) = \left|\bar{X}_1 - \bar{X}_2\right| + s_{\bar{X}_1 - \bar{X}_2} t_{\alpha,\nu}. \tag{20}$$

In the overall procedure suggested by Seaman and Serlin, this interval is replaced by the traditional $1-\alpha$ confidence interval from Equations (13) and (14) when the traditional null hypothesis of no difference $H_0 : \theta_1 - \theta_2 = 0$ (see Equation (2) with $a = 0$)

is rejected. Although Seaman and Serlin's overall procedure does not suggest testing the null and alternative hypotheses from Equations (4) and

Error! Reference source not found., I would like to discuss in the following what would happen if we used this interval for testing the equivalence null hypotheses. Taking a closer look at Seaman and Serlin's overall procedure, we would have to realize that they still suggest accept-support testing. A detailed discussion of their overall procedure will not be included here.

Evaluating a Testing Procedure

The first question that needs to be asked when we would like to compare procedures for equivalence testing is "Do all procedures on average arrive at the same conclusion given a certain set of data?" This is the same as asking whether the procedures have the same *power function*.

"The *power function* of a hypothesis test with rejection region R is the function of θ defined by $\beta(\theta) = P_\theta(\mathbf{X} \in R)$ (Casella & Berger, 2002, page 383)." Further, "if C is a class of tests all testing the same null hypothesis about a parameter θ, then a test in C with power function $\beta(\theta)$ is a *uniformly most powerful* (UMP) *class C test* if $\beta(\theta) \geq \beta'(\theta)$ for every θ in the parameter space and every $\beta'(\theta)$ that is a power function of a test in class C. (Casella & Berger, 2002, page 388)."

When the null hypothesis is true, the maximum of the power function over all values of θ in the null hypothesis gives us the probability of a Type I error. When the null hypothesis is false, the power function gives us the value for power dependent on θ. Usually we would like $\beta(\theta)$ to be small when the null hypothesis is true and large when

the alternative hypothesis is true (this corresponds to a small Type I error and large power).

The maximum type I error rates across all possible parameter values for the traditional $1-2\alpha$ confidence interval, Westlake's symmetrical confidence interval, and Seaman and Serlin's interval, are known to be .05. However, Type I error rates can be substantially less. There are two separate reasons for this reduction in type I error rate, depending on the procedure. The $1-2\alpha$ CI and Seaman and Serlin's interval will have the same type I error rate as the TOST which will be significantly less than α in the following situation: Consider equations (6) and (7). Examination of these equations reminds us that for a rejection to occur, *regardless of the value of* $\hat{\theta}_1 - \hat{\theta}_2$, $t_{\alpha,\nu} s_{\hat{\theta}_1 - \hat{\theta}_2}$ must be smaller than Δ. However, $t_{\alpha,\nu} s_{\hat{\theta}_1 - \hat{\theta}_2}$ might in fact be larger than Δ, in which case one will not reject any of the null hypotheses, no matter what the observed value of $\hat{\theta}_1 - \hat{\theta}_2$ *or* the true size of $\theta_1 - \theta_2$ is. Thus, due to insufficient precision (and a large value of $\sigma_{\hat{\theta}_1 - \hat{\theta}_2}$), we might have nearly zero power and nearly zero α, a fact that will be reviewed in the discussion of the power formula (see below). This is also know as the bias of the TOST (e.g. Brown, Hwang, & Munk, 1997) and has been seen as a disadvantage by, e.g., Anderson and Hauck (1983) and Rocke (1984), which led to the development of their procedures. See Brown et al. (1997) for a discussion on the topic. In addition to low type I error rate and power caused by a lack of precision, Westlake's procedure will be conservative and have an α-level close .025 when the true difference $\theta_1 - \theta_2$ is close to zero, independently of how large an *n* we choose (Westlake, 1981). We can summarize that using the $1-2\alpha$ confidence interval and Seaman and Serlin's interval for testing the

equivalence null hypotheses will exactly mirror the decision arrived at by the TOST, while Westlake's interval will not.

Power of the TOST

We can derive a general power formula for the equivalence test for the difference between two parameters. First, let $Q(\mathbf{X}, \theta)$ be a *pivot*, i.e., a function of the data and its distributional parameters such that the distribution of $Q(\mathbf{X}, \theta)$ may be written in a form that is independent of these parameters. As an example, the variable $Z = \left(\overline{X} - \mu\right)/\sigma$ has a standard normal distribution independent of the mean and standard deviation μ and σ; another example is the random variable $T = \left(\overline{X} - \mu\right)/s$ with a *t*-distribution with *n* degrees of freedom. Considering Equation (8), we can see that the probability of rejecting the null hypothesis is

$$P_{\theta_1-\theta_2}\left(\left(\hat{\theta}_1 - \hat{\theta}_2\right) \in \left[-\Delta + t_{\alpha,\nu}s,\ \Delta - t_{\alpha,\nu}s\right]\right), \tag{21}$$

where *s* is the estimated standard error for the difference between two means from Equation (12). This is the same as

$$P\left(\frac{-\Delta + t_{\alpha,\nu}s - \left(\theta_1 - \theta_2\right)}{s} \le \frac{\left(\hat{\theta}_1 - \hat{\theta}_2\right) - \left(\theta_1 - \theta_2\right)}{s} \le \frac{\Delta - t_{\alpha,\nu}s - \left(\theta_1 - \theta_2\right)}{s}\right). \tag{22}$$

When testing for equivalence of the difference between two means, the pivot

$$Q = \frac{\left(\hat{\theta}_1 - \hat{\theta}_2\right) - \left(\theta_1 - \theta_2\right)}{s} \tag{23}$$

is distributed as a noncentral *t* random variable with $n_1 + n_2 - 2$ degrees of freedom and cdf F_T, where $\hat{\theta}_1 - \hat{\theta}_2 = \bar{X}_1 - \bar{X}_2$ and $\theta_1 - \theta_2 = \mu_1 - \mu_2$. Given an estimate *s* for the standard deviation, we can approximate power by

$$\beta(\theta_1 - \theta_2) \approx F_T\left(\frac{\Delta - t_{\alpha,\nu}s - (\theta_1 - \theta_2)}{s}\right) - F_T\left(\frac{-\Delta + t_{\alpha,\nu}s - (\theta_1 - \theta_2)}{s}\right). \quad (24)$$

Power estimates from Equation (24) will have negative values whenever $\Delta - t_{\alpha,\nu}s - (\theta_1 - \theta_2)$ is smaller than $-\Delta + t_{\alpha,\nu}s - (\theta_1 - \theta_2)$. In that case, it is sufficient to set the value to zero. It is especially insightful to consider why power can be zero if we look at the $1-2\alpha$ confidence interval procedure. The null hypothesis will only be rejected when the confidence interval lies entirely inside the equivalence region, that is, when $-\Delta < \left(\hat{\theta}_1 - \hat{\theta}_2\right) \pm t_{\alpha,\nu}s < \Delta$. When *n* is small enough, $2t_{\alpha,\nu}s$ (the width of the confidence interval) might be larger than 2Δ (the width of the equivalence region). Certainly, the confidence interval cannot lie within $[-\Delta, \Delta]$ then. Finding a test that will maximize $\beta(\theta_1 - \theta_2)$ when $|\theta_1 - \theta_2| \leq \Delta$ while keeping it smaller or equal to .05 as long as $|\theta_1 - \theta_2| > \Delta$ has inspired most recent research efforts (e.g. Berger & Hsu, 1996).

Power analysis has, in general, received inadequate attention in applied statistics literature (Cohen, 1962; Sedlmeier & Gigerenzer, 1989), and equivalence testing is no exception. Only a few articles on the power of equivalence tests have found their way into psychological journals. Phillips (1990) constructed power curves for equivalence testing of the difference between two means, finding encouraging results that seem to promise sufficient power at sample sizes as small as 30. However, his results are based on very wide equivalence regions, sometimes as wide as four times the size of the

standard deviation, which corresponds to $\Delta = 2$ standardized effect sizes. When $\Delta = 2/3$ of a standardized effect size, a much more realistic width for an equivalence region, power results look less impressive. However, with large *N*, e.g. $N \geq 100$, power might still be sufficient to justify equivalence testing. Other power analyses have focused on hypotheses inherent in biomedical research (see, e.g. Feng & Liang, 2006) and do not seem easily transferable to research situations in the social sciences. Certainly, power for testing the equivalence of two means can easily be estimated using Equation (24). While this article is not on testing the difference between two means in particular, I do provide power graphs for correlational equivalence testing later.

Properties of Confidence Intervals in Equivalence Testing

When we replace the TOST with a confidence interval procedure, we need to take not only α-level and power into account, but also the properties of confidence intervals themselves. Point estimators are evaluated with respect to their bias, sufficiency, and efficiency, and significance tests differ with respect to their power function. Similarly, confidence intervals have a set of characteristics on which they need to be compared. So far we have seen that empirical Type I error rate is controlled satisfactorily when an equivariant $1 - 2\alpha$ confidence interval is used for equivalence testing of the difference between two means (Berger & Hsu, 1996). Westlake has shown that his procedure controls Type I error rate at or below the nominal level as well. Again assuming that we used Seaman and Serlin's interval to test the hypotheses from Equations (4) and **Error! Reference source not found.**, we can see that Type I error rate will be controlled at or below the nominal level since its construction involves pivoting the cdf for the

TOST. Therefore it seems feasible and necessary to compare the procedures on other characteristics. I base my discussion of characteristics for confidence intervals on the excellent summary that is given in Casella and Berger's *Statistical Inference* (2002).

Confidence intervals have three main properties: *Coverage rate* (or *coverage probability*), *bias* and *width* (Casella & Berger, 2002) .

Coverage probability is defined as follows (see Casella & Berger (2002), page 418):

Definition 1: "For an interval estimator $\left[L(\mathbf{X}),U(\mathbf{X})\right]$ of a parameter θ, the *coverage probability* of $\left[L(\mathbf{X}),U(\mathbf{X})\right]$ is the probability that the random interval $\left[L(\mathbf{X}),U(\mathbf{X})\right]$ covers the true parameter, θ. It is denoted by

$$P_\theta\left(\theta\in\left[L(\mathbf{X}),U(\mathbf{X})\right]\right)." \tag{25}$$

For a given $1-\alpha$ confidence interval, we expect that it covers the true parameter at least $100(1-\alpha)\%$ of the time, independently of the true value of the parameter. The definition of the confidence coefficient provides us with an adequate short description of this requirement:

Definition 2: "For an interval estimator $\left[L(\mathbf{X}),U(\mathbf{X})\right]$ of a parameter θ, the *confidence coefficient* of $\left[L(\mathbf{X}),U(\mathbf{X})\right]$ is the infimum of the coverage probabilities."

That is, for a $100(1-\alpha)\%$ confidence interval, we expect that

$$\inf_\theta P_\theta\left(\theta\in\left[L(\mathbf{X}),U(\mathbf{X})\right]\right)=1-\alpha\,. \tag{26}$$

At the same time, we would consider a confidence interval optimal that covers the true parameter "as often as possible," but does not cover other values "too often." We write θ' for a value that is not equal to θ, $\theta' \neq \theta$. Then *false coverage* is defined as (Casella & Berger, 2002, page 444):

Definition 3:

$$P_\theta\left(\theta' \in C(\mathbf{X})\right), \theta \neq \theta' \text{ if } C(\mathbf{X}) = \left[L(\mathbf{X}), U(\mathbf{X})\right] \tag{27}$$

for a two-sided confidence interval,

$$P_\theta\left(\theta' \in C(\mathbf{X})\right), \theta' < \theta \text{ if } C(\mathbf{X}) = \left[L(\mathbf{X}), -\infty\right) \tag{28}$$

for a one-sided confidence interval with a lower bound, and

$$P_\theta\left(\theta' \in C(\mathbf{X})\right), \theta' > \theta \text{ if } C(\mathbf{X}) = \left(-\infty, U(\mathbf{X})\right] \tag{29}$$

for a one-sided confidence interval with an upper bound.

To give an example, the probability that the value $\theta' = 0.5$ will be covered by a two-sided confidence interval $C(\mathbf{X})$ that is computed from sample values $\mathbf{X}$ when $\theta = 1$ is the true value of the parameter will be $P_1\left(0.5 \in C(\mathbf{X})\right)$; if $C(\mathbf{X})$ is unbiased, this will be smaller than $P_1\left(1 \in C(\mathbf{X})\right)$. Equations (28) and (29) define a one-sided CI to have false coverage only when a value θ' closer to the respective bound is covered more often than θ. It is questionable whether anyone will be able to find a one-sided confidence interval that has false coverage greater than coverage rate of the true parameter. However, as we will see later on, such a definition will help making an argument with respect to the three confidence intervals we are comparing in the study.

The relationship between coverage rate and false coverage leads us to the definition of bias for two-sided confidence intervals: An unbiased two-sided confidence interval for θ will cover no value more often than it covers θ:

Definition 4: "A $1-\alpha$ confidence set $C(\mathbf{x})$ is *unbiased* if $P_\theta\left(\theta' \in C(\mathbf{x})\right) \leq 1-\alpha$ for all $\theta \neq \theta'$." (Casella & Berger (2002), page 446)

Two-sided intervals for the mean given by

$$C(\bar{x}) = \bar{x} \pm z_\alpha \sigma_{\bar{x}} \tag{30}$$

and

$$C(\bar{x}) = \bar{x} \pm t_{\alpha,\nu} s_{\bar{x}} \tag{31}$$

are unbiased (Casella & Berger, 2002).

The definition for bias from above can be understood easily and it can sometimes be relatively easy to demonstrate that a given confidence interval *is* biased. However, showing that any given confidence interval is unbiased in practice proves to be more complicated. At this point it would be very useful to be able to make additional statements of the form "when condition *XY* is satisfied, the present (two-sided) CI will be unbiased". One possibility that seems worth exploring is to show that a confidence interval that misses the true parameter equally often on both sides will be unbiased. Such a confidence interval will be said to have *coverage balance*, while *coverage imbalance* occurs when a confidence interval does not miss the true parameter equally often above and below.

The third property of confidence intervals, width, directly addresses the length $U(\mathbf{X}) - L(\mathbf{X})$ of a confidence interval. A two-sided normal confidence interval for the

mean symmetric about the parameter estimate as in Equation (30) is the shortest length $1-\alpha$ confidence interval and so is its counterpart for unknown σ (Casella & Berger, 2002, page 443).

Evaluating the Properties of the Replacement Procedures

After reiterating classical properties of confidence intervals, I would like to compare our three options for replacing the TOST with confidence intervals with respect to these properties. While defining the properties of interest, we have seen that the traditional normal two-sided $1-2\alpha$ confidence interval will perform well. It is easy to verify that it has a confidence coefficient of $1-2\alpha$ and the proof shall not be repeated here. Similarly, Westlake has shown in his 1976 article that his procedure has a confidence coefficient of $1-\alpha$. Seaman and Serlin (1998) use Monte Carlo simulations to demonstrate that their confidence interval has a confidence coefficient of $1-\alpha$. Here is a quick outline of why this is true: Considering Equations (19) and (20), the confidence interval will not cover the true parameter when $\theta_1-\theta_2$ lies outside of

$$\pm\left(\left|\hat{\theta}_1-\hat{\theta}_2\right|+s_{\hat{\theta}_1-\hat{\theta}_2}t_{\alpha,\nu}\right). \tag{32}$$

Assuming that $\theta_1-\theta_2>0$, the probability that it lies outside the interval from Equation (32) will be

$$
\begin{aligned}
&P\left(\theta_1-\theta_2>\left|\hat{\theta}_1-\hat{\theta}_2\right|+s_{\hat{\theta}_1-\hat{\theta}_2}t_{\alpha,\nu}\right)\\
&=P\left(-t_{\alpha,\nu}>\frac{\left|\hat{\theta}_1-\hat{\theta}_2\right|-\left(\theta_1-\theta_2\right)}{s_{\hat{\theta}_1-\hat{\theta}_2}}\right)\\
&\le P\left(-t_{\alpha,\nu}>\frac{\left(\hat{\theta}_1-\hat{\theta}_2\right)-\left(\theta_1-\theta_2\right)}{s_{\hat{\theta}_1-\hat{\theta}_2}}\right)\\
&=\alpha.
\end{aligned}
\tag{33}
$$

Conversely, when $\theta_1-\theta_2<0$, the probability will be

$$
\begin{aligned}
&P\left(\theta_1-\theta_2<-\left|\hat{\theta}_1-\hat{\theta}_2\right|-s_{\hat{\theta}_1-\hat{\theta}_2}t_{\alpha,\nu}\right)\\
&=P\left(t_{\alpha,\nu}<\frac{-\left|\hat{\theta}_1-\hat{\theta}_2\right|-\left(\theta_1-\theta_2\right)}{s_{\hat{\theta}_1-\hat{\theta}_2}}\right)\\
&\le P\left(t_{\alpha,\nu}<\frac{\left(\hat{\theta}_1-\hat{\theta}_2\right)-\left(\theta_1-\theta_2\right)}{s_{\hat{\theta}_1-\hat{\theta}_2}}\right)\\
&=\alpha.
\end{aligned}
\tag{34}
$$

Hence,

$$
\sup_{\theta_1-\theta_2}P\left(\theta_1-\theta_2\notin\pm\left(\left|\hat{\theta}_1-\hat{\theta}_2\right|+s_{\hat{\theta}_1-\hat{\theta}_2}t_{\alpha,\nu}\right)\right)=\alpha\,. \tag{35}
$$

The second property, bias, turns out to be of considerably more interest. As mentioned above, the normal two-sided confidence interval from Equations (30) and (31) will be unbiased. Hence, in the case where we are making inferences on parameter estimates whose pivot is distributed as a *z* random variable or *t* random variable, we can safely assume that the interval will not be biased.

