EQUIVALENCE TESTS FOR REPEATED MEASURES

VICTORIA NG

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

MASTER OF ARTS

GRADUATE PROGRAM IN DEPARTMENT OF PSYCHOLOGY
YORK UNIVERSITY
TORONTO, ONTARIO

Abstract

Equivalence tests from the null hypothesis significance testing framework are appropriate alternatives to difference tests for demonstrating lack of difference. For determining equivalence among more than two repeated measurements, recently developed equivalence tests include the omnibus Hotelling T^2, the pairwise standardized test, the pairwise unstandardized test, and the two one-sided test for negligible trend. With Monte Carlo simulations, the current research evaluated Type I error rates and power rates for these equivalence tests to inform an applied data analytic strategy. Because results suggest that there is no one statistical test that is optimal across all situations, I compare the tests' statistical behaviours to provide guidance in test selection. Specifically, test selection should be informed by the measurement level of the repeated outcome, correlation structure, and precision.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Rob Cribbie, for his ongoing feedback and support. With his encouragement, openness, and guidance, I have learned to explore on my own and to pursue collaboration. I would also like to thank the committee members, Dr. David Flora, Dr. Justin Podur, and Dr. Christopher Green, for their helpful comments and questions during the defense. My thanks also go to Phil Chalmers for his generosity in sharing resources and ideas and to Dr. Hugh McCague for his insight into aspects of theoretical and applied statistics. I would also like to acknowledge the fellow students in the program, as well as the departmental professors, for their combined ability to build a collegial atmosphere of learning and sharing.

Contents

List of Tables

List of Figures

Equivalence Tests for Repeated Measures

Beyond the hypotheses of change and difference that are prevalent in repeated measures studies of psychology, hypotheses for lack of change also arise. For research questions of equivalence in longitudinal settings, for example, Davidson et al. (2003) assessed the influence of meditation on affect and immune functions, and they stringently hypothesized that the control group would show no change over time; Graney and Engle (2000) sought to show that a single assessment of daily living would be as accurate as intensively repeated assessments by providing evidence that all responses were practically equivalent; and in a developmental study, Buist, Reitz, and Deković (2004) used correlation analyses and tests of slope to demonstrate the stability of internalizing behaviour scores over several years. Research questions of equivalence also appear in experimental research, particularly when stimuli sets require validation. For example, the study of affect and odour pleasantness requires that odour intensities be controlled, so ratings for perceived intensities would need to be practically equivalent across odours in order to isolate the effect of pleasantness (Winston, Gottfried, Kilner, & Dolan, 2005); similarly, the study of selective activation and letter processing in the brain also requires that certain subsets of experimental conditions are statistically equivalent in order to conclude that brain regions of study are equally recruited (Joseph, Cerullo, Farley, Steinmetz, & Mier, 2006).

To establish equivalence, all these aforementioned studies used inferential tests of difference. However, the establishment of equivalence based on statistical non-significance from difference tests is problematic. Logically, rejection of a null hypothesis implies that there is evidence against it. Failure to reject the null hypothesis does not, however, imply that there is evidence for the null hypothesis. Statistically, seeking equivalence via difference tests yields lower power with increased sample size and increased power with lower sample size. Drawing the conclusion for equivalence via difference tests that are designed to test for differences, particularly with small sample sizes, is therefore both illogical and anathema to research ideals.

These issues are addressed by equivalence tests. Equivalence tests from the framework of null hypothesis significance testing determine the strength of evidence against the null hypothesis that population parameters are non-equivalent. With the null hypothesis stated as a form of non-equivalence and the alternate hypothesis as a statement of equivalence, a Type I error event is defined as an incorrect conclusion for equivalence. A power event is then defined as a correct conclusion for equivalence. Unlike the implementation of difference tests, usage of equivalence tests requires the additional and substantive definition of 'equivalence'. Generally, population parameters are considered equivalent if their differences are so small that they are considered negligible. The degree of this negligibility is defined by some equivalence interval, say $\pm\epsilon$, which is prespecified *a priori* by the researcher. For example, if a difference of two units is considered to be the minimum effect deemed to indicate meaningful difference, then the equivalence interval may be ± 2. The equivalence interval then deems any difference that falls within the boundaries of $\pm\epsilon$ (e.g., > -2 or < 2) to be negligible, or essentially, practically equivalent.

The majority of equivalence tests have been developed in the context of biopharmaceutical studies, but their applications have steadily gained prominence in the behavioural sciences. Most studies on equivalence tests for the behavioural sciences have focused on independent groups designs (Gruman, Cribbie, & Arpin-Cribbie, 2007; Koh & Cribbie, 2013; Schuirmann, 1987; Seaman & Serlin, 1998). Fewer studies have examined equivalence tests for designs with repeated measures. Within psychology, only equivalence tests for paired samples have been empirically evaluated (Mara & Cribbie, 2012).

Currently, no clear recommendation has been made for testing the overall equivalence of more than two repeated measures, either in experimental or longitudinal contexts of psychology. There are three tests of mean equivalence and one test of negligible trend that invite evaluation and discussion. Wellek (2010) introduced three tests of mean equivalence: an adapted Hotelling T^2 test and two pairwise-based procedures using the intersection-union principle (explicated later). For longitudinal contexts, the test for negligible trend is an adapted equivalence test that exploits model fitting procedures.