Turning to Westlake's and Seaman and Serlin's confidence intervals, which are symmetric about zero, we can quickly see that we might run into an issue here. It has been satisfactorily shown that their CIs have confidence coefficients $\le 95\%$. However,

the intervals will also always cover zero, although zero is virtually never going to be the true parameter value (see e.g. Kirkwood, 1981), and thus, the interval is biased – if we follow the definition of bias for a two-sided CI from Equation (30). It is amazing that this property of both procedures has not found its way into the discussion surrounding equivalence testing, but in fact additional and only slightly different procedures are suggested (e.g. Berger & Hsu, 1996). As this study goes on, I will suggest an alternative perspective on both Seaman and Serlin's and Westlake's symmetric CIs. Looked at from this different perspective, their intervals will no longer be biased, but in my opinion display their true nature more accurately. Before giving a formal discussion of width for all three confidence interval procedures, I would like to demonstrate their construction with two examples, one with large *N* and one with small *N*. This will also facilitate understanding why Westlake received strong criticism from several other authors (Kirkwood, 1981; Mantel, 1977).

Examples for all Three Procedures

Below I have constructed the Westlake CI, the confidence interval suggested by Seaman and Serlin, the $1-2\alpha$ traditional CI, and the $1-\alpha$ traditional confidence interval for some example data.

Example 1: Let's assume that $\bar{X}_1 = 53$, $\bar{X}_2 = 50$, $N_1 = N_2 = 300$, $\Delta = 5$, $\alpha = .05$, $t_{\alpha,\nu} \approx 1.64741$, $t_{\alpha/2,\nu} \approx 1.96394$ and $s_1 = s_2 = 8$ which leads to $s_{\bar{X}_1 - \bar{X}_2} \approx 0.653197$. Then

(a) Westlake $\quad [-4.0761,\ 4.0761]$

(b) Seaman & Serlin $\quad [-4.0761,\ 4.0761]\ (\rightarrow [1.7172,\ 4.2828])$

(c) $1-2\alpha$ *traditional CI* $\quad [1.9239,\ 4.0761]$

(d) $1-\alpha$ *traditional CI* $\quad [1.7172,\ 4.2828]$

When reading Seaman and Serlin's 1998 article on equivalence testing, one might get the impression that they are suggesting an interval that is substantively different from Westlake's. However, for a number of conditions, Westlake's and Seaman and Serlin's intervals are going to be essentially equal numerically. In this example, all four intervals when used for testing the null hypothesis from Equation (4) would yield the conclusion that there is practical equivalence. In their overall procedure, Seaman and Serlin would retain the $1-\alpha$ traditional confidence interval. The $1-2\alpha$ confidence interval lies entirely within the intervals proposed by both Westlake (a) and Seaman and Serlin (b) and its width is only 26.4% of the width of either (a) or (b). Although the $1-\alpha$ traditional confidence interval has the most extreme endpoint (4.2828) among all four procedures, its width is still only a fraction of that of (a) and (b).

Example 2: If we choose a smaller sample size, say $N_1 = N_2 = 10$, the relationships are still somewhat similar. Let's assume that $\bar{X}_1 = 50$, $\bar{X}_2 = 53$,

$N_1 = N_2 = 10$, $\Delta = 5$, $\alpha = .05$, $t_{\alpha,\nu} \approx 1.73406$, $t_{\alpha/2,\nu} \approx 2.10092$ and $s_1 = s_2 = 8$ which leads to $s_{\bar{X}_1 - \bar{X}_2} \approx 3.5777$. Then

(a) Westlake $\quad [-9.264,\ 9.264]$

(b) Seaman & Serlin $\quad [-9.204,\ 9.204]$

(c) $1-2\alpha$ *traditional CI* $\quad [-3.204,\ 9.204]$

(d) $1-\alpha$ *traditional CI* $\quad [-4.516,\ 10.516]$

None of the procedures would allow us to conclude in favor of practical equivalency, due to a lack of precision. However, this example shows that Seaman and Serlin's interval is somewhat narrower than Westlake's. The $1-2\alpha$ confidence interval lies entirely inside the Westlake confidence interval as well as the Seaman and Serlin interval.

Similar illustrations opened the door for early comments on Westlake's procedure. The first to critique the symmetric confidence interval was Mantel (1977) and Kirkwood (1981) who mentioned that Westlake's procedure will yield meaningless intervals when the difference between the two sample parameters is large. Even if the true difference $\mu_1 - \mu_2$ still lies within $[-\Delta, \Delta]$, Westlake's interval ignores information about the difference, as basically only one of the confidence limits is constructed while the other is simply a mirror image of the first. This argument largely applies to Seaman and Serlin's interval as well, since it only differs by a small amount from Westlake's procedure.

Westlake's symmetric interval has been criticized for the rate at which it misses the true value of the parameter as well. We can observe that when we construct Westlake's (and Seaman and Serlin's) interval, it will miss the true parameter either always below or always above. When θ is positive, the probability that the interval lies *above* θ is zero: $P_{\theta>0}\left(\theta < \left[L(\mathbf{X}), U(\mathbf{X})\right]\right) = 0$, when θ is negative, the probability that the interval lies *below* θ will be zero: $P_{\theta<0}\left(\theta > \left[L(\mathbf{X}), U(\mathbf{X})\right]\right) = 0$. Either way, whenever (a) and (b) miss the true parameter, they will inevitably underestimate and never overestimate its absolute value. With a confidence coefficient of 5%, this means that up to 5% of the time, intervals (a) and (b) will suggest a range for reasonable values for the parameter not including the same whose absolute value will be larger than estimated. The traditional $1-2\alpha$ confidence interval, however, will produce a range of values for the true parameter that is too large 5% of the time and too small another 5% of the time. Hence, 5% of the time the intervals symmetric about zero underestimate the absolute value of true difference, while the $1-2\alpha$ confidence interval underestimates the absolute value of true difference only 5% of the time as well.

So far we have seen that for large samples, Westlake's and Seaman and Serlin's procedures yield basically the same result. We also saw that the traditional $1-2\alpha$ and $1-\alpha$ confidence intervals can be substantially narrower than the symmetric intervals and that both Westlake's and Seaman and Serlin's intervals will always underestimate the absolute value of the true difference.

Comparing the Widths of the Symmetric and the Traditional Confidence Intervals

A complete coverage of the topic width for confidence intervals is beyond the scope of this study, but I would like to direct the reader's attention to an interesting detail regarding the width of traditional and symmetric CIs. Seaman and Serlin (1998) seem to claim that the interval from Equations (19) and (20) will yield a $1 - \alpha$ confidence interval that will be narrower than the traditional $1 - \alpha$ confidence interval for small values of $\mu_1 - \mu_2$. I will show that even when $\mu_1 - \mu_2 = 0$, on average the interval from Equation (19) and (20) will be wider than the traditional $1 - \alpha$ CI for the most common values for Type I error rate α.

Let's assume $\mu_1 - \mu_2 = 0$, then its estimator $\bar{X}_1 - \bar{X}_2$ will have a sampling distribution with mean 0 and a standard error that can be estimated with $s_{\bar{X}_1 - \bar{X}_2}$. The traditional $1 - \alpha$ CI is calculated as in Equations (13) and (14). Dividing by the standard error $\sigma_{\bar{X}_1 - \bar{X}_2}$, which will be the same for both intervals, we have width

$$w_t = 2t_{\alpha/2,\nu}. \tag{36}$$

Seaman and Serlin's interval is calculated as in Equations (19) and (20), and width divided by standard error will be

$$w_{S\&S} = 2\frac{\left|\bar{X}_1 - \bar{X}_2\right|}{s_{\bar{X}_1 - \bar{X}_2}} + 2t_{\alpha,\nu}. \tag{37}$$

We would like to find the average length of both w_t and $w_{S\&S}$ and show that, on average, $E(w_t) < E(w_{S\&S})$. $E(w_t)$ will be a constant for any given *n* and chosen α-level. The expected value of $w_{S\&S}$ is a little more difficult to obtain:

$$E(w_{S\&S}) = 2E\left(\frac{\left|\bar{X}_1 - \bar{X}_2\right|}{s_{\bar{X}_1-\bar{X}_2}}\right) + 2E\left(t_{\alpha,\nu}\right). \tag{38}$$

$E\left(t_{1-\alpha,\nu}\right)$ will be a constant, and we are left with finding the expected value of the first term in (38). $\left|\bar{X}_1 - \bar{X}_2\right| / s_{\bar{X}_1-\bar{X}_2}$ is distributed as the absolute value of a *t*-random variable with $\nu = n_1 + n_2 - 2$ degrees of freedom which has expected value ≥ 0.7979:

$$\inf_{\nu\to\infty} E\left(\frac{\left|\bar{X}_1 - \bar{X}_2\right|}{s_{\bar{X}_1-\bar{X}_2}}\right) = \inf_{\nu\to\infty} E\left(\frac{|Z|}{\sqrt{\chi_\nu^{\,2}/\nu}}\right) = E\left(|Z|\right) = 0.7979. \tag{39}$$

Choosing 0.7979 as the lower limit, we find that

$$E(w_{S\&S}) \;\geq\; 1.5958 + 2t_{\alpha,\nu}. \tag{40}$$

The term in Equation (40) will, for common α-levels, be larger than the expected value for width of the traditional $1-\alpha$ CI from Equation (36).

Assuming a different point of view, we can ask ourselves how often Seaman and Serlin's CI will be narrower than the traditional $1-\alpha$ CI, given that $\mu_1 - \mu_2 = 0$. For example, let's assume that $\alpha = .05$ and $\nu = 10$, then we can calculate the probability that Seaman and Serlin's interval will be narrower than the traditional interval: It will be narrower when $T = \left|\bar{X}_1 - \bar{X}_2\right| / \sigma_{\bar{X}_1-\bar{X}_2}$ is smaller than the absolute value of the difference

between $t_{.025,10}$ and $t_{.05,10}$. For $\alpha = .05$ and $\nu = 10$, T will be smaller than $2.306 - 1.85955 = 0.44645$ about 33.5% of the time.

On a final note, we can say that the results for Seaman and Serlin's interval apply to Westlake's interval as well since it will be wider than the former.

Reconstructing the Symmetric Intervals

We have previously observed that the symmetric intervals will always cover zero, independently of what the true value of the parameter is. If we conceive of these intervals as two-sided confidence intervals, they will be heavily biased following Definition 4. According to Kirkwood (1981), such bias might facilitate a misunderstanding regarding the parameter of interest that will have practical consequences for the average data analysis consumer. Equating the center of a confidence interval with the best estimate of a parameter, he seems to imply that for the symmetric interval, zero might be understood to be the best estimate of the parameter. I do not feel convinced that the best estimate of a parameter ought to be the center of its confidence interval. This is certainly not practical in situations where the parameter estimate has a bounded, strongly skewed distribution such as a sample population proportion $\hat{p}$ e.g., where $p = .10$ with small *n*. However, we can assume that the difference between two sample means or similar parameter estimates will have symmetrical, unbounded distributions and a confidence interval whose center will, on average, be the true value of the parameter seems to provide more information than an interval that is not centered on the parameter. While contemplating this, I developed the idea that the information the symmetric intervals provide us with is information on the absolute value of the difference between two means.

The Symmetric Intervals as one-sided $1-\alpha$ *CIs for* $|\mu_1-\mu_2|$

For demonstrative purposes, let $C_{S\&S}\left(T\left(\mathbf{X}\right)\right)$ be Seaman and Serlin's symmetric confidence interval from Equations (19) and (20) on some statistic $T\left(\mathbf{X}\right)$ distributed as a *t* random variable. Let $C_a\left(T\left(\mathbf{X}\right)\right)$ be a confidence interval for the absolute value of $T\left(\mathbf{X}\right)$ such that

$$L_a\left(T\left(\mathbf{X}\right)\right)=0 \tag{41}$$

and

$$U_a\left(T\left(\mathbf{X}\right)\right)=U_{S\&S}\left(T\left(\mathbf{X}\right)\right). \tag{42}$$

Then $C_a\left(T\left(\mathbf{X}\right)\right)$ has confidence coefficient $1-\alpha$ and will cover $\left|T\left(\mathbf{X}\right)\right|$ as often as $C_{S\&S}\left(T\left(\mathbf{X}\right)\right)$ covers $T\left(\mathbf{X}\right)$. I have shown above shown (1976) that

$$\inf_{\mu_1-\mu_2} P\left(\mu_1-\mu_2\in\left[L_{S\&S}\left(\bar{X}_1-\bar{X}_2\right),\ U_{S\&S}\left(\bar{X}_1-\bar{X}_2\right)\right]\right)=1-\alpha\,. \tag{43}$$

From that we can derive the confidence coefficient of $C_a\left(T\left(\mathbf{X}\right)\right)$

$$\begin{aligned}
&1-\alpha\\
&=\inf_{\mu_1-\mu_2} P\left(\mu_1-\mu_2\in\left[L_{S\&S}\left(\bar{X}_1-\bar{X}_2\right),\ U_{S\&S}\left(\bar{X}_1-\bar{X}_2\right)\right]\right)\\
&=\inf_{\mu_1-\mu_2} P\left(L_{S\&S}\left(\bar{X}_1-\bar{X}_2\right)\le\mu_1-\mu_2\le U_{S\&S}\left(\bar{X}_1-\bar{X}_2\right)\right)\\
&=\inf_{\mu_1-\mu_2} P\left(-U_{S\&S}\left(\bar{X}_1-\bar{X}_2\right)\le\mu_1-\mu_2\le U_{S\&S}\left(\bar{X}_1-\bar{X}_2\right)\right)\\
&=\inf_{\mu_1-\mu_2} P\left(0\le\left|\mu_1-\mu_2\right|\le U_{S\&S}\left(\bar{X}_1-\bar{X}_2\right)\right)\\
&=\inf_{\mu_1-\mu_2} P\left(L_a\left(\bar{X}_1-\bar{X}_2\right)\le\left|\mu_1-\mu_2\right|\le U_a\left(\bar{X}_1-\bar{X}_2\right)\right).
\end{aligned} \tag{44}$$

Further,

$$
\begin{aligned}
&P\left(\mu_1-\mu_2\in\left[L_{S\&S}\left(\bar{X}_1-\bar{X}_2\right),\ U_{S\&S}\left(\bar{X}_1-\bar{X}_2\right)\right]\right)\\
&=P\left(\left|\mu_1-\mu_2\right|\in\left[0,\ U_a\left(\bar{X}_1-\bar{X}_2\right)\right]\right).
\end{aligned}
\tag{45}
$$

It is easy to see that the same argument holds for Westlake's confidence interval, which is symmetric about zero and has confidence coefficient = $1-\alpha$ as well.

If we adopt the view that what Westlake and Seaman and Serlin have constructed are one-sided confidence intervals for the absolute value of the difference between two means, their procedure is no longer biased, considering Definitions 3 and 4. I find this new perspective on the two procedures especially helpful, since the symmetry of the null hypotheses and the obtained intervals is retained while avoiding certain expectations that are routinely connected to the construction of two-sided confidence intervals (e.g. unbiasedness). The name symmetric interval seems to imply that the interval provides information with respect to two bounds for parameter estimates while technically only one side is determined from the data, the other side is merely a duplication of the first. Westlake's and Seaman and Serlin's intervals are symmetric about zero, but this symmetry is arbitrary with respect to explanatory power because the interval is going to be symmetric about zero independently of the data. Table 1 summarizes the properties of confidence intervals that have been suggested for substitution.