Equivalence Tests for More than Two Measurement Occasions

Omnibus Hotelling T^2

The Hotelling T^2 test is a multivariate approach that treats k repeated measurement occasions as k separate dependent variables. Consequently, there are no assumptions made for correlational structure or sphericity, which are inherent in univariate approaches like the traditional, repeated-measures analysis of variance.

The equivalence interval ϵ is specified in terms of Mahalanobis distance, which is a multivariate measure of distance from a data set's centroid that accounts for differences in scale among the dependent variables. For k means, a contrast matrix $\mathbf{C}$ specifies comparisons for $k-1$ adjacent mean differences δ_j such that

$$\delta_j = \mu_{j+1} - \mu_j,\ j = 1, \ldots, k-1. \tag{1}$$

Thus, the row vector of $(k-1)$ mean differences δ is calculated as $\delta = (\mathbf{C}M)'$, where M is a column vector of k means. For the $k \times k$ covariance matrix $\mathbf{\Sigma}$, the (k-1) × (k-1) covariance matrix of the differences is $\mathbf{\Sigma_D} = \mathbf{C\Sigma C}'$. The equivalence hypotheses are thus expressed as follows:

$$H_0 :\ \delta(\mathbf{\Sigma}_D^{-1})\delta' \geq \epsilon^2 \tag{2}$$

$$H_1 :\ \delta(\mathbf{\Sigma}_D^{-1})\delta' < \epsilon^2. \tag{3}$$

Taken together, $\delta(\mathbf{\Sigma}_D^{-1})\delta'$ expresses the Mahalanobis distance for adjacent pairs of means. For the set of sample means $\bar{\mathbf{X}}$ and sample covariance matrix for the differences $\mathbf{S}_D$, the sample T^2 statistic, then, is

$$T^2 = n\left(\mathbf{C}\bar{\mathbf{X}}\right)' \mathbf{S}_D^{-1}\, \mathbf{C}\bar{\mathbf{X}}. \tag{4}$$

The null hypothesis is rejected if T^2 falls in the critical region, which is given by

$$T^2 < ((n-1)(k-1)/(n-k+1))\, F_{k-1,n-k+1;\alpha}\left(n\epsilon^2\right), \tag{5}$$

where n is the sample size and α is the nominal Type I error probability. The term $F_{k-1,n-k+1;\alpha}\left(n\epsilon^2\right)$ denotes the lower $100\alpha\%$ point of the F-distribution with the noncentrality parameter of $n\epsilon^2$ and the degrees of freedom of $k-1$ and $n-k+1$.

It would be expected that, in the context of an equivalence testing problem, an omnibus procedure should have greater power over pairwise procedures, as was demonstrated in the independent groups case (Cribbie, Arpin-Cribbie, & Gruman, 2009). Despite this expected power advantage, it should be noted that the shape of the theoretical equivalence region — the region which contains the set of mean difference vectors that are considered practically equivalent — depends on the correlation structure of intraindividual differences. Suppose that the equivalence interval ϵ is 0.50 (or equivalently, $\epsilon^2 = 0.25$). When intraindividual differences are uncorrelated, the equivalence region is spherical (Figure 1). Within this region is the set of adjacent mean difference vectors that are considered practically equivalent. As the magnitude of the correlation of the difference scores increases, however, the equivalence region shape becomes increasingly elongated and elliptic (Figure 2). At the narrow ends of these ellipses, seemingly similar sets of adjacent mean differences may be associated with different results; one vector of adjacent mean differences may lie within the region while a similar vector lies outside.

It is arguable that that this test should not be recommended because the dependency of the equivalence region on correlation structures can make for problematic interpretations, especially if highly correlated differences do occur frequently in practice. However, it is possible that such

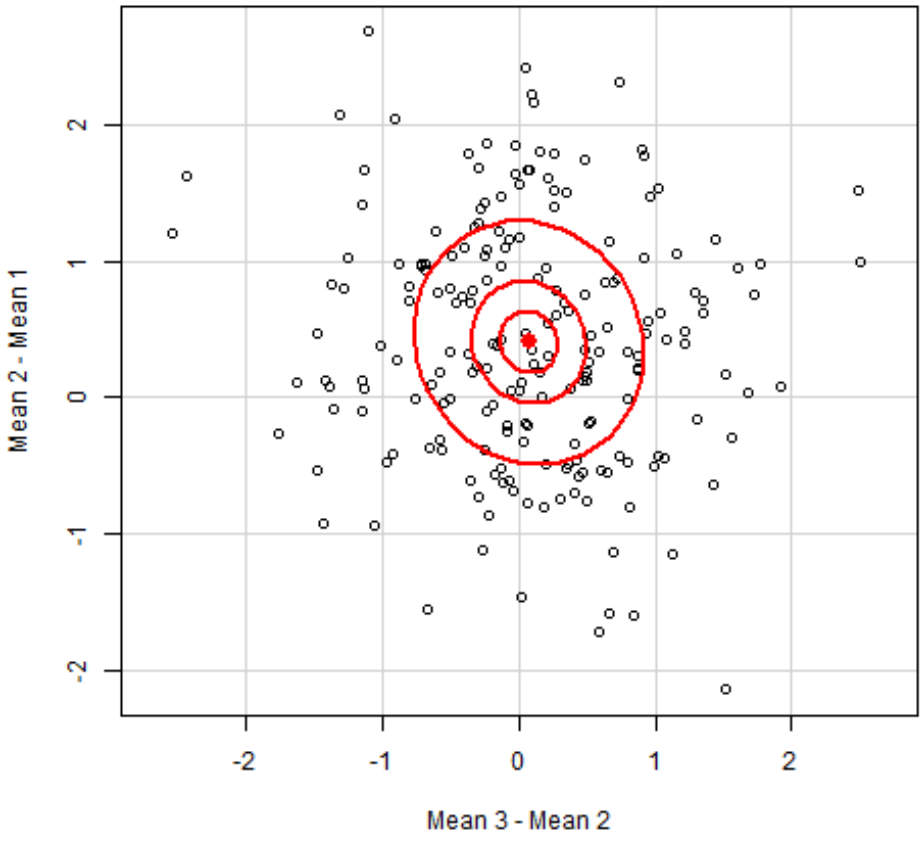

Figure 1: Equivalence region for uncorrelated difference scores. The correlation for the set of two adjacent mean difference scores is - .08.

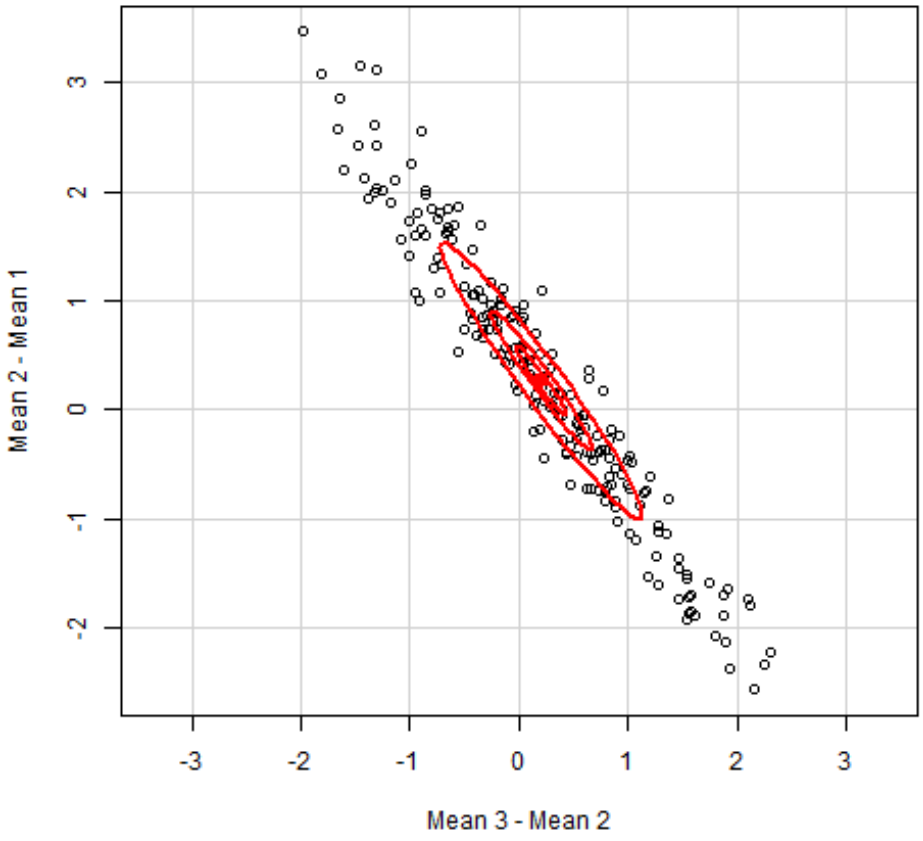

Figure 2: Equivalence region for highly correlated difference scores.The correlation for the set of two adjacent mean difference scores is - .97.

correlation structures rarely appear in practice and that applied researchers would be able to identify whether the Hotelling T^2 is appropriate for their data by computing the correlation structure of adjacent mean differences. The slight computational inconvenience could be arguably justified by any substantial power advantage. Of course, any argumentative justification must come after evidence. Thus, an empirical demonstration for highlighting consistent relationships between repeated measures data and corresponding correlations among intraindividual differences would elucidate, at least to the extent that simulation conditions provide, just how prevalent elongated equivalence regions may be in actual practice.

Pairwise Procedures using the Intersection-Union Principle

For k measurement occasions, the pairwise approach — which includes the standardized and unstandardized forms — involves the comparison of all possible pairs among j and l, where j and l each index some measurement occasion, $1 \leq j \leq k$, $1 \leq l \leq k$, and $j \neq l$, totaling $k(k-1)/2$ pairwise comparisons. This set of comparisons is embedded within a composite hypothesis test consisting of $k\,(k-1)\,/2$ sub-hypotheses.