Table 1

Summary of confidence intervals suggested for equivalence testing

	Coverage Rate	Confidence Coefficient	Bias	Width	Relationship to TOST/other
Westlake	Between $1-\alpha$ and 1	$1-\alpha$	Biased when conceptualized as a two-sided CI but unbiased when thought of as one-sided CI for the absolute value	Wider than Seaman and Serlin's CI	Not equivalent to TOST, too conservative, the same as a one-sided CI for $\lvert\theta\rvert$ with confidence coefficient $1-\alpha$
Seaman and Serlin	Between $1-\alpha$ and 1	$1-\alpha$		On average wider than the traditional $1-\alpha$ CI	Equivalent to TOST, the same as a one-sided CI for $\lvert\theta\rvert$ with confidence coefficient $1-\alpha$
Traditional $1-2\alpha$ CI	$1-2\alpha$	$1-2\alpha$	Unbiased	Narrowest	Equivalent to TOST
Traditional $1-\alpha$ CI	$1-\alpha$	$1-\alpha$	Unbiased	On average narrower than Seaman and Serlin	Not Equivalent to TOST, too conservative

CHAPTER III

CORRELATIONAL EQUIVALENCE TESTING

In the previous sections I have attempted to shed some light on the properties of several confidence interval procedures that have been proposed as replacements for the TOST. The second part is concerned with the application of equivalence testing utilizing confidence intervals to the difference between two correlations. I decided to derive a test and examine its performance based on the traditional confidence interval from Equations (10) and (11) for several reasons: (1) the traditional confidence interval is the only unbiased two-sided confidence interval; (2) The other two techniques are either strongly biased or have to be conceived as one-sided confidence intervals; (3) Further, it is the narrowest interval that allows testing both equivalency null hypotheses in Equation (4) simultaneously at the .05 Type I error rate.

There are many possible settings in which one might want to test the difference between two correlations for equivalence which I try to summarize in Figure 1. I constructed a statistic for testing the difference between two independent correlations coming from normally distributed data. Furthermore, the statistic will be asymptotically normally distributed and not exact.

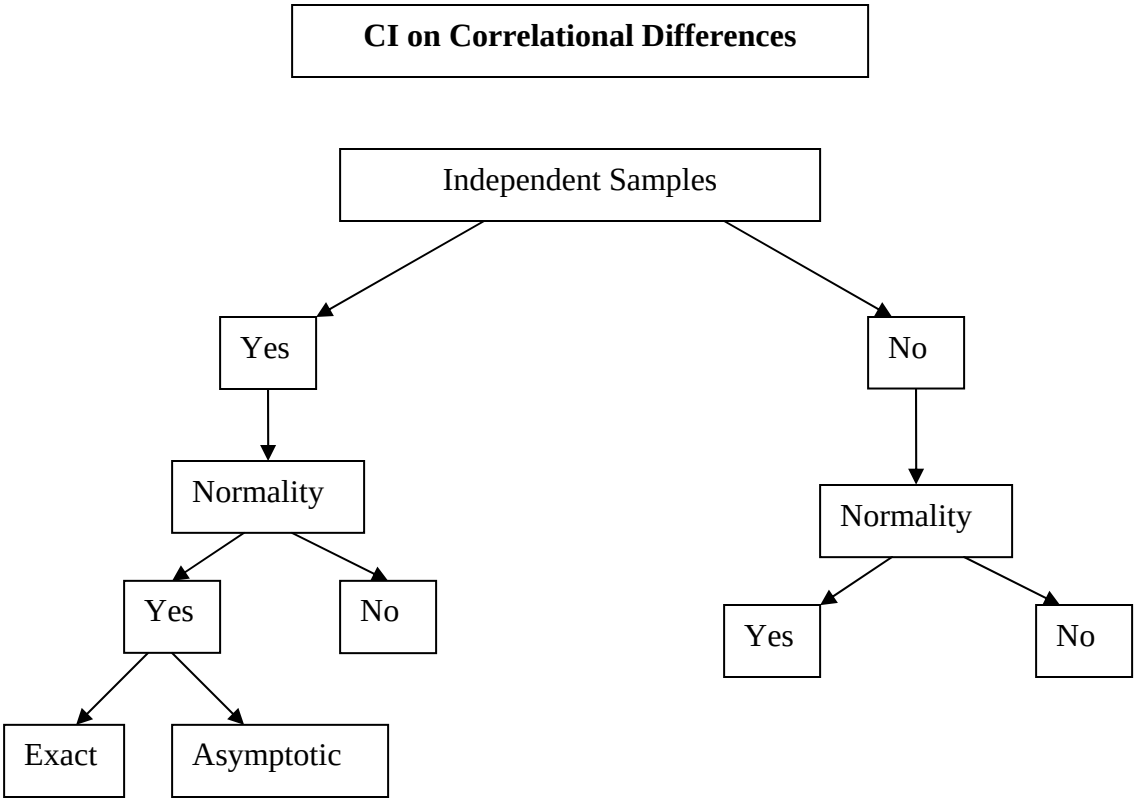

Figure 1. Tree diagram for confidence intervals for correlational differences, distinguishing between intervals for independent samples, normal populations and exact vs. asymptotic intervals

Derivation of the Statistic and the Confidence Interval

For simplicity, designate the larger one of the two correlations with ρ_1 and the smaller one with ρ_2 such that $\rho_1 - \rho_2 > 0$. Estimating ρ_1 and ρ_2 in the sample, we obtain $\hat{\rho}_1 = r_1$ and $\hat{\rho}_2 = r_2$, the familiar sample correlations. The asymptotic variance of a single sample correlation is:

$$\sigma^2{}_r = \frac{\left(1 - r^2\right)^2}{\left(n - 2\right)} \tag{46}$$

Assuming that r_1 and r_2 are independent, the standard deviation of $r_1 - r_2$ can be estimated as (see, e.g. Olkin & Finn, 1995):

$$s_{r_1 - r_2} = \sqrt{\frac{\left(1 - r_1^2\right)^2}{\left(n_1 - 2\right)} + \frac{\left(1 - r_2^2\right)^2}{\left(n_2 - 2\right)}} \tag{47}$$

If we wish to conduct the TOST for correlational equivalence testing, we need to construct test statistics corresponding to Equations (6) and (7). $H_{0_a} : \rho_1 - \rho_2 \leq -\Delta$ and $H_{0_b} : \rho_1 - \rho_2 \geq \Delta$ can be tested using the following test statistics:

$$Z_l = \frac{\Delta + \left(r_1 - r_2\right)}{s_{r_1 - r_2}} \tag{48}$$

and

$$Z_u = \frac{\Delta - \left(r_1 - r_2\right)}{s_{r_1 - r_2}}, \tag{49}$$

where $s_{r_1 - r_2}$ is the standard deviation estimate from equation (47). We are assuming here that Z_l and Z_u will be close to normally distributed and hence the values resulting from (48) and (49) will be compared to a critical value from the standard normal table. As an example, we are going to reject both null hypotheses and accept the alternative hypothesis at a .05 Type I error level when both Z_l and Z_u are greater than or equal to 1.645.

To construct a traditional $1 - 2\alpha$ confidence interval around the parameter estimate $r_1 - r_2$, we utilize the estimate for the standard deviation from equation (47):

$$\left(r_1 - r_2\right) \pm z_\alpha s_{r_1 - r_2} \tag{50}$$

The null hypotheses will be rejected when the confidence interval in (50) lies entirely inside the equivalence region $\left[-\Delta, \Delta\right]$.

Why do we not use the Fisher z-transform for this test?

When testing a single correlation, the Fisher z–transform is widely used for its accuracy and comparable simplicity. Restating the Fisher z-transform, we see that it is a non-linear transformation:

$$f(x) = \frac{1}{2}\ln\left(\frac{1+x}{1-x}\right) \tag{51}$$

The standard deviation for $f\left(r\right)$ can be approximated by $\sqrt{1/(n-3)}$, so that we can test a point value null hypothesis $H_0 : \rho = a$ with the test statistic

$$Z = \frac{f\left(r\right) - f\left(a\right)}{\sqrt{1/(n-3)}} \tag{52}$$

which is approximately $\sim N\left(0,1\right)$. Applying this transformation to a single correlation yields impressive results. Further, the z-transform can be used to test whether two correlations are the same $H_0 : \rho_1 = \rho_2$, which is the same as testing whether the difference between two correlations is zero $\left(H_0 : \rho_1 - \rho_2 = 0\right)$. The test statistic will be

$$\frac{f\left(r_1\right) - f\left(r_2\right)}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \tag{53}$$

and is approximately normally distributed with mean 0 and SD 1. These two tests belong to the class of pattern hypotheses for correlations which test whether a set of correlations is equal to each other or to some point value. In general, we can test $H_0 : \rho_1 = ... = \rho_m = a$ using the Fisher z-transform (e.g., Hedges & Olkin, 1983). By contrast, linear hypotheses are hypotheses about linear combinations of correlations. The Fisher z-transform cannot be used on linear hypotheses, therefore, we cannot test whether the difference between two correlations is equal to some point value other than zero, e.g. $H_0 : \rho_1 - \rho_2 = a$, where $a \neq 0$. Obviously, the two one-sided tests used in equivalence testing are of this form and the z-transform cannot be applied.

Power Calculations

Once the size of Δ has been chosen, power for the correlational equivalence test depends on two quantities: The actual difference between the two correlations and the sample sizes in the two samples the correlations are calculated from. The estimate for $\sigma_{r_1 - r_2}$ from Equation (47) solely depends on the size of $\rho_1 - \rho_2$ as well as n_1 and n_2. For a given size of our parameter $\rho_1 - \rho_2$, we can approximate power analogously to Equation (21) by

$$P\left(\left[-\Delta + \sigma_{r_1 - r_2} z_\alpha \leq \ r_1 - r_2 \ \leq \Delta - \sigma_{r_1 - r_2} z_\alpha\right] | \, \rho_1 - \rho_2\right), \tag{54}$$

which is equal to 0 when $z_\alpha \sigma_{r_1 - r_2} \geq \Delta$ and

$$\Phi\left[\frac{\left(\Delta - z_\alpha \sigma_{r_1 - r_2}\right) - \left(\rho_1 - \rho_2\right)}{\sigma_{r_1 - r_2}}\right] - \Phi\left[\frac{\left(-\Delta + z_\alpha \sigma_{r_1 - r_2}\right) - \left(\rho_1 - \rho_2\right)}{\sigma_{r_1 - r_2}}\right] \tag{55}$$

when $\left|z_\alpha \sigma_{r_1-r_2}\right| < \Delta$, where $\Phi[x]$ is the cdf of the normal distribution up to the point *x*.

If we choose $H_a : \rho_1 - \rho_2 = 0$, Equation (55) can be reduced to

$$2\Phi\left(\frac{|\Delta|}{\sigma_{r_1-r_2}} - z_\alpha\right) - 1. \tag{56}$$

Formula for required N when $N_1 = N_2$

As in traditional hypothesis testing, we need to keep a balance between Type I and Type II error and for a fixed α we might want to ascertain a minimum level of power by choosing a sample size large enough. Unfortunately, this is not straight forward for our equivalence test of the difference between two correlations. When trying to calculate a general formula for required N for given ρ_1 and ρ_2, notice that the function for power (Equation (55)) cannot simply be inverted: $\Phi^{-1}\left(\Phi\left(a\right) - \Phi\left(b\right)\right) \neq \left(a\right) - \left(b\right)$!

However, we can provide two less general formulas that can serve as guidelines. The first formula covers the situation when $H_a : \rho_1 - \rho_2 = 0$. Then the simplified power function from Equation (56) applies which can be inverted and the formula for required *N* in this special case will be:

$$N = \frac{\left(2\left(1-\rho^2\right)^2\right)\left(\Phi^{-1}\left(\frac{Power+1}{2}\right) + z_\alpha\right)^2}{\Delta^2} + 2 \tag{57}$$

(compare to Liu & Chow, 1992).

The second formula will allow us to find the *N* necessary to have estimated power at least minimally greater than zero for any combination of ρ_1 and ρ_2. In the discussion

above I mentioned that estimated power will be virtually zero whenever $\sigma_{r_1-r_2} z_\alpha \geq \Delta$. Thus, we need to find the *N* necessary to make the confidence interval narrow enough that it will potentially "fit into" the equivalence region. Solving for N_1, we get:

$$N_1 \geq \frac{\left(1-\rho_1^2\right)^2}{\left(\frac{\Delta}{z_\alpha}\right)^2 - \frac{\left(1-\rho_2^2\right)^2}{N_2-2}} + 2. \tag{58}$$

Estimated power will be equal to 0 for any N_1 smaller than the term on the right. Equivalently, we could solve for N_2.

In order to give potential users of correlational equivalence testing a guideline as to how large *N* needs to be to achieve desired power, I used a FindRoot routine in the mathematical package *Mathematica* to solve for required *N*. I varied both the values for the difference between the correlations, $\rho_1 - \rho_2$ and the average size of the correlations $\left(\rho_1 + \rho_2\right)/2$. A table for required *N* when desired power is equal .80 is provided.

Monte Carlo Analyses

We can construct a *T* statistic with known distribution as well as an exact interval with confidence bounds as in Equations (13) and (14) for the difference between two means. Such a traditional equivariant $1-2\alpha$ CI will have confidence coefficient $1-2\alpha$ and be unbiased (Casella and Berger, 2002, page 446; Berger & Hsu, 1996). The test statistics from Equations (48) and (49) and the confidence interval from Equation (50) for the difference between two correlations on the other hand are only approximations. Since $\hat{\rho}_1 - \hat{\rho}_2$ is only asymptotically normally distributed with mean $\rho_1 - \rho_2$ and standard

deviation $\sigma_{\rho_1-\rho_2}$, the confidence interval may be biased or not have confidence coefficient $1-2\alpha$. It seems advisable to assess the quality of performance of the estimate for the $1-2\alpha$ confidence interval using Monte Carlo analyses.

I will examine Type I error rate, power, coverage rate, and make an attempt at investigating bias. Further, I will provide tables with values for required *N*. The investigation of Type I error rate, power, and coverage rate is rather straight forward, and can be performed by simple counting. However, the examination of bias according to definition raises a problem since it would be difficult to show that the given interval does not cover any value more often than the true parameter. Therefore, I decided to measure coverage balance, the tendency of the confidence interval to miss the true parameter equally often above and below.

Unbiased two-sided equivariant confidence intervals for a statistic from a normal distribution miss the true parameter equally often on both sides. Although the equivalence of unbiasedness and missing equally often on both sides has not been established, it still makes sense intuitively that an unbiased CI for a normally distributed parameter should and will miss it equally often. Since the confidence interval from Equation (50) is for an asymptotically normally distributed statistic, showing that it misses the true parameter equally often above and below seems to add validity to the claim that it is unbiased. The presence of such coverage balance will be of interest to the applied statistician in its own right, even if it is not the same as finding unbiasedness. Hence, I counted the number of times the confidence interval missed the true parameter above and below.

Method

For the Monte Carlo simulations, I used the software package *Mathematica 5.2*, writing code to produce random correlations and the confidence interval, and count rejection rates for 100,000 replications. For each case, two sample correlations r_1 and r_2 from the underlying population with specified ρ_1 and ρ_2 were simulated. After the confidence interval was constructed, the program verified whether it was contained inside $[-\Delta, \Delta]$ and whether it contained the underlying true difference or was lying completely below or above $\rho_1 - \rho_2$. As a result, the program would return counts for the power function, coverage rate, missed below, and missed above.

The random number generator in Mathematica is a Marsaglia-Zaman subtract-with-borrow generator (e.g. Marsaglia and Zaman, 1991) for real numbers. The random number seed was 9164297.

Random Correlations:

A random correlation for a sample with $N = 1000$ and an underlying $\rho = 0.5$ can be obtained by creating 1000 random values from a bivariate normal distribution and computing the sample correlation for these values. Obviously, the larger *N*, the more time this will take, since values from random variables need to be drawn and the correlation between all values needs to be computed. There is another way to simulate random sample correlations. In order to accelerate the generation of randomly distributed sample correlations with sample size *n*, we can use the fact that a $p \times p$ matrix $\mathbf{S}$ of random

covariances for deviation scores, multiplied by *n*, has a *p*-variate Wishart distribution, more specifically, $\mathbf{S} \sim (1/n)\, W\left(\mathbf{I}, N-1\right)$ (see, e.g. Browne, 1968).

Let $\mathbf{S}$ be a $p \times p$ matrix of random covariances for samples of size *n*, then

$$E\left(\mathbf{S}\right) = \mathbf{\Sigma}, \tag{59}$$

where the ij^{th} element of $\mathbf{\Sigma}$ is σ_{ij}, $i = 1,..., p$ and $j = 1,..., p$. When $\mathbf{\Sigma} = \mathbf{I}$, the $p \times p$ identity matrix, $\mathbf{S}$ can be written as

$$\mathbf{S} = \frac{1}{n}\mathbf{T}\mathbf{T}'. \tag{60}$$

If we let $\mathbf{T}$ be a lower triangular $p \times p$ matrix, its diagonal elements are χ random variables with $n - i$ degrees of freedom and its off-diagonal elements are standard normal random variables. Hence, if we would like to produce random values for correlations between three variables that are uncorrelated with each other and have unit variance in the population with $n = 20$ each, we would only need to produce values for three χ random variables (with degrees of freedom 20, 19, and 18) and three normal random variables. Then $\left(1/n\right)\mathbf{T}\mathbf{T}'$ with

$$\mathbf{T} = \begin{bmatrix} t_{1,1} & 0 & 0 \\ t_{2,1} & t_{2,2} & 0 \\ t_{3,1} & t_{3,2} & t_{3,3} \end{bmatrix} \tag{61}$$

with $t_{1,1} \sim \chi_{20}$, $t_{2,2} \sim \chi_{19}$, $t_{3,3} \sim \chi_{18}$, and $t_{2,1} \sim t_{3,1} \sim t_{3,2} \sim N\left(0,1\right)$ would provide a matrix of random covariances where $n = 20$ and $\mathbf{\Sigma} = \mathbf{I}$.

However, if $\mathbf{\Sigma} \neq \mathbf{I}$, we can still use this result and produce random covariances in a way similar to the one just described. Let $\mathbf{CC}'$ be a Cholesky decomposition of $\mathbf{\Sigma}$ such that $\mathbf{CC}' = \mathbf{\Sigma}$. Then

$$\mathbf{S}^* = \frac{1}{n}\mathbf{CTT}'\mathbf{C}' \tag{62}$$

will be distributed as the Maximum Likelihood estimate of $\mathbf{\Sigma}$ with $E(\mathbf{S}*) = \mathbf{\Sigma}$. Random correlations ρ_{ij} can be obtained from random covariance σ_{ij} by dividing by the respective standard deviations σ_i and σ_j, taking the square root of the diagonal elements of $\mathbf{S}^*$. Utilizing this technique for the simulation of random correlations shortened the amount of time necessary for computation by a factor of five, on average.

Selection of Cases

In order to investigate empirical Type I error rate, I varied (1) Sample size (where $N_1 = N_2 = N$); (2) The size of the equivalence region, choosing values $\Delta = 0.05$, $\Delta = 0.1$, and $\Delta = 0.2$; (3) The average size of the correlations $(\rho_1 + \rho_2)/2$ and (4) Nominal Type I error rate ($\alpha = 0.05$ and $\alpha = 0.01$).Values of $\Delta = 0.05$ and $\Delta = 0.1$ seem loosely justified as 0.1 has been suggested as a small effect size for the difference between two correlations (Cohen, 1962), $\Delta = 0.2$ was included as well to have results for a larger range of values.

As mentioned above, I conducted Monte Carlo analyses to monitor the performance of the power formula from Equation (55). The following parameters were varied: (1) Sample size (same as above); (2) The average size of the correlations

$(\rho_1 + \rho_2)/2$ and (3) The difference between the correlations $\rho_1 - \rho_2$. The size of the equivalence region and nominal Type I error rate stayed fixed at $\Delta = 0.1$ and $\alpha = 0.05$, respectively.

For the investigation of coverage rate and the frequency with which the traditional equivariant $1-2\alpha$ confidence interval misses the true parameter below and above, I varied (1) Sample size (same as above); (2) The average size of the correlations $(\rho_1 + \rho_2)/2$ and (3) The difference between the correlations $\rho_1 - \rho_2$.

Results

Alpha:

Results for empirical alpha level are summarized in Tables 2 through 7. As one might have expected, the smaller Δ and α, the fewer (false) rejections. Further, several cells contain the value zero when *N* is small or $(\rho_1 + \rho_2)/2$ close to zero and thus $\sigma_{\hat{\rho}_1 - \hat{\rho}_2} z_\alpha$ is larger than Δ.

Table 2

Empirical Type I error rates when $\Delta = \rho_1 - \rho_2 = 0.05$ and $\alpha = 0.05$.

	$\frac{\rho_1 + \rho_2}{2}$									
Sample Size	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
20	0	0	0	0	0	0	0	0	0	.0018
40	0	0	0	0	0	0	0	0	0	.0049
60	0	0	0	0	0	0	0	0	0	.0143
80	0	0	0	0	0	0	0	0	0	.0297
100	0	0	0	0	0	0	0	0	0	.0440
125	0	0	0	0	0	0	0	0	0	.0542
150	0	0	0	0	0	0	0	0	0	.0580
175	0	0	0	0	0	0	0	0	.0001	.0579
200	0	0	0	0	0	0	0	0	.0004	.0592
250	0	0	0	0	0	0	0	0	.0045	.0565
300	0	0	0	0	0	0	0	0	.0165	.0564
350	0	0	0	0	0	0	0	0	.0313	.0575
400	0	0	0	0	0	0	0	0	.0404	.0568
500	0	0	0	0	0	0	0	.0010	.0498	.0557
600	0	0	0	0	0	0	0	.0124	.0502	.0554
750	0	0	0	0	0	0	.0001	.0354	.0503	.0550
1000	0	0	0	0	0	0	.0179	.0490	.0529	.0551
1250	0	0	0	0	0	.0059	.0384	.0501	.0512	.0532
1500	0	0	0	0	.0009	.0272	.0470	.0524	.0509	.0532
2000	0	0	.0010	.0171	.0324	.0446	.0498	.0493	.0509	.0541

Table 3

Empirical Type I error rates when $\Delta = \rho_1 - \rho_2 = 0.05$ and $\alpha = 0.01$.

	$\frac{\rho_1 + \rho_2}{2}$									
Sample Size	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
20	0	0	0	0	0	0	0	0	0	.0001
40	0	0	0	0	0	0	0	0	0	.0001
60	0	0	0	0	0	0	0	0	0	.0002
80	0	0	0	0	0	0	0	0	0	.0004
100	0	0	0	0	0	0	0	0	0	.0015
125	0	0	0	0	0	0	0	0	0	.0035
150	0	0	0	0	0	0	0	0	0	.0072
175	0	0	0	0	0	0	0	0	0	.0105
200	0	0	0	0	0	0	0	0	0	.0118
250	0	0	0	0	0	0	0	0	0	.0126
300	0	0	0	0	0	0	0	0	0	.0128
350	0	0	0	0	0	0	0	0	0	.0130
400	0	0	0	0	0	0	0	0	0	.0127
500	0	0	0	0	0	0	0	0	.0007	.0126
600	0	0	0	0	0	0	0	0	.0046	.0124
750	0	0	0	0	0	0	0	0	.0090	.0122
1000	0	0	0	0	0	0	0	.0001	.0110	.0119
1250	0	0	0	0	0	0	0	.0051	.0100	.0111
1500	0	0	0	0	0	0	0	.0095	.0103	.0117
2000	0	0	0	0	0	0	.0057	.0099	.0103	.0114

Table 4

Empirical Type I error rates when $\Delta = \rho_1 - \rho_2 = 0.1$ and $\alpha = 0.05$.

	$\frac{\rho_1 + \rho_2}{2}$									
Sample Size	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
20	0	0	0	0	0	0	0	0	.0023	.0834
40	0	0	0	0	0	0	0	0	.0053	.0872
60	0	0	0	0	0	0	0	.0001	.0166	.0808
80	0	0	0	0	0	0	0	.0003	.0347	.0765
100	0	0	0	0	0	0	0	.0014	.0486	.0734
125	0	0	0	0	0	0	0	.0069	.0555	.0714
150	0	0	0	0	0	0	.0001	.0188	.0570	.0699
175	0	0	0	0	0	0	.0006	.0324	.0575	.0681
200	0	0	0	0	0	0	.0032	.0415	.0574	.0671
250	0	0	0	0	0	.0001	.0204	.0494	.0580	.0647
300	0	0	0	0	0	.0049	.0365	.0525	.0562	.0635
350	0	0	0	0	.0005	.0218	.0445	.0538	.0550	.0640
400	0	0	0	0	.0092	.0338	.0490	.0523	.0545	.0627
500	0	0	.0020	.0165	.0336	.0457	.0507	.0523	.0546	.0608
600	.0162	.0186	.0252	.0344	.0439	.0477	.0503	.0520	.0536	.0590
750	.0366	.0382	.0420	.0470	.0488	.0500	.0506	.0520	.0529	.0586
1000	.0476	.0466	.0480	.0505	.0485	.0497	.0506	.0531	.0546	.0584
1250	.0497	.0492	.0491	.0502	.0494	.0512	.0515	.0515	.0526	.0560
1500	.0485	.0499	.0502	.0503	.0496	.0495	.0511	.0530	.0522	.0565
2000	.0487	.0502	.0495	.0505	.0498	.0501	.0505	.0498	.0523	.0564

Table 5

Empirical Type I error rates when $\Delta = \rho_1 - \rho_2 = 0.1$ and $\alpha = 0.01$.

	$\frac{\rho_1 + \rho_2}{2}$									
Sample Size	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
20	0	0	0	0	0	0	0	0	.0001	.0258
40	0	0	0	0	0	0	0	0	.0001	.0288
60	0	0	0	0	0	0	0	0	.0002	.0262
80	0	0	0	0	0	0	0	0	.0007	.0246
100	0	0	0	0	0	0	0	0	.0019	.0221
125	0	0	0	0	0	0	0	0	.0046	.0207
150	0	0	0	0	0	0	0	0	.0090	.0203
175	0	0	0	0	0	0	0	0	.0111	.0196
200	0	0	0	0	0	0	0	.0001	.0130	.0182
250	0	0	0	0	0	0	0	.0013	.0126	.0167
300	0	0	0	0	0	0	0	.0057	.0128	.0161
350	0	0	0	0	0	0	0	.0097	.0122	.0162
400	0	0	0	0	0	0	.0005	.0102	.0122	.0161
500	0	0	0	0	0	0	.0062	.0105	.0117	.0156
600	0	0	0	0	0	.0010	.0088	.0109	.0113	.0150
750	0	0	0	0	.0006	.0078	.0101	.0107	.0110	.0142
1000	0	0	.0005	.0053	.0082	.0100	.0107	.0115	.0119	.0137
1250	.0062	.0069	.0080	.0092	.0095	.0094	.0111	.0105	.0109	.0126
1500	.0087	.0091	.0096	.0100	.0102	.0099	.0100	.0107	.0110	.0132
2000	.0095	.0101	.0099	.0105	.0104	.0106	.0105	.0102	.0110	.0126

Table 6

Empirical Type I error rates when $\Delta = \rho_1 - \rho_2 = 0.2$ and $\alpha = 0.05$.

	$\frac{\rho_1 + \rho_2}{2}$									
Sample Size	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.85
20	0	0	0	0	0	0	.0015	.0157	.0807	.1146
40	0	0	0	0	0	.0004	.0066	.0422	.0811	.0920
60	0	0	0	0	.0002	.0030	.0267	.0592	.0751	.0832
80	0	0	0	0	.0018	.0177	.0462	.0603	.0722	.0790
100	0	0	0	.0010	.0123	.0370	.0530	.0619	.0698	.0757
125	0	.0001	.0034	.0167	.0359	.0477	.0549	.0589	.0677	.0720
150	.0148	.0168	.0248	.0344	.0444	.0517	.0557	.0583	.0663	.0719
175	.0319	.0336	.0378	.0441	.0487	.0538	.0547	.0580	.0653	.0700
200	.0402	.0415	.0438	.0473	.0501	.0515	.0538	.0572	.0639	.0678
250	.0475	.0474	.0497	.0497	.0511	.0515	.0541	.0561	.0634	.0657
300	.0493	.0492	.0496	.0509	.0504	.0511	.0535	.0568	.0620	.0636
350	.0500	.0501	.0500	.0500	.0517	.0512	.0523	.0563	.0602	.0635
400	.0485	.0499	.0488	.0507	.0523	.0512	.0528	.0548	.0593	.0626
500	.0489	.0490	.0500	.0496	.0516	.0512	.0526	.0554	.0583	.0600
600	.0506	.0507	.0498	.0498	.0507	.0499	.0514	.0548	.0566	.0602
750	.0495	.0490	.0501	.0519	.0502	.0506	.0520	.0537	.0563	.0586
1000	.0498	.0484	.0492	.0511	.0490	.0503	.0517	.0552	.0574	.0556
1250	.0500	.0492	.0493	.0499	.0500	.0514	.0518	.0528	.0542	.0564
1500	.0483	.0497	.0495	.0499	.0495	.0499	.0517	.0535	.0540	.0549
2000	.0488	.0501	.0495	.0503	.0501	.0509	.0509	.0514	.0538	.0550

Table 7

Empirical Type I error rates when $\Delta = \rho_1 - \rho_2 = 0.2$ and $\alpha = 0.01$.

	$\frac{\rho_1 + \rho_2}{2}$									
Sample Size	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.85
20	0	0	0	0	0	0	.0001	.0013	.0229	.0513
40	0	0	0	0	0	0	0	.0028	.0256	.0375
60	0	0	0	0	0	0	.0002	.0073	.0240	.0299
80	0	0	0	0	0	0	.0009	.0125	.0221	.0271
100	0	0	0	0	0	.0001	.0037	.0139	.0200	.0244
125	0	0	0	0	0	.0005	.0087	.0135	.0186	.0228
150	0	0	0	0	.0001	.0034	.0112	.0136	.0189	.0213
175	0	0	0	0	.0009	.0079	.0112	.0138	.0174	.0205
200	0	0	0	.0001	.0034	.0099	.0116	.0135	.0175	.0193
250	0	0	.0011	.0054	.0091	.0109	.0119	.0127	.0165	.0176
300	.0049	.0053	.0071	.0090	.0106	.0110	.0118	.0133	.0159	.0172
350	.0088	.0088	.0093	.0095	.0100	.0106	.0113	.0133	.0148	.0169
400	.0096	.0095	.0093	.0101	.0101	.0102	.0115	.0122	.0147	.0161
500	.0094	.0098	.0101	.0097	.0106	.0103	.0109	.0120	.0135	.0160
600	.0103	.0097	.0098	.0099	.0102	.0100	.0107	.0120	.0131	.0153
750	.0100	.0094	.0099	.0101	.0091	.0105	.0106	.0116	.0131	.0145
1000	.0099	.0097	.0097	.0101	.0096	.0102	.0113	.0125	.0131	.0140
1250	.0099	.0101	.0099	.0098	.0099	.0098	.0113	.0112	.0122	.0129
1500	.0094	.0098	.0098	.0099	.0101	.0102	.0105	.0114	.0121	.0132
2000	.0094	.0101	.0099	.0104	.0105	.0107	.0108	.0108	.0119	.0132

Type I error rate is controlled satisfactorily for almost all cases when $\Delta = 0.05$ (Tables 2 and 3). For larger Δ, the test becomes liberal for the difference between two large correlations (see, e.g. Tables 6 and 7). For example, when $\Delta = 0.2$, $\alpha = 0.01$ and $(\rho_1 + \rho_2)/2 = 0.85$, the proportion of falsely rejected non-equivalencies can be 0.0513, more than five times the nominal Type I error rate. Note that here, $\rho_1 = 0.95$ and $\rho_2 = 0.75$, that is, one correlation is very close to one.