The composite hypothesis test is constructed as an intersection-union test, which expresses the null hypothesis as a union of subsets and the alternative hypothesis as an intersection of subsets. In other words, the null hypothesis states that at least one sub-hypothesis is true, while the alternative hypothesis states that all sub-hypotheses are true (Berger, 1982). Thus, in the context of the current equivalence problem, the composite null hypothesis states that at least one pair of means, out of all pairwise comparisons, is not equivalent. Thus, if all sub-hypotheses stating the null condition are

rejected in favour of pairwise equivalence, then the composite null hypothesis of non-equivalence is also rejected. If even one sub-hypothesis stating the null condition fails to be rejected, then the composite null hypothesis also fails to be rejected.

It would seem that multiplicity control is required, for an increase in tests typically leads to an increase in familywise Type I error rate, the probability of falsely rejecting at least one null hypothesis in a family of hypotheses. The composite null hypothesis can be considered as a family of hypotheses; intuitively, it would seem that as the number of sub-hypotheses increases, the familywise error rate would also increase. This would only be the case if one were to conclude on equivalence for only a subset of all the null hypotheses instead of all possible null hypotheses. Thus, to maintain the familywise error rate at or below the α, the test result must conclude on either the equivalence of all possible pairs or on non-equivalence, but not the equivalence of only a subset smaller than the set of all possible pairs.

Overall, for procedures that are constructed as intersection-union tests, the global procedure does not require multiplicity control (Berger & Hsu, 1996). To show this more formally, Berger (1982) proved the theorem that if all sub-hypotheses of the null hypothesis are tested at the α-level, for which the rejection region is size α, then the intersection-union test for all sub-hypotheses will have a rejection region that is an intersection of the rejection regions of all those sub-hypotheses. In other words, the rejection region does not contain more than $\alpha\%$ of test statistics that are incorrectly rejected.

Intersection-union tests are considered exactly valid, in that the risk of a Type I error is maintained below α. With multiple sub-hypotheses, they are suggested to be prone to conservatism

(Berger & Hsu, 1996). However, there are conditions under which the intersection-union hypothesis tests do have accurate Type I error rates (Berger, 1982) and adequate power (Laska & Meisner, 1989). For the present problem of equivalence for multiple repeated measures, empirical work has not yet described the extent of this conservatism and the pattern in which this conservatism may exhibit itself.

Standardized Pairwise

One variant of the pairwise procedure is the standardized pairwise test. The hypothesis test is framed in terms of the standardized mean difference and uses the paired *t*-test of equivalence for each comparison. The population mean difference is standardized by the standard deviation of the pair's population difference scores. Thus, the equivalence interval ϵ can be specified by the researcher in two ways: as some arbitrary constant greater than 0 involving no other unknown parameter or as multiples of the theoretical standard deviations of the differences, $\sigma_{(j,l)}$ (Wellek, 2010). The hypotheses are the following:

$$H_0\text{: } \frac{|\mu_j - \mu_l|}{\sigma_{(j,l)}} \geq \epsilon, \text{ for at least one pairwise comparison} \tag{6}$$

$$H_1 : \frac{|\mu_j - \mu_l|}{\sigma_{(j,l)}} < \epsilon, \text{ for all pairwise comparisons} \tag{7}$$

where $\sigma_{(j,l)}$ is the population standard deviation of the differences for measurement occasions j and l. $\sigma_{(j,l)}$ also relates to the two marginal population variances, σ_j^2 and σ_l^2, and the covariance, σ_{jl},

as,

$$\sigma_{(j,l)} = \sqrt{\sigma_j^2 + \sigma_l^2 - 2\sigma_{jl}}. \tag{8}$$

The null hypothesis is rejected if, for all pairwise comparisons,

$$\frac{\sqrt{n}|\bar{X}_j - \bar{X}_l|}{S_{(j,l)}} < \sqrt{F_{1,n-1;\alpha\left(n\epsilon^2\right)}}. \tag{9}$$

The critical region is based on the noncentral F-distribution, with 1 and $n-1$ as the degrees of freedom and $n\epsilon^2$ as the noncentrality parameter. Because the square root of a random variable following the F-distribution with a noncentrality parameter of ϵ^2 gives a random variable that follows the noncentral t-distribution with $n-1$ degrees of freedom and a noncentrality parameter of ϵ (Johnson et al., 1995), the critical value in Equation 9 can be equivalently written as $t_{n-1,\,\alpha(\sqrt{n}\epsilon)}$. It should be noted that the distribution used is the noncentral F-distribution (or similarly, the noncentral t-distribution) because the parameter of interest is a standardized mean difference instead of a raw difference.

The standardized pairwise test should be invariant to differences in correlation structure. Loosely translated from mathematical theory, the distribution of each standardized mean difference is equal in distribution to $\sigma_0\sqrt{1-r_{(j,l)}}\left(Z_{ji}-Z_{li}\right)$, where all Zs are standard normally distributed variates that are mutually independent, σ_0 is the standard deviation for each measurement occasion, and $r_{(j,l)}$ is the correlation of two measurement occasions; this implies that the joint distribution of the $k\left(k-1\right)/2$ vector of pairwise t-statistics from the sample is of the same distribution as those calculated from unit normal vectors. Because unit normal test statistics do not depend on correlation structure, neither does the standardized test depend on it (Wellek, 2010).