Power

Figures *2* through *6* are power graphs for $\Delta = 0.1$ and $\alpha = .05$ and provide a comparison between power values calculated using the formula from Equation (55) and power values obtained with Monte Carlo simulations. Continued lines represent power calculated with Equation (55) and symbols not connected with a line represent Monte Carlo results. Symbols for Monte Carlo results lying on the continued lines for estimated power indicate good performance of Equation (55). Overall, a good fit between the values given by the formula and the Monte Carlo results can be found, suggesting that the normal approximation worked reasonably well. When *N* is small, minor deviations can be observed, especially for large correlations, as in Figures 5 and *6*, for example, where power values generated by the formula tend to underestimate or overestimate true power slightly.

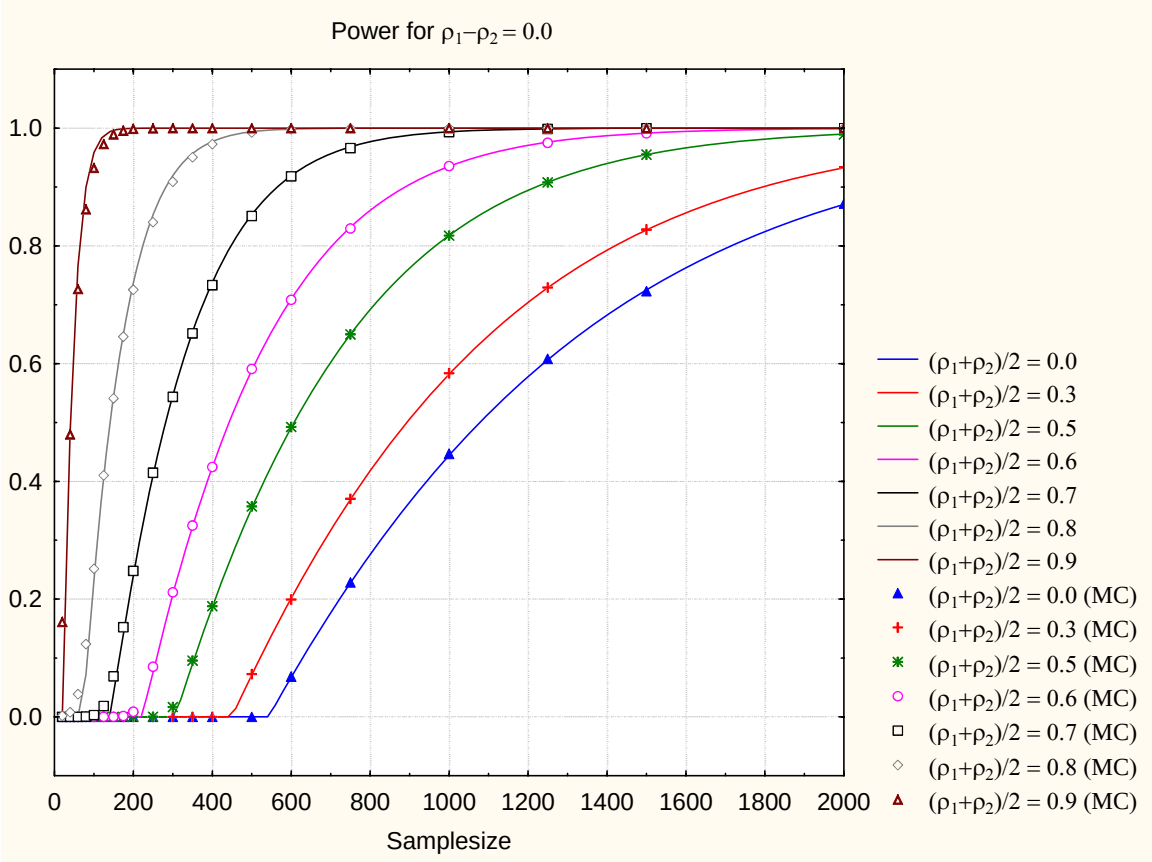

Figure 2. Power graph for testing correlational equivalence when $\rho_1 - \rho_2 = 0$ and $(\rho_1 + \rho_2) = 0.0, 0.3, 0.5 - 0.9$, including data points from Monte Carlo analyses for comparison

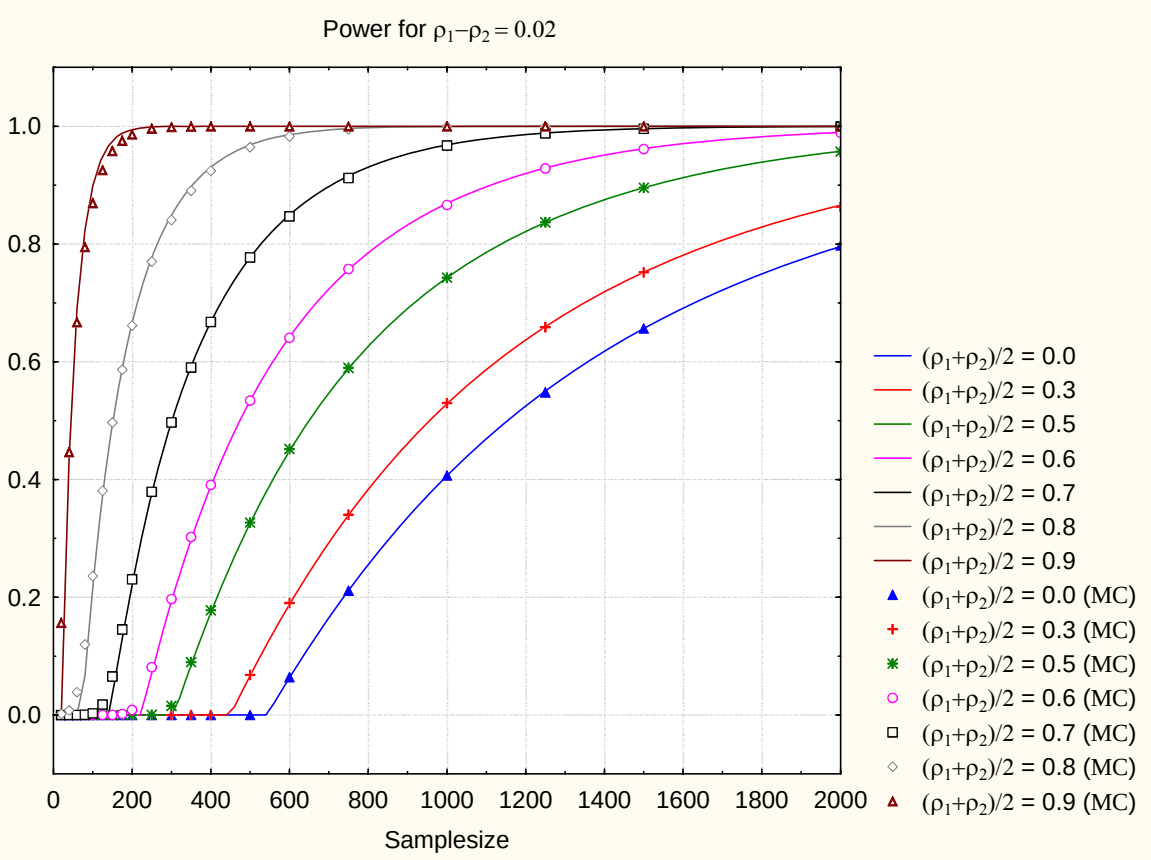

Figure 3. Power graph for testing correlational equivalence when $\rho_1 - \rho_2 = 0.2$ and $(\rho_1 + \rho_2) = 0.0, 0.3, 0.5 - 0.9$, including data points from Monte Carlo analyses for comparison

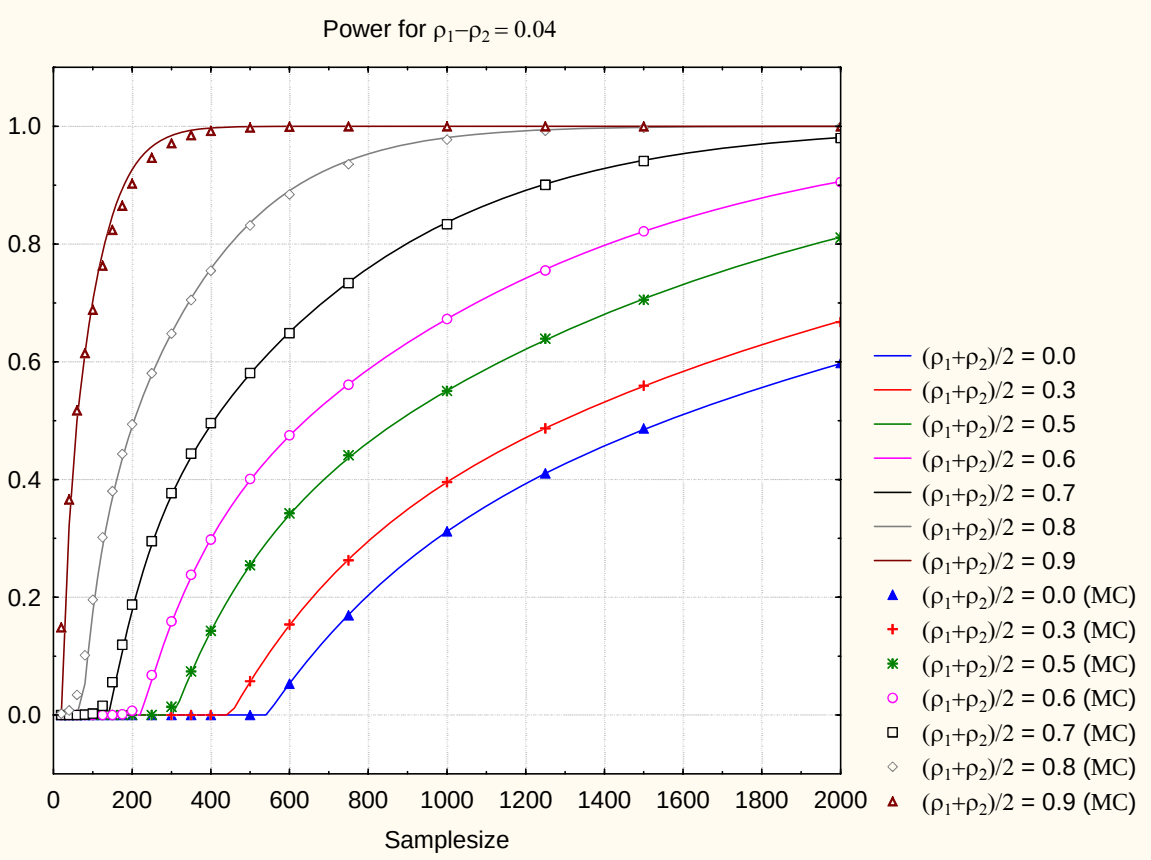

Figure 4. Power graph for testing correlational equivalence when $\rho_1 - \rho_2 = 0.4$ and $(\rho_1 + \rho_2) = 0.0, 0.3, 0.5 - 0.9$, including data points from Monte Carlo analyses for comparison

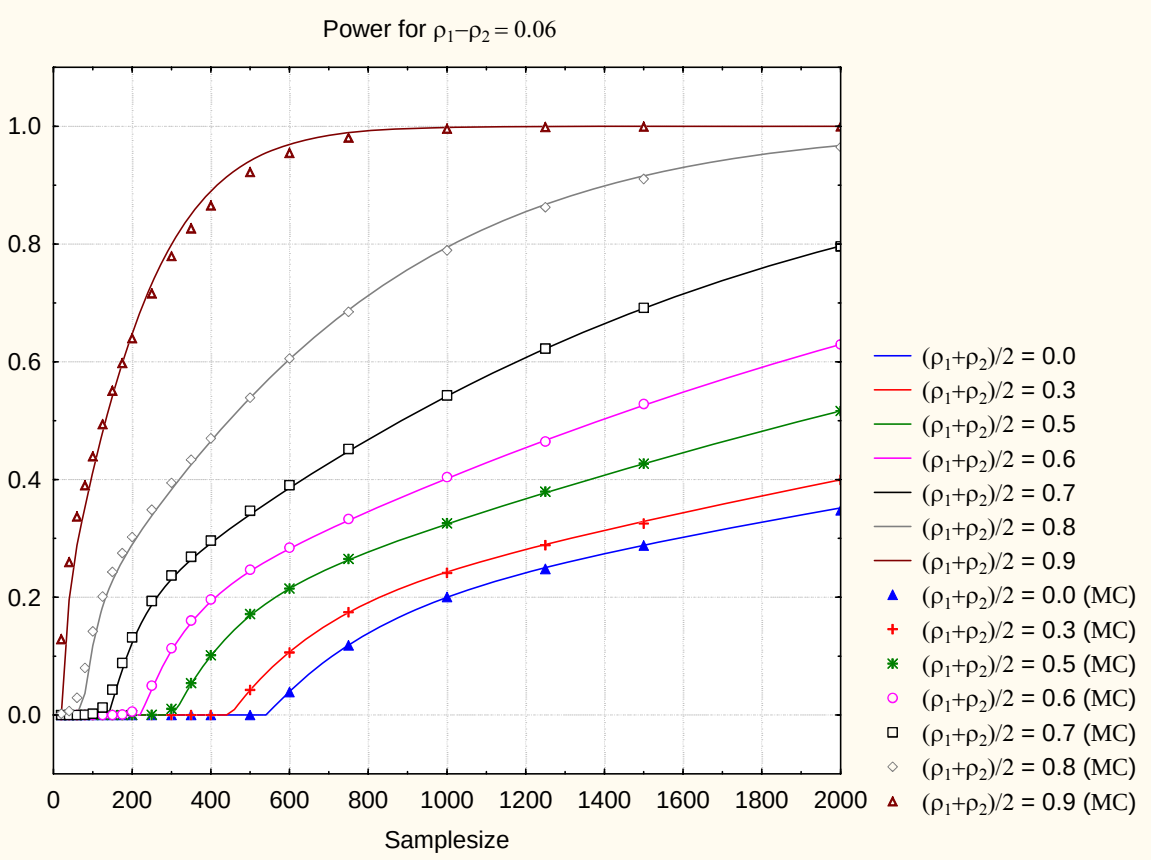

Figure 5. Power graph for testing correlational equivalence when $\rho_1 - \rho_2 = 0.6$ and $(\rho_1 + \rho_2) = 0.0, 0.3, 0.5 - 0.9$, including data points from Monte Carlo analyses for comparison

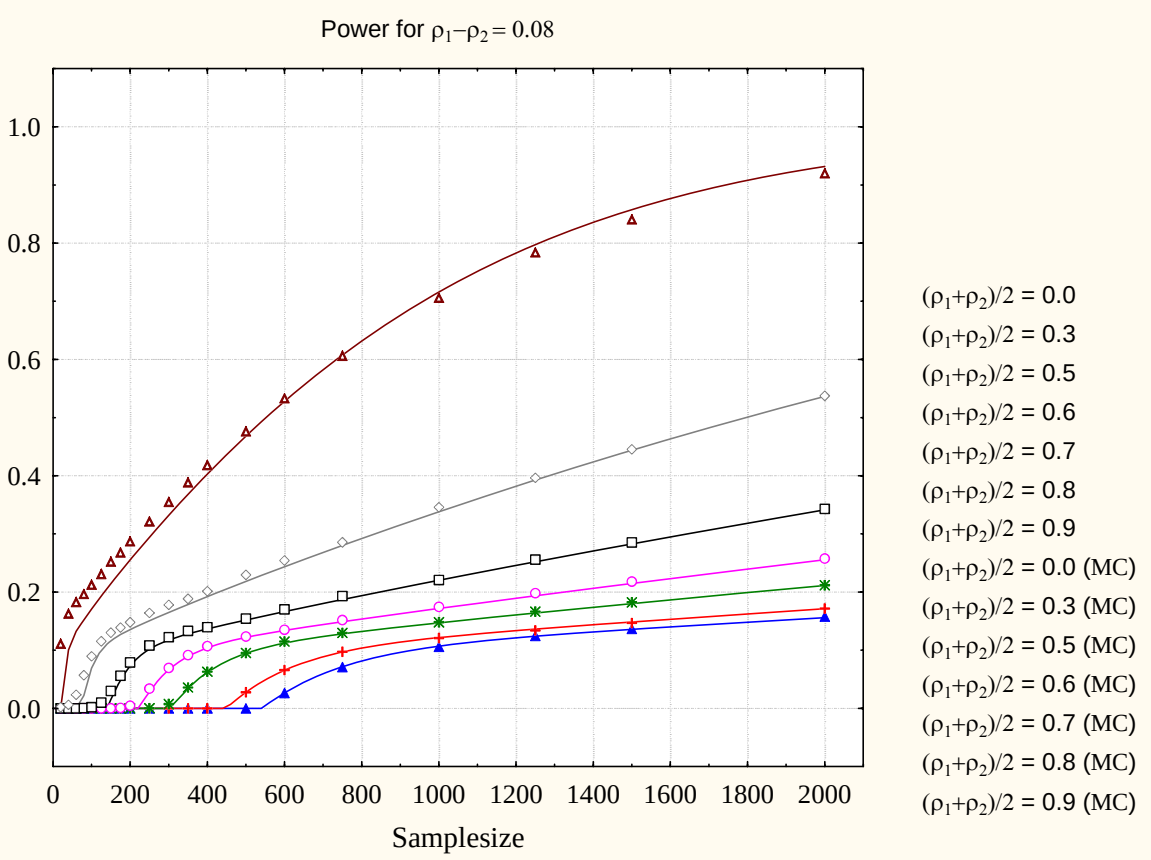

Figure 6. Power graph for testing correlational equivalence when $\rho_1 - \rho_2 = 0.8$ and $(\rho_1 + \rho_2) = 0.0, 0.3, 0.5 - 0.9$, including data points from Monte Carlo analyses for comparison

Whether a researcher can expect good power when conducting correlational equivalence testing depends on sample size, but to a large extent on the size of the correlations tested as well. Assuming that the true difference between two correlations is zero, if both correlations are small, we will need large sample sizes to achieve a power of .80 for detecting this equivalence, a value that has often been described as a lower limit for satisfactory power. On the other hand, to have a good chance of discovering a true difference of zero between two large correlations does by far not require as large a sample size.

In general, we have to realize that with sample sizes common in psychological research, there only is a realistic chance of (correctly!) finding a negligible difference between two correlations when this difference is small (0.0 – 0.04) and both correlations are fairly large $\left(\left(\rho_1 + \rho_2\right)/2 \geq 0.7\right)$.

Required N

Table 8 provides values for required *N* when the desired power is equal to .80. Required sample size could not be found analytically (see above), however, it was possible to utilize a root finding function in *Mathematica*. The FindRoot routine provided by *Mathematica* did not converge properly at the first attempt in several cases and it was necessary to try out different starting values for *N*.

Required *N* obviously mirror results from the power calculations. With large correlations and small differences between the two correlations, the *N* necessary to ascertain satisfactory power can be quite reasonable and even small. However, other

combinations, e.g. small correlations and larger differences between them, can lead to astronomical numbers for required *N*.

Table 8

Required *N* for testing correlational equivalence with power $= .80$, $\Delta = .10$, and $\alpha = .05$

		$(\rho_1 + \rho_2)/2$						
		0	.1	.2	.3	.4	.5	.55
$\rho_1 - \rho_2$	0	1715	1681	1581	1421	1211	966	836
	.005	1731	1697	1596	1434	1222	975	844
	.01	1782	1747	1643	1476	1258	1004	868
	.015	1875	1838	1728	1553	1324	1056	913
	.02	2021	1981	1863	1674	1427	1138	985
	.025	2236	2192	2061	1852	1579	1259	1089
	.03	2535	2485	2337	2100	1790	1428	1235
	.035	2929	2871	2700	2426	2068	1649	1427
	.04	3435	3366	3166	2845	2425	1934	1673
	.045	4086	4005	3766	3384	2885	2301	1991
	.05	4942	4844	4555	4094	3489	2783	2408
	.055	6099	5978	5622	5052	4306	3435	2972
	.06	7717	7563	7113	6392	5448	4346	3761
	.065	10075	9875	9286	8346	7114	5675	4911
	.07	13708	13435	12635	11355	9679	7722	6684
	.075	19731	19339	18187	16345	13933	11117	9623
	.08	30816	30204	28405	25529	21763	17366	15033
	.085	54760	53672	50475	45367	38676	30866	26721
	.09	123153	120707	113519	102032	86991	69432	60114
	.095	492377	482602	453868	407957	347837	277662	240426

Table 8 continued

Required *N* for testing correlational equivalence with power $= .80$, $\Delta = .10$, and $\alpha = .05$

		$(\rho_1 + \rho_2)/2$							
		.6	.65	.7	.75	.8	.85	.9	.95
$\rho_1 - \rho_2$	0	704	574	448	330	224	134	64	19
	.005	711	579	452	333	227	136	65	19
	.01	732	596	466	343	233	140	67	20
	.015	770	627	490	361	245	147	70	21
	.02	829	676	528	389	265	158	76	22
	.025	918	748	584	431	293	175	84	25
	.03	1040	848	662	488	332	199	96	29
	.035	1202	980	765	564	384	230	111	33
	.04	1410	1149	897	662	450	270	130	40
	.045	1677	1367	1068	788	536	322	156	48
	.05	2029	1654	1292	953	649	391	290	60
	.055	2504	2041	1595	1178	802	484	236	76
	.06	3169	2584	2019	1491	1017	614	301	100
	.065	4139	3375	2638	1949	1330	805	397	135
	.07	5633	4594	3592	2655	1814	1100	547	191
	.075	8111	6616	5175	3828	2618	1591	796	286
	.08	12672	10339	8089	5987	4100	2499	1260	465
	.085	22529	18384	14390	10658	7308	4466	2270	864
	.09	50691	41375	32400	24015	16489	10110	5183	2034
	.095	202773	165555	129700	96210	66159	40698	21052	8518

To test the accuracy of these results, I reentered the results for required *N* from Table 8 into Monte Carlo code for finding power (the same code that has been used for testing the performance of the power formula), expecting power values around .80. For small to medium *N*, it seemed that the empirical power was somewhat lower than expected (the lowest being .70), suggesting that the actual required *N* are a little larger than what is given in Table 8. This might indicate that the approximation to a power formula from Equation (55) with a simple formula for $s_{r_1-r_2}$ is not optimal when *N* is small.

Coverage Rate and Coverage Balance:

Figures 7 through 11 and Tables 9 and 10 summarize results for coverage rate and coverage balance. Tables 9 and 10 contain results for the smallest sample size $\left(N_1 = N_2 = 20\right)$ for all combinations. Only a list of representative cases, highlighted yellow in Table 9, is displayed in the graphs. Further, coverage rate was subtracted from 1 to give what might be called a "miss-rate". While information on miss rates above and below the true parameter are kept separate and not combined into one index in the figures, I computed the ratio of miss rate below/miss rate above as an index of coverage balance in the tables:

$$\text{Balance Index} = \frac{\text{Miss Rate Below}}{\text{Miss Rate Above}}. \tag{63}$$

A value close to 1 stands for a "balanced" CI, while values diverging from 1 stand for coverage imbalance.

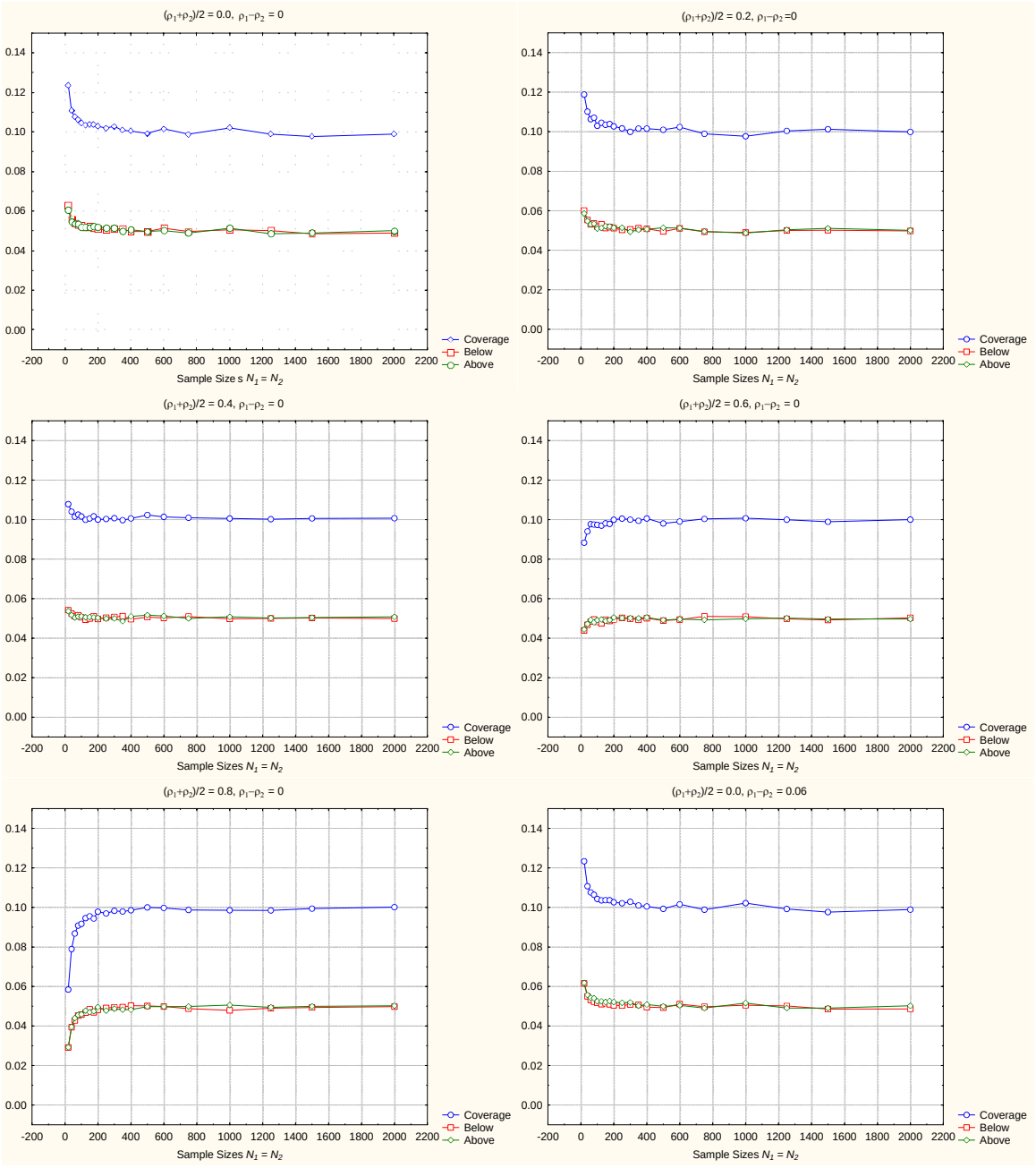

Figure 7. Representative graphs of Monte Carlo analyses of 1 – coverage rate and coverage balance for correlational equivalence testing. Cells 1 through 5: $\rho_1 - \rho_2 = 0.0$ and $(\rho_1 + \rho_2) = 0.0, 0.2, 0.4, 0.6, 0.8$, cell 6: $\rho_1 - \rho_2 = 0.06$ and $(\rho_1 + \rho_2) = 0.0$

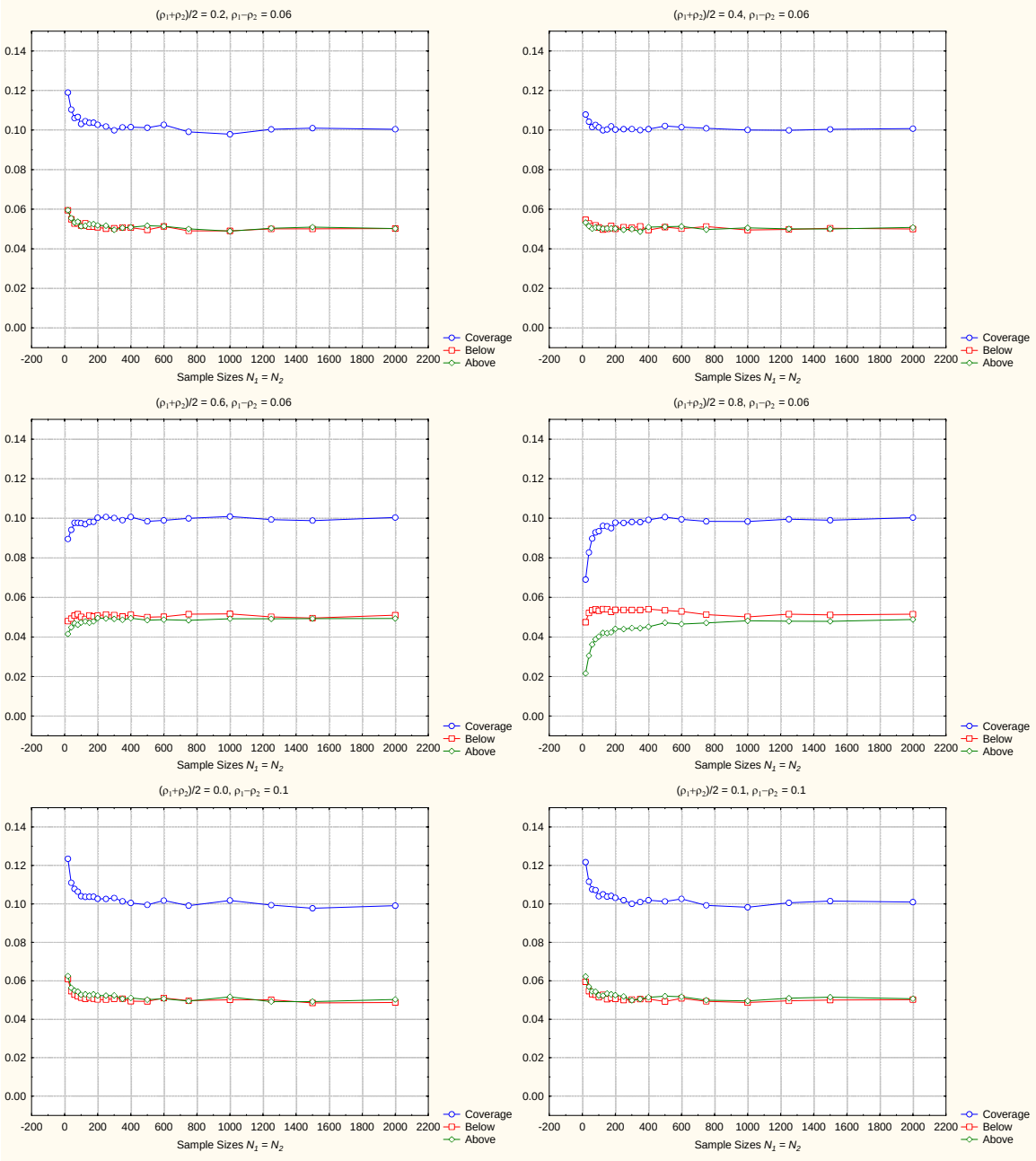

Figure 8. Representative graphs of Monte Carlo analyses of 1 – coverage rate and coverage balance for correlational equivalence testing. Cells 1 through 4: $\rho_1 - \rho_2 = 0.06$ and $(\rho_1 + \rho_2) = 0.2, 0.4, 0.6, 0.8$, cell 5 and 6: $\rho_1 - \rho_2 = 0.1$ and $(\rho_1 + \rho_2) = 0.0, 0.1$