Unstandardized Pairwise

The second pairwise procedure frames the hypothesis test with a parameter for the raw mean difference, as opposed to the standardized mean difference. The hypotheses are the following:

$$H_0 : |\mu_j - \mu_l| \geq \epsilon, \text{for at least one pairwise comparison} \tag{10}$$

$$H_1 : |\mu_j - \mu_l| < \epsilon, \text{ for all pairwise comparisons.} \tag{11}$$

Each pairwise comparison is conducted with the two one-sided test (TOST) procedure, originally popularized by Schuirmann (1987) and recently adapted for paired samples by Mara and Cribbie (2012). The TOST approach for establishing equivalence of two related means tests two sub-hypotheses in relation to the equivalence interval ϵ. The hypotheses are the following:

$$H_{TOST_{01}} : \mu_2 - \mu_1 \geq \epsilon, \quad H_{TOST_{02}} : \mu_2 - \mu_1 \leq -\epsilon \tag{12}$$

$$H_{TOST_{11}} : \mu_2 - \mu_1 < \epsilon, \quad H_{TOST_{12}} : \mu_2 - \mu_1 > -\epsilon. \tag{13}$$

Denoting S_D as the sample standard deviation of the difference, two sample t-statistics are computed as

$$t_{TOST_1} = \frac{\bar{X}_2 - \bar{X}_1 - \epsilon}{s_D/\sqrt{n}} \tag{14}$$

$$t_{TOST_2} = \frac{\bar{X}_2 - \bar{X}_1 - (-\epsilon)}{s_D/\sqrt{n}}. \tag{15}$$

Equivalence for the paired sample is established when both $H_{TOST_{11}}$ and $H_{TOST_{12}}$ are rejected, which occurs when both test statistics fall in their critical regions,

$$t_{TOST_1} \leq -t_{\alpha,n-1}, \text{ and } t_{TOST_2} \geq t_{1-\alpha,n-1}. \tag{16}$$

Because the TOST is a simple form of the intersection-union test (Berger & Hsu, 1996), the composite hypothesis test of the TOST has accurate Type I error rates, even when both sub-hypotheses are tested at α level. This becomes more intuitive when it is seen that $H_{TOST_{01}}$ and $H_{TOST_{02}}$ cannot be simultaneously true, so the effective critical region is of size α and not 2α (Blair & Cole, 2002).

The paired TOST is applied for every pairwise comparison. By the intersection-union principle, the null hypotheses for each and every one of the pairwise comparisons must be rejected to conclude in favour of equivalence for all k occasions. The decision rule can also be written as

$$\frac{\epsilon - |\bar{X}_j - \bar{X}_l|}{s_{(j,l)}/\sqrt{n}} > t_{1-\alpha,n-1,} \text{ for all pairwise comparisons.} \tag{17}$$

where $s_{(j,l)}$ is the standard deviation of the differences for some pair.

Using an unstandardized mean difference, the test statistic is distributed as a central t-distribution. Because the parameter of interest is in unstandardized scale, the test should have properties that

vary by the spread of each repeated measurement and by the correlation structures (Wellek, 2010). This behaviour is similar to that of difference tests using unstandardized parameters and is quite unlike the aforementioned standardized test. The variation in its statistical behaviours across correlation structures can be explained by an alternative expression of the rejection region. Consider the population variance of the difference scores $\sigma^2_{(j,l)}$, given as

$$\sigma^2_{(j,l)} = \sigma^2_j + \sigma^2_l - 2\sigma_j\sigma_l\rho_{(j,l)} \tag{18}$$

where the correlation coefficient of the two dependent measurements is $\rho_{(j,l)}$. It is seen here that $\sigma^2_{(j,l)}$ decreases as the correlation $\rho_{(j,l)}$ increases. More specifically in the case where all variances of the measurement occasions are one, the rejection region of Equation 17 can be expressed asymptotically as

$$\frac{\epsilon - |\bar{X}_j - \bar{X}_l|}{\sqrt{2(1-r_{(j,l)})}/\sqrt{n}} > u_{1-\alpha}. \tag{19}$$

where $u_{1-\alpha}$ is the critical value at α level for the standard normal distribution. It can be seen once again that as $r_{(j,l)}$, the sample correlation coefficient of two dependent measurements, increases, the standard error decreases, thus yielding a test statistic that is more likely to fall past the critical value. Indeed, Mara and Cribbie's (2012) comparison of the standardized and unstandardized paired samples tests showed that the power of the unstandardized form increases with stronger correlations.

Two One-Sided Test (TOST) for Negligible Trend

In longitudinal analyses for linear trends, a phenomenon of change can be quantified as some change in outcome over a unit of time — essentially, a slope. To determine if the change over time is statistically significant, one tests whether the estimate of the mean of the slope is significantly different from zero. To determine if the change over time is only negligibly different, it is still the case that statistical non-significance on a difference test is inadequate evidence for demonstration of lack of change.

Dixon and Pechmann (2005) recognized the applicability of negligible trends in ecology; in time-series studies for population trends of species, it is sometimes the case that population growth is hypothesized to be negligible or that growth is not substantively meaningful. Reframing the TOST for the context of trends, the TOST for negligible trend differs from the previous tests in that the equivalency of aggregate means is not tested; rather, this test determines whether the rate of change, from one measurement occasion to the next, is small enough to be considered negligible. Thus, the bounds of the equivalence interval no longer pertain to mean differences but rather to the rate of change across some ordered predictor, such as time.

Following the logic of the intersection-union TOST, the set of null hypotheses for the test of negligible trend states that the population slope is or exceeds the minimum rate of change considered to be meaningful, while the alternate hypotheses states that the population slope falls within the range of negligible values, as follows:

$$H_{01} : \beta \geq \epsilon, \quad H_{02} : \beta \leq -\epsilon \tag{20}$$

$$H_{11} : \beta > -\epsilon, \quad H_{12} : \beta < \epsilon. \tag{21}$$

The two test statistics are computed as

$$t_1 = \frac{b - \epsilon}{SE_b} \tag{22}$$

$$t_2 = \frac{b - (-\epsilon)}{SE_b} \tag{23}$$

where b is the estimated model slope and SE_b is the associated standard error with the slope. The overall null hypothesis that the slope falls outside the equivalence interval is rejected if both test statistics are rejected,

$$t_1 \leq -t_{\alpha, df}, \text{ and } t_2 \geq t_{1-\alpha, df} \tag{24}$$

where *df* refers to the appropriate degrees of freedom from some model.

The TOST for negligible trend is proposed to be conducted after fitting an appropriate model, which should provide the slope, standard error, and the degrees of freedom for inference. In their paper, Dixon and Pechmann (2005) fitted autoregressive time-series models and simply incorporated the estimated slopes, standard errors, and Kenward-Roger approximations for degrees of

freedom (Kenward & Roger, 1997) into the TOST framework. They also suggested that with other types of data, standard errors and degrees of freedom could be obtained from whichever model may be most suitable, such as linear mixed models (Dixon & Pechmann, 2005).

Because mixed models have gained popularity for longitudinal studies in psychology, it would seem advantageous to couple the flexibility of alternative modeling approaches with the intuitive implementation of the TOST. However, there have been few applied examples of coupling mixed models with the TOST and no simulation that has demonstrated the statistical properties of this procedure. To ensure that the TOST for negligible trend is useful, it would be necessary to ascertain that the TOST behaves as expected when using fixed effects estimates and degrees of freedom from external models. Specifically, one must ascertain that the degrees of freedom for the test of slope is applicable in the TOST framework, as was initially proposed.

Current Study

The aforementioned tests are potential avenues for the problem of testing equivalence in longitudinal or repeated-measures research, but it is unclear as to how behavioural researchers should choose from among the options given some of the issues discussed earlier (e.g., the Hotelling T^2's equivalence region shape, the pairwise procedures' conservatism). Beyond the exposition of theoretical properties from the perspective of mathematical statistics, there has been no empirical comparison for the power of the equivalence tests for repeated measures under a common set of conditions, no description of the conservatism of the pairwise procedures, and no clear recommendation for analysis. In particular, the Hotelling T^2 is an omnibus procedure that should optimize

power, but it has an equivalence region that is dependent upon the correlational structure of the data; and, although the pairwise procedures are not afflicted with any dependence on correlation structure, conservative Type I error rates could be associated with diminished power. It is quite possible that there is no one optimal solution, but empirical demonstrations of these procedures' statistical behaviours may provide insight into the conditions in which each test may be most useful.

The current research aims to empirically evaluate the following: Type I error rates for each test of mean equivalence (Hotelling T^2, standardized pairwise, and unstandardized pairwise), power rates for the tests of mean equivalence under a common set of conditions, Type I error and power rates for the test of negligible trend using linear mixed models (random intercept model), and the extent to which the problem of the theoretical equivalence region arises in practice. With the criteria that the recommended solution is valid in Type I error rate, adequate in power, and relatively invariant across data situations, this study is expected to provide insight as to how researchers should choose among the existing equivalence tests for repeated measures.

Method

R Software (R Development Core Team, 2015) was used for programming each statistical test and for running Monte Carlo simulations. Two major sets of simulations were conducted: One set of separate simulations determined the Type I error rate of each hypothesis test; the second set compared power rates among the tests of mean equivalence and calculated the power rate for the test of negligible trend. Type I error rates were considered acceptable with liberal bounds of $\pm.50\alpha$ (Bradley, 1978), yielding bounds of .025 and .075 for $\alpha = .05$. Conditions for calculating Type I error and power rates were replicated 2000 times.

To determine the extent to which the problem of the Hotelling T^2's equivalence region arises in practice, a supplementary simulation also examined the correlation of difference scores corresponding to a variety of correlation structures.

Data Generation

Tests of Mean Equivalence

Conditions for the tests of mean equivalence were manipulated by the number of repeated measurements (3, 5, 7), sample size (10, 50, 90), equivalence interval width (0.25, 0.50, 0.75),

equal standard deviations across the repeated measurements (denoted as σ_0; 0.70, 1.00, 1.30), and correlation structures. Three different sets of correlation structures were created: autocorrelated, equicorrelated, and mixed. The autocorrelated structure had a starting correlation of .50 (code obtained from Rizopoulos, 2007). Equicorrelated structures had matrices with equal off-diagonals (.25, .50, .90), while two mixed correlation structures differed in off-diagonals. Inter-sample correlations for mixed structures are detailed in Table 1. Together, the defined conditions yielded 486 unique scenarios. For every replication, a population correlation structure was defined and unstandardized into a population variance-covariance matrix by σ_0. With the variance-covariance matrix and all other attributes of the condition, data were simulated as multivariate normal with equal variances for each of the repeated points.