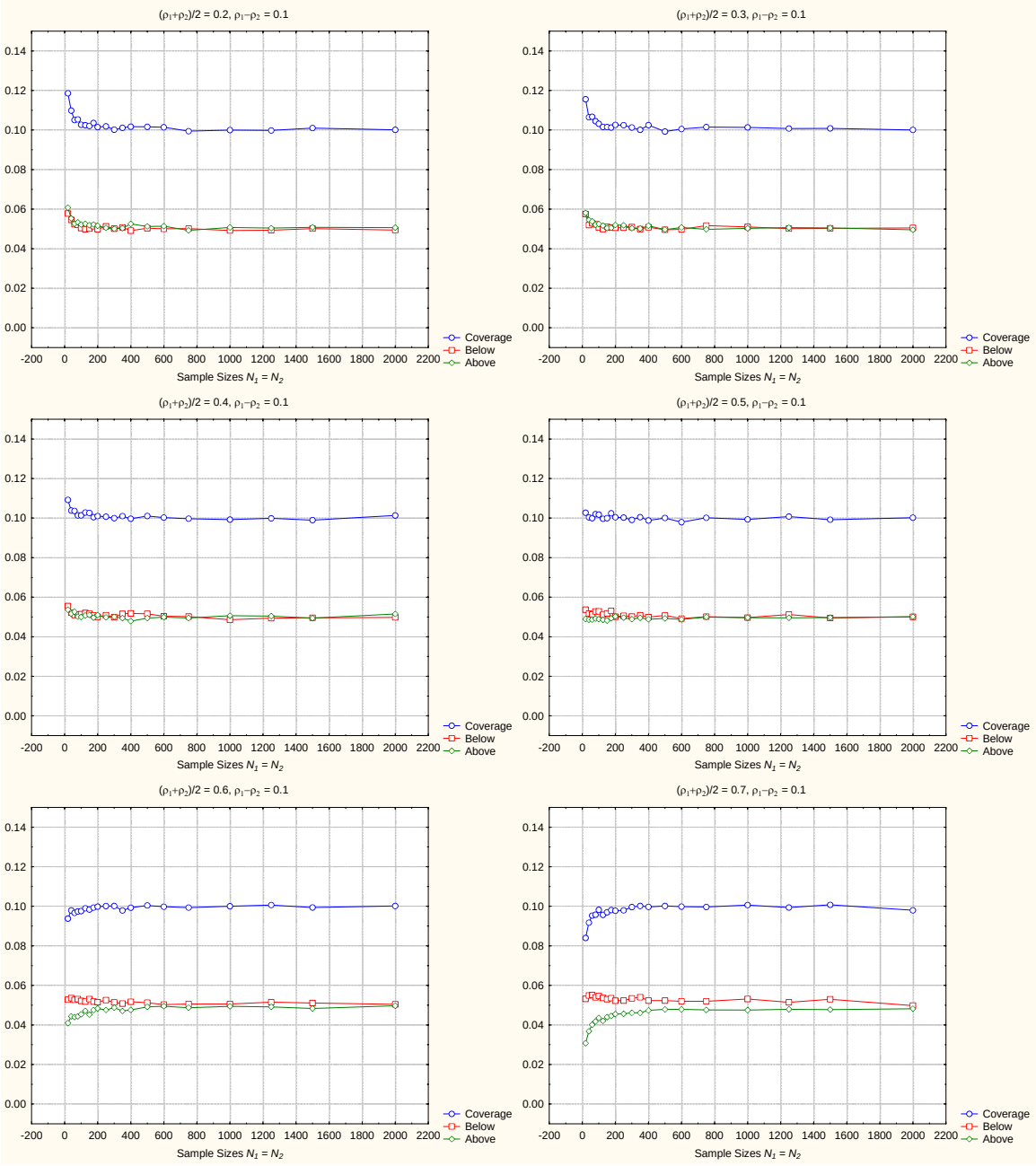

Figure 9. Representative graphs of Monte Carlo analyses of 1 – coverage rate and coverage balance for correlational equivalence testing. Cells 1 through 6: $\rho_1 - \rho_2 = 0.1$ and $(\rho_1 + \rho_2) = 0.2 - 0.7$

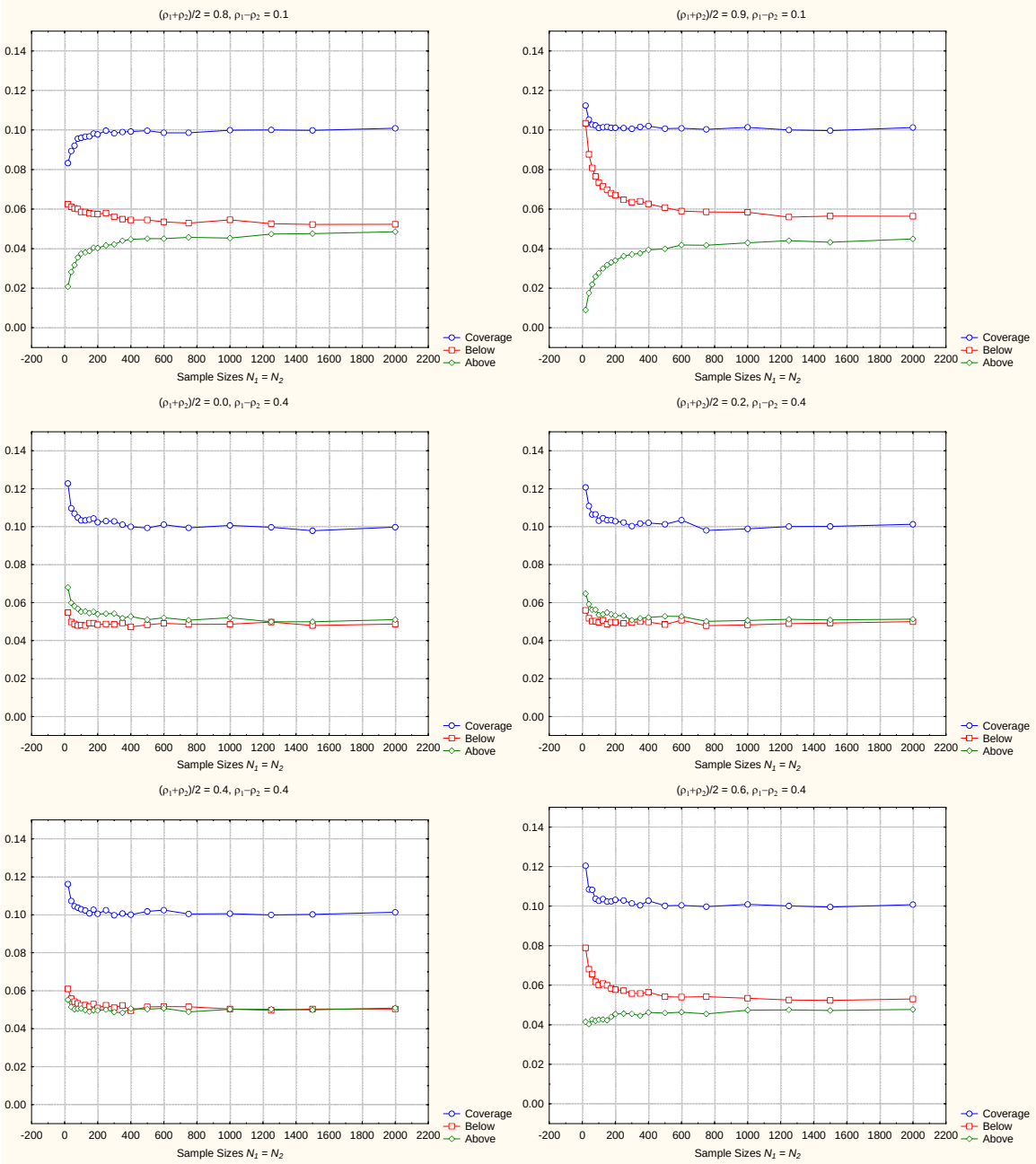

Figure 10. Representative graphs of Monte Carlo analyses of 1 – coverage rate and coverage balance for correlational equivalence testing. Cells 1 and 2: $\rho_1 - \rho_2 = 0.1$ and $(\rho_1 + \rho_2) = 0.8, 0.9$, cell 3 through 6: $\rho_1 - \rho_2 = 0.4$ and $(\rho_1 + \rho_2) = 0.0, 0.2, 0.4, 0.6$

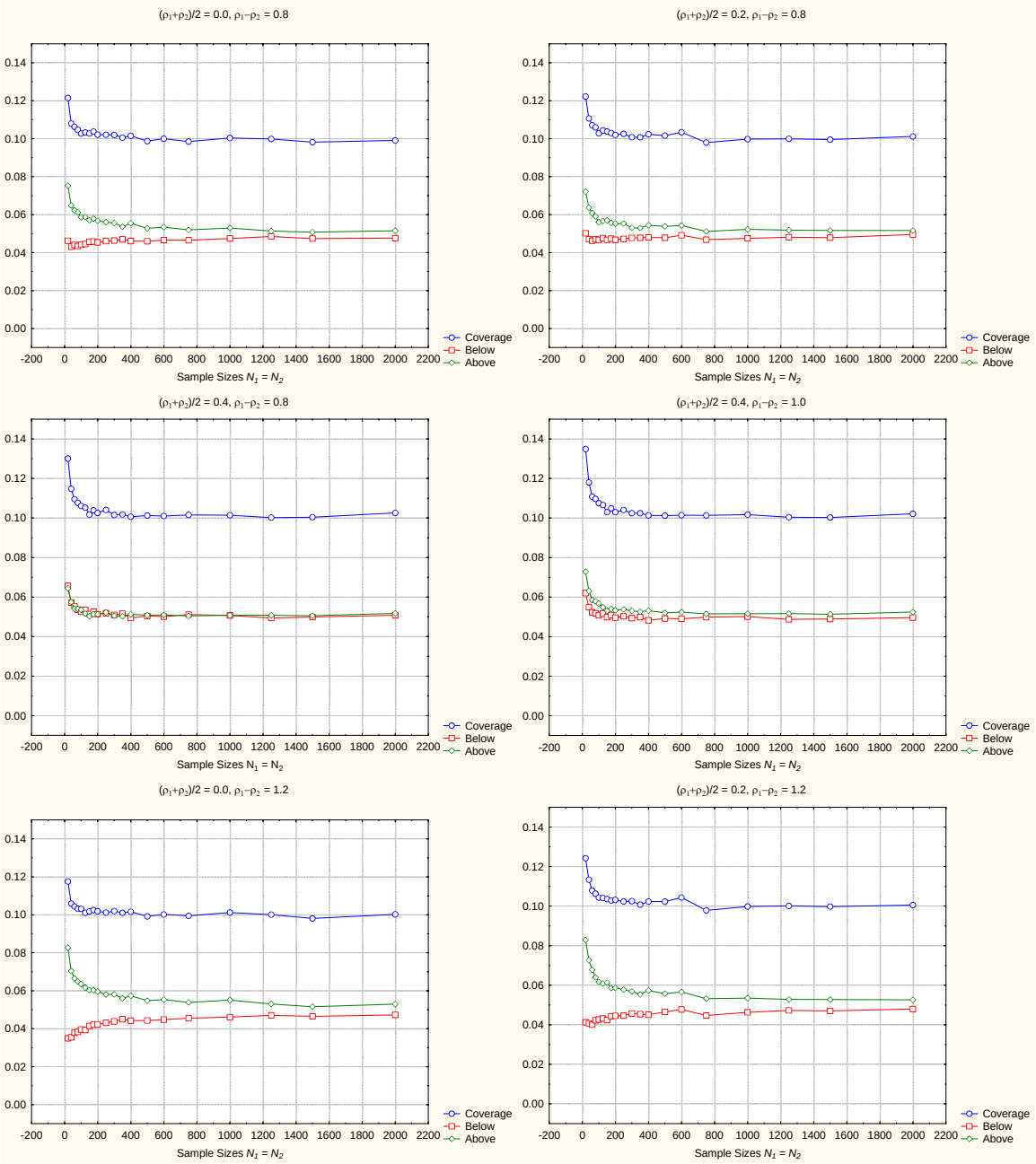

Figure 11. Representative graphs of Monte Carlo analyses of 1 – coverage rate and coverage balance for correlational equivalence testing. Cell 1: $\rho_1 - \rho_2 = 0.6$ and $(\rho_1 + \rho_2)/2 = 0.6$, cell 2 and 3: $\rho_1 - \rho_2 = 0.8$ and $(\rho_1 + \rho_2)/2 = 0.0, 0.2$, cell 4: $\rho_1 - \rho_2 = 1.0$ and $(\rho_1 + \rho_2)/2 = 0.4$, cell 5 and 6: $\rho_1 - \rho_2 = 1.2$ and $(\rho_1 + \rho_2)/2 = 0.0, 0.2$

Table 9

1 – Coverage rate (top value) and coverage balance (bottom value) for $N_1 = N_2 = 20$. Combinations of $(\rho_1 + \rho_2)/2$ and $\rho_1 - \rho_2$ in yellow cells are also displayed in Figures 7 through 11.

	$\frac{\rho_1 + \rho_2}{2}$										
$\rho_1 - \rho_2$		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	0	0.1235 1.0415		0.1188 1.0258		0.1078 1.0097		0.0883 0.9892		0.0584 0.9966	
	0.02	0.1235 1.0238		0.1188 1.0166		0.1079 1.0183		0.0884 1.0411		0.0598 1.3288	
	0.04	0.1234 1.0127		0.1186 1.0090		0.1078 1.0219		0.0888 1.0960		0.0638 1.7302	
	0.06	0.1234 0.9956		0.1190 0.9995		0.1079 1.0259		0.0895 1.1563		0.0690 2.1963	
	0.08	0.1235 0.9866		0.1192 0.9940		0.1084 1.0361		0.0904 1.2057		0.0749 2.6855	
	0.1	0.1234 0.9755	0.1217 0.9563	0.1186 0.9511	0.1156 0.9890	0.1092 1.0343	0.1028 1.0961	0.0938 1.2890	0.0840 1.7298	0.0833 3.0029	0.1124 11.5531
	0.2	0.1235 0.9174		0.1195 0.9420		0.1098 1.0760		0.1014 1.5243		0.1161 5.0648	
	0.4	0.1228 0.8057		0.1207 0.8633		0.1162 1.1042		0.1204 1.9012			
	0.6	0.1223 0.7117		0.1211 0.7812		0.1234 1.0891		0.1347 1.8490			
	0.8	0.1215 0.6148		0.1223 0.6956		0.1301 1.0186					
	1.0	0.1203 0.5183		0.1227 0.5996		0.1349 0.8516					
	1.2	0.1176 0.4228		0.1243 0.4975							

Table 10

Summary of coverage rate and coverage balance when $N_1 = N_2 = 20$. Minuses stand for empirical coverage below the nominal value, pluses for coverage above the nominal value. Blue cells indicate combinations where the CI misses the true parameter more often by lying above it, reddish cells indicate combination where the CI misses the true parameter more often by lying below it. Color intensity indicates severity of coverage imbalance.

	$\frac{\rho_1 + \rho_2}{2}$										
$\rho_1 - \rho_2$		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	0	–		–				+		+	
	0.02	–		–				+		+	
	0.04	–		–				+		+	
	0.06	–		–				+		+	
	0.08	–		–						+	
	0.1	–	–	–	–		+		+	+	–
	0.2	–		–						–	
	0.4	–		–		–		–			
	0.6	–		–		–		–			
	0.8	–		–		–					
	1.0	–		–		–					
	1.2	–		–							

Construction of Table 10: The definition of a “medium or large deviation from nominal coverage rate” is of course somewhat arbitrary. I chose the following guidelines: Coverage rate shows a medium deviation from nominal coverage rate when the proportion of covered true parameters ranges from 0.91 to 0.925 (high coverage rate, symbolized by a small +) or 0.89 to 0.875 (low coverage rate, symbolized by a small –). Extreme deviations from nominal coverage rate are encountered when the proportion of covered parameters is above 0.925 (large +) or below 0.875 (large –).

To find decision rules as to what constitutes medium and large deviations from missing the true parameter equally often above and below, one may consider that a ratio of 3 : 2 means that 60 percent of the misses the confidence interval was lying entirely below the true parameter and 40 percent of the misses it was lying above. A ratio of 3 : 2 seemed reasonable to me as a lower bound for the risk of systematically under- or overestimating the true parameter. Hence, values for ratios from Equation (63) ranging from 1.5 to 3.0 and from 0.67 to 0.33 indicate medium coverage imbalance, while values greater than 3 or smaller than 0.33 indicate strong imbalance (see Table 10). Applying these guidelines, I summarized coverage rate and balance results in Table 10 to give an overview as to which combinations might lead to high or low coverage rate and a miss rate ratio far from 1 using visual cues such as color.

Coverage rate for small sample sizes showed almost always some deviation (too high or too low) from the nominal level. For the 90% confidence interval, coverage rate could be as high as 94% (Figure 11, cell 5) or as low as 86.5% (Figure 11, cell 4). While no or very little coverage imbalance was found in most cases, substantial imbalance was observed in some; an overview of combinations that led to medium and severe imbalance

can be gained from Table 10 and the figures. As *N* increases, coverage rate and balance approximate their nominal values.

Conclusions

Correlational equivalence testing using the approximate $1 - 2\alpha$ confidence interval provides satisfactory Type I error control at the .05 level for a wide range of correlations. When both correlations are fairly large (e.g., $(\rho_1 + \rho_2)/2 = 0.8$), however, the tests can be too liberal. These results are repeated at the $\alpha = .01$ level, where large correlations again lead to overly high rejection rates. For small sample sizes, the empirical rejection rate was up to five times as high as the nominal α-level. This turns out to make correlational equivalence testing utilizing the approximate procedures developed above less feasible, as power rates for small or medium sample sizes tend to be acceptable only when both correlations are large. Testing differences between two small correlations will require sample sizes that are rarely found in psychological research.