The Type I error simulation for the Hotelling T^2 differed slightly in that there was no manipulation for either σ_0 (held constant at one) or variants of mixed correlation structures because of the way in which the Type I error conditions were simulated (detailed later).

Three measurement occasions

Matrix	r_{12}	r_{13}	r_{23}
Mixed A	.15	.25	.55
Mixed B	.70	.65	.60

Five measurement occasions

Matrix	r_{12}	r_{13}	r_{14}	r_{15}	r_{23}	r_{24}	r_{25}	r_{34}	r_{35}	r_{45}
Mixed A	.15	.25	.30	.50	.55	.40	.30	.45	.25	.70
Mixed B	.70	.65	.60	.55	.60	.55	.50	.65	.65	.85

Seven measurement occasions

Matrix	r_{12}	r_{13}	r_{14}	r_{15}	r_{16}	r_{17}	r_{23}	r_{24}	r_{25}	r_{26}	r_{27}	r_{34}	r_{35}	r_{36}	r_{37}	r_{45}	r_{46}	r_{47}	r_{56}	r_{57}	r_{67}
Mixed A	.15	.25	.20	.25	.30	.45	40	.45	.35	.25	.10	.45	.25	.20	.15	.50	.35	.25	.40	.45	.35
Mixed B	.70	.65	.60	.55	.50	.45	.60	.55	.50	.45	.40	.75	.65	.50	.35	.85	.70	.40	.60	.55	.40

Table 1: Elements of the population mixed correlation matrices

TOST for Negligible Trend

Manipulations included the number of repeated measurements (3, 5, 7), sample size (10, 50, 90), and equivalence interval width (0.25, 0.50, 0.75); unlike simulations for the tests of mean equivalence, this simulation manipulated neither σ_0 nor correlation structure. Data were simulated exactly from the following random intercept model (code obtained from Center for Statistical Consultation and Research 2011):

$$Y_{ij} = \pi_{0i} + \pi_{1i}(time) + \epsilon_{ij} \tag{25}$$

$$\pi_{0i} = \gamma_{00} + \zeta_{0i} \tag{26}$$

$$\pi_{1i} = \gamma_{10} \tag{27}$$

π_{0i} and π_{1i} are individual-level intercepts and slopes, respectively. γ_{00} is the population average intercept, and γ_{10} is the population average slope. ϵ_{ij} is the individual-level error term, and ζ_{0i} is the error term for the varying intercepts; both error terms are distributed as $N(0, 0.25)$. To ensure correct data generation for the true model, Type I error for the test of difference on the slope was checked for accuracy.

To conduct the actual TOST for negligible trend, a random intercept model was fit to these data using restricted maximum likelihood with 'time' as a predictor and with Satterthwaite approximated degrees of freedom. From the model, the fixed effect slope, its standard error, and the

associated degrees of freedom for inference on the slope were submitted to the TOST for negligible trend.

Population Mean Configurations

Tests of Mean Equivalence

The power condition was set such that all population means were 0. For the Type I error rate conditions of each test, the population mean configuration was set such that the parameter of interest lies at the boundary of the equivalence interval.

For Hotelling T^2, the Type I error condition is such that the Mahalanobis distance, $\delta\left(\mathbf{\Sigma}_D^{-1}\right)\delta'$, is equal to ϵ^2. To satisfy this criterion, a set of adjacent mean differences was iteratively found for some combination of ϵ and a variance-covariance matrix of differences. For any such combination, a mean difference — say d — was found such that each k occasion's mean (for $k > 1$) was d units away from the previous mean. The Type I error configuration can be generalized as $0, d, 2d, \ldots, kd$. Table 2 lists the differences in each adjacent mean required to create a Mahalanobis distance equal to ϵ^2 for particular combinations of the standardized variance-covariance matrix and ϵ values. As correlations increase, the adjacent mean difference d decreases.

For both pairwise procedures, the null condition states that at least one pairwise comparison yields some mean difference that falls at or beyond the equivalence interval; thus, for both pairwise procedures, exactly one pair, μ_1 and μ_2, had a population mean difference at a boundary. For the standardized pairwise test, the mean difference was to be standardized, and so the population standard deviation of the difference for this pair of means was calculated from the population

Table 2: Type I error conditions for the Hotelling T^2. The distance between adjacent means results in a Mahalanobis distance equal to ϵ^2.