The results for coverage rate and coverage balance seem to justify the simple approximations utilized in this study somewhat. Except for extreme combinations with one correlation close to 1, empirical coverage rate does not diverge too strongly from nominal coverage rate and the $1 - 2\alpha$ confidence interval misses the true parameter about equally often on both sides. Improvements on this performance might be easy to achieve with more accurate formulas for the standard error of $r_1 - r_2$.

CHAPTER IV

DISCUSSION

Several procedures for testing the difference between two parameters for equivalence have been suggested in the past 35 years. As there has been a considerable amount of dispute which procedures are useful and valid, it seems necessary to investigate their characteristics and performance carefully to be able to choose an appropriate method for the research question at hand. I have attempted to compare three procedures that utilize confidence intervals for equivalence testing. An application of the traditional equivariant $1-2\alpha$ confidence interval to correlational equivalence testing with a simple approximation to the standard error of $\rho_1 - \rho_2$ was examined in Monte Carlo simulations.

Do all Confidence Procedures perform equally well?

Confidence intervals have been suggested as a replacement for, or at least addition to, significance tests for a long time and by a large number of sources (see, e.g. Wilkinson & The APA Task Force on Statistical Inference, 1999), as they add information regarding the parameter in form of a range of reasonable values. Typically, using a confidence interval will result in the same conclusion regarding the hypotheses at hand that would have been reached had a significance test been used. It seems that the traditional *reject-support* testing situation that taught introductory statistics classes poses a convenient coincidence: An α-level test will arrive at the same conclusion as the

appropriate decision rule utilizing a $1-\alpha$ confidence interval. This correspondence need not always be the case! For the test of equivalence for two parameters, an equivariant $1-2\alpha$ confidence interval will reach the same conclusion as an α- level test (e.g. Berger & Hsu, 1996). Other authors have mentioned that the lack of a complete correspondence between the α- level of a test and the size of a confidence interval might be too confusing for the average consumer of statistical analyses (Westlake, 1976). However, in the pharmaceutical sciences, familiarization with the issue has long taken place. I believe that the social sciences would also be well advised to ponder this non-correspondence. There have been numerous efforts to keep this correspondence, but properties of new procedures always need to be carefully examined.

Comparing the intervals that are available on the main properties coverage rate, bias, and width enables us to make better informed decisions as to which procedure is most helpful. Foremost of all, I find that the replacement procedure should arrive at the same conclusion as the original procedure. Out of the three procedures investigated, Westlake's symmetric interval was not equivalent to the hypothesis tests it was supposed to replace. When the true parameter is close to zero, i.e. when practical equivalence is given, his symmetric interval will be too conservative. The two other intervals, Seaman and Serlin's and the traditional $1-2\alpha$ confidence interval when used to test Equations (4), will arrive at the same conclusions as the TOST. For all three intervals, satisfactory coverage rate has been proven or observed, however, both Westlake's and Seaman and Serlin's interval are dramatically biased by always including zero.

The bias of a confidence interval has not received a lot of attention in the past and most Monte Carlo studies examining confidence interval performance report coverage

rates but no measure of bias. One possible reason might be that it has only recently become computationally feasible to investigate bias as in Definition 4 in a Monte Carlo study. A large number of simulations for an even larger number of cases will be necessary to assess the coverage rate for a representative number of values in the parameter space.

Although no correspondence between bias and coverage balance has been proven, there might be at least two reasons why we should investigate coverage balance: (1) For some parameter distributions (e.g. symmetric distributions), there might be a direct connection between coverage imbalance and bias of a confidence interval; (2) Two-sided confidence intervals with good coverage balance are desirable. A confidence interval provides a range of "good suggestions" for the true value of the parameter. The conclusion we draw from a $1-\alpha$ confidence interval is that with $100(1-\alpha)\%$ confidence, we have covered the true parameter. In general, it might not be appropriate to connect any expectation of the form "the values that are in the middle of the CI are more likely to be the parameter than the values at either end of the CI" to the interval estimate. However, when an interval estimator systematically misses the true parameter on only one side, i.e. it over- or underestimates the true parameter, it also consistently suggests ranges of values that lie above (below) the true parameter more often.

It would be very desirable to find a way of investigating bias in a form other than the one from the definition. One suggestion might be to see whether covering no value more often than the true parameter and missing the true parameter equally often on both sides (i.e., exhibiting coverage balance) are equivalent at least for a subset of interval

estimators. Then an investigation of bias would only involve collecting and displaying two values per situation. Unfortunately, such an equivalence has not been established yet.

Which interval should we use?

Differences between the three confidence intervals investigated in this study have been highlighted several times and are summarized in Table 1. Preference for one interval or the other should be guided by the research question. Often, when a choice between several procedures available for testing one set of hypotheses has to be made, we look for the uniformly most powerful test (UMP) that successfully controls Type I error rate at the desired level. Certainly, Westlake's symmetric interval provides less power than when Seaman and Serlin's interval or the traditional equivariant $1-2\alpha$ confidence interval are used to test for equivalence and it does not seem to offer any other advantages.

When we construct a confidence interval, we hope to gain additional information with respect to the parameter of interest as opposed to the information a simple hypothesis test can provide us with. While both the traditional $1-2\alpha$ CI and Seaman and Serlin's CI will reject the null hypotheses from Equation (4) at the same rate, the $1-2\alpha$ CI has a confidence coefficient of $100(1-2\alpha)\%$ and Seaman and Serlin's CI has a confidence coefficient of $100(1-\alpha)\%$. However, these additional $\alpha\%$ might be bought at a high cost: Their CI will always be wider than the $1-2\alpha$ CI and in many cases it might be substantially wider. Example 1 above demonstrated a possible situation in which Seaman and Serlin's interval was 3.79 times wider than the traditional $1-2\alpha$ CI.

An alternative suggestion might be to test the two null hypotheses with a $1-2\alpha$ confidence interval or the TOST procedure and then present the traditional $1-\alpha$

confidence interval for the parameter, if an interval with $100(1-\alpha)\%$ confidence coefficient is desired. I would like to note that a correspondence between the choice of Type I error rate for any hypothesis test and the confidence coefficient of the interval that can be used to test the hypothesis is not mandatory.

As a two-sided CI, Seaman and Serlin's interval systematically underestimates the absolute value of the true parameter $\mu_1 - \mu_2$, but can be perceived of as a $1-\alpha$ one-sided confidence interval for the absolute value of $\mu_1 - \mu_2$. In equivalence testing, we might ask what is more dangerous, over- or underestimating the true difference. Equivalence testing tries not to wrongly conclude that the two parameters are practically equal and hence, the only dangerous situation from a "protecting the null hypothesis" perspective is to underestimate the parameter. Both Seaman and Serlin's interval and the $1-2\alpha$ CI are going to underestimate the parameter equally often.

An important reminder seems necessary here: Seaman and Serlin do not suggested using the procedure from Equations (19) and (20) to test the equivalence null hypotheses. Instead, they recommend a traditional hypothesis test followed by construction of a traditional $1-\alpha$ confidence interval or their equivalence procedure depending on the outcome of the hypothesis test. I have not discussed their complete procedure in this thesis, although a thorough discussion might promise interesting insights. Rather, the intention here was to discuss the interval properties of the new confidence interval they constructed and compare it to other confidence intervals.

Summarizing all aspects, the traditional $1-2\alpha$ confidence interval seems to be the most attractive alternative for a number of reasons: It is the only unbiased two-sided confidence interval. The other two techniques are either strongly biased or else have to be

conceived as one-sided confidence intervals. Further, it is the narrowest interval that allows testing both equivalency null hypotheses simultaneously at the .05 Type I error rate. Of course one may wish to construct a one-sided interval, that is, one may only be interested in the size of the absolute value. However, it is important to notice that this interval is only gaining limited additional information above the TOST compared to the $1-2\alpha$ CI (e.g. Meredith & Heise, 1996).

Performance of the equivalence test for the difference between two correlations

The construction of test statistics and the $1-2\alpha$ confidence interval for the difference between two correlations from independent normal random samples from Equations (48), (49) and (50) are only asymptotic. Since the statistics are not exact, it was necessary to investigate whether Type I error rate is controlled at acceptable levels and whether the computed values for power agree with empirical values. Empirical Type I error rate was found to be controlled at an acceptable level for $\alpha = .05$ and $\alpha = .01$ in most cases, although testing two large correlations could lead to a significant deviation from the nominal Type I error level, especially as the size of the equivalence region gets larger. Computed power values and nominal coverage rate seemed to agree quite well with empirical values for power and coverage rate, while coverage imbalance was substantial in some cases. I have neither considered exact formulas nor non-normal data or dependent samples. The formula for the pdf of a correlation is considerably complex to make hand computations hard if not impossible. If asymptotic formulas perform well, the gain in performance that can be achieved by exact formulas might not be worth the effort. Non-normal data on the other hand might alter the distribution of $\hat{\rho}_1 - \hat{\rho}_2$ to a degree that

does not warrant the use of asymptotic formulas any longer. When testing correlations from dependent samples, power might increase. All these issues should be investigated in the future since they might make equivalence testing a much more viable technique in some situations while the use of asymptotic formulas might forbid itself in others. Bootstrap and Maximum Likelihood Estimation might offer valuable alternatives.

Is Equivalence Testing for the Difference Between Two Correlations Worth Doing?

When a researcher would like to show that the difference between two parameters is trivially small, equivalence testing offers an attractive alternative to previous options. Efforts to show practical equivalence between two parameters have included the conduction of *accept-support* testing in the past. *Accept-support* testing will often suffer from insufficient power which will lead to non-rejection of the null hypothesis (in favor of the researcher's interests) not due to evidence in favor of the null but simply due to a lack of desirable precision. At the same time, it is possible to have "too much" power or precision such that an inconsiderable difference will still lead to a rejection of the null hypothesis. Only in few situations will *accept-support* testing be a satisfactory alternative for showing that the difference between two parameters is practically trivial. Equivalence testing can provide evidence for practical equivalence utilizing the *reject-support* testing logic that rewards increasing precision. Under these circumstances it seems that equivalence testing is indispensable for anyone investigating whether the difference between two parameters is practically trivial. This compelling advantage in terms of hypothesis testing logic applies to correlational equivalence testing as well. Yet it was necessary to determine whether correlational equivalence tests are feasible in practice, i.e.

what the cost for required precision is. Test about two correlations require relatively large sample sizes in order to provide sufficient power (see, e.g. Cohen, Cohen, West & Aiken, 2002), and correlational equivalence testing is not an exception to this rule. The tests developed in this study performed satisfactorily in most cases, and we can make a general statement on the feasibility of equivalence testing for the difference between two correlations. Power will be sufficient at small to medium sample sizes only when the two correlations involved are large, but when the two correlations are close to zero, sample sizes of 1500 or more are required to assure power of 0.8. Researchers who hope to find equivalence between two correlations should keep this in mind. The consequences of unawareness of insufficient power can be devastating for a research program. Unfortunately, the asymptotic formulas developed here tend to perform suboptimally exactly in those situations where sufficient power would be available, namely when the two correlations are large. I would recommend a search for methods with better Type I error control and coverage rate (and possibly coverage balance) to test differences between two large correlations.

Why Is There Coverage Imbalance In The Confidence Interval?

The equivariant $1-2\alpha$ confidence interval from Equation (50) does not miss the true parameter equally often above and below. I was interested in understanding the reason for such imbalance, since this might suggest ways to find better confidence interval estimates. The construction of the CI is based on the assumption that $r_1 - r_2$ is normally distributed with mean $\rho_1 - \rho_2$ and a standard deviation from Equation (47).

That a supposedly equivariant confidence interval around θ (constructed as $\hat{\theta} \pm z_{\alpha} * \widehat{\sigma}$) misses the true parameter more often on one side than on the other could have the following reasons (as well as their interaction): (1) The interval will, on average, have the same width across the whole parameter space, however, the distribution of the parameter estimate is skewed while we are assuming it is normal. The distributions of both confidence interval limits will be skewed as well and this will lead to the bias. (2) The parameter estimate is more or less normally distributed, however it is *not* independent from its standard error. In this case, the width of the CI varies together with the value we observed for the parameter estimate. If the size of the standard error is systematically larger for sample values from one side of the distribution of the parameter estimate, the CIs for those values will be wider and therefore contain the true parameter more often while CIs on the other side of the distribution will be narrower and contain the true parameter less often, which creates the bias.

In general, the distribution of *r* is not normal and thus, the distribution of $r_1 - r_2$ might be normal either. Further, the value for $r_1 - r_2$ and it's standard deviation estimate from Equation (47) are not independent, in fact, in some settings, the correlation between $r_1 - r_2$ and $s_{r_1 - r_2}$ can be quite large (e.g. when $r_1 = 0.5$ and $r_2 = -0.5$).

Future Directions

I would like to continue examining the relationship between bias from Definition 4 and coverage balance both analytically and with Monte Carlo simulations. Further, it would be of interest to look at previous publications that utilize confidence intervals and investigate whether undetected bias or coverage imbalance of two-sided CIs led to

erroneous conclusions, determining how seriously the neglect of these two concepts affects the research process. Better estimates for the $1-2\alpha$ confidence interval around the difference between two correlations might lead to improved Type I error control and possibly coverage rate. Maximum Likelihood estimation might provide means to create asymmetrical intervals with better properties.

REFERENCES

Anderson, S. & Hauck, W. W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics — Theory and Methods, 12* (23), 2663–2692.

Berger, R. L., & Hsu, J.C. (1996). Bioequivalence Trials, Intersection-Union Tests and Equivalence Confidence Sets. *Statistical Science 11*(4), 283–302.

Brown, L. D., Hwang, J. T. G., & Munk, A. (1997). An unbiased test for the bioequivalence problem. *The Annals of Statistics, 25* (6), 2345–2367.

Browne, M. W., (1968). A Comparison of Factor Analytic Techniques. *Psychometrika 33*(3), 267–334.

Casella, G., & Berger, R. L., (2002). Statistical inference (2nd ed.). Pacific Grove, CA: Duxbury.

Cohen, J., (1962). The Statistical Power of Abnormal-Social Psychological Research: A Review. *Journal of Abnormal and Social Psychology, 65*(3), 145–153.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2002). Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Feng, S., Liang, Q., Kinser, R. D., Newland, K., & Guilbaud, R. (2006). Testing equivalence between two laboratories or two methods using paired-sample analysis and interval hypothesis testing. *Analytical and Bioanalytical Chemistry, 385*, 975–981.

Hedges, L. V., & Olkin, I. (1983). Regression models in research synthesis. *The American Statistician, 37* (2), 137–140.

Kirkwood, T. B. L. (1981). Bioequivalence testing — A need to rethink [Letter to the Editor]. *Biometrics 37*, 589–594.

Liu, J. P. & Chow, S. C. (1992). Sample size for the two one-sided tests procedure in bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics, 20* (1), 101–104.

Mantel, N. (1977). Do we want confidence intervals symmetrical about the null value? [Letter to the Editor]. *Biometrics*, 759–760.

Marsaglia, G., & Zaman, A., (1991). A new class of random number generators. *The Annals of Applied Probability, 1* (3), 462–480.

Meredith, M. P. & Heise, M. A. (1996). [Comment]. *Statistical Science, 11*(4), 304–306.

Olkin, I., & Finn, J. D. (1995). Correlations redux. *Psychological Bulletin, 118* (1), 155–164.

Phillips, K. F., (1990). Power of the two one-sided tests procedure in bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics, 18*(2).

Rocke, D. M. (1984). On Testing for bioequivalence. *Biometrics 40,* 225–230.

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin 113*(3), 553–565.

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*(6).

Seaman, M. A., & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods 3*(4), 403–411.

Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin 105*(2), 309–316.

Stegner, B. L., Bostrom, A. G., & Greenfield, T. K. (1996). Equivalence testing for use in psychosocial and services research: An introduction with examples. *Evaluation and Program Planning, 19*(3), 193–198.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative methods of conducting null hypothesis statistical tests. *Psychological Methods 6*(4), 371–386.

Westlake, W. J. (1972). Use of confidence intervals in analysis of comparative bioavailability trials. *Journal of Pharmaceutical Sciences, 61* (8), 1340–1341.

Westlake, W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics, 32*(4), 741–744.

Westlake, W.J. (1981). Bioequivalence testing – A need to rethink – Reply [Letter to the editor]. *Biometrics, 37* (3), 591–593.

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