		Equivalence Interval = 0.50
k	Correlation Matrix	Distance Between Adjacent Means, d
3	Equi .25	0.3061862
	Equi .50	0.2500000
	Equi .90	0.1118034
5	Equi .25	0.1369306
	Equi .50	0.1118034
	Equi .90	0.0500000
7	Equi .25	0.08183171
	Equi .50	0.06681531
	Equi .90	0.02988072

Test	Type I Error Configuration
Hotelling T^2	$0,\ d,\ 2d, \ldots, kd$
Standardized Pairwise	$0,\ \epsilon\sigma_{(1,2)},\ {}^{\epsilon\sigma_{(1,2)}}/_{2}, \ldots, {}^{\epsilon\sigma_{(1,2)}}/_{2}$
Unstandardized Pairwise	$0,\ \epsilon,\ {}^{\epsilon}/_{2}, \ldots, {}^{\epsilon}/_{2}$
TOST for Negligible Trend	$0,\ \epsilon,\ 2\epsilon, \ldots, (k-1)\,\epsilon$

Table 3: Mean configurations for Type I error conditions

variance-covariance matrix by Equation 8. For the null hypothesis to be true then, the difference between μ_1 and μ_2 is $\epsilon\sigma_D$ while all other mean differences are smaller than that of ϵ. Thus, the mean configuration is generalized as $0, (\epsilon\sigma_{D,}), (0.5\epsilon\sigma_D), \ldots, (0.5\epsilon\sigma_D)$. For the unstandardized pairwise test, the mean configuration was $0,\ \epsilon,\ {}^{\epsilon}/_{2}, \ldots, {}^{\epsilon}/_{2}$.

A summary of Type I error mean configurations for the tests of equivalence is found in Table 3.

TOST for Negligible Trend

The power condition was set such that the population fixed slope parameter was 0. For the Type I error condition, the null hypothesis states that the slope parameter falls at the boundary of the equivalence interval. For some rate of negligible change, ϵ, the mean configuration would be $0,\ \epsilon,\ 2\epsilon, \ldots, (k-1)\,\epsilon$.

Correlation of Difference Scores

This supplementary simulation explored the relation between the correlation of observed values and the correlation of their adjacent difference scores. Generally, from population correlation matrices, adjacent difference scores were calculated along with the correlations of these difference scores (i.e., the correlation between the first intraindividual difference and the second), resulting in empirical sampling distributions of correlations for intraindividual differences.

Multivariate normal data were simulated as observed scores for three measurement occasions, each with a population mean of 0, standard deviation of 1, and sample size of 1000. Three population correlations were chosen: the low correlation of .10, the medium correlation of .50, and the high correlation of .90. Given three measurement occasions, all possible permutations of these three base correlation magnitudes resulted in 27 conditions. Each condition was replicated 5000 times. For each dataset of observed scores, difference scores between adjacent vectors were computed; that is, for $1 < i < n$ observations on three measurement occasions, the two difference scores were $x_{3i} - x_{2i}$ and $x_{2i} - x_{1i}$. The correlational magnitude of these intraindividual differences ($|r_{(x_3-x_2,x_2-x_1)}|$=$|r_{diff}|$) was extracted on each replication; thus, for every condition (i.e., every tested correlation structure for observed data) with 5000 replicates, a distribution of $|r_{diff}|$s was obtained, and the mean magnitude of these correlations was calculated.

Based on a visual appraisal of the rejection region as detailed in Wellek (2010), it seems that correlations of intraindividual differences well beyond .50 would yield problematic interpretations. To describe the expected $|r_{diff}|$ associated with each tested correlation structure, proportions were obtained for the $|r_{diff}|$s falling into each of the following intervals: $\left(|r_{diff}| \leq .30\right)$,

$\left(.30 < |r_{diff}| \leq .50\right)$, $\left(.50 < |r_{diff}| \leq .70\right)$, and $\left(|r_{diff}| > .70\right)$. Thus, if some particular condition yielded a distribution that was centrally located either between .50 and .70 or beyond .70, the correlation structure of that condition could be said to frequently yield intraindividual correlations that are problematic for the Hotelling T^2.

Results

Tests of Mean Equivalence

Empirical Type I Error Rates

The Hotelling T^2 maintained adequate Type I error rates (.033 - .063) across all conditions (Table 4). As expected, the two pairwise procedures were prone to conservatism. For both procedures, the degree of conservatism was exacerbated as k increased. For both, the empirical Type I error rate approached α at higher sample sizes and with wider equivalence intervals (Figure 3). Generally, the conditions with adequate Type I error rates (.025 to .075) were characterized by equivalence interval widths of at least .50 and sample sizes of at least 50. For the unstandardized pairwise, higher standard deviations were associated with more conservatism; for the equicorrelated structure with 0.90 off-diagonals, nominal rates were achieved with three measurement occasions.

Empirical Power Rates

For all tests of mean equivalence, higher sample sizes increased power (Figure 4), more measurement occasions slightly decreased power (Figure 5), and wider equivalence intervals increased power (Figure 5). Generally, the Hotelling T^2 had the power advantage, followed by the standard-

		Rejection Rates		
Measurement Occasions	Off-diagonal Correlations	$n = 10$	$n = 50$	$n = 90$
3	0.25	0.045	0.051	0.043
3	0.50	0.046	0.045	0.045
3	0.90	0.063	0.048	0.051
5	0.25	0.052	0.051	0.041
5	0.50	0.052	0.049	0.045
5	0.90	0.033	0.055	0.050
7	0.25	0.061	0.055	0.045
7	0.50	0.053	0.049	0.055
7	0.90	0.052	0.049	0.049

Table 4: Type I error rates for the Hotelling T^2. Results belong to conditions with the equivalence interval width of 0.50. Off-diagonal values pertain to the identical elements of the population correlation matrix.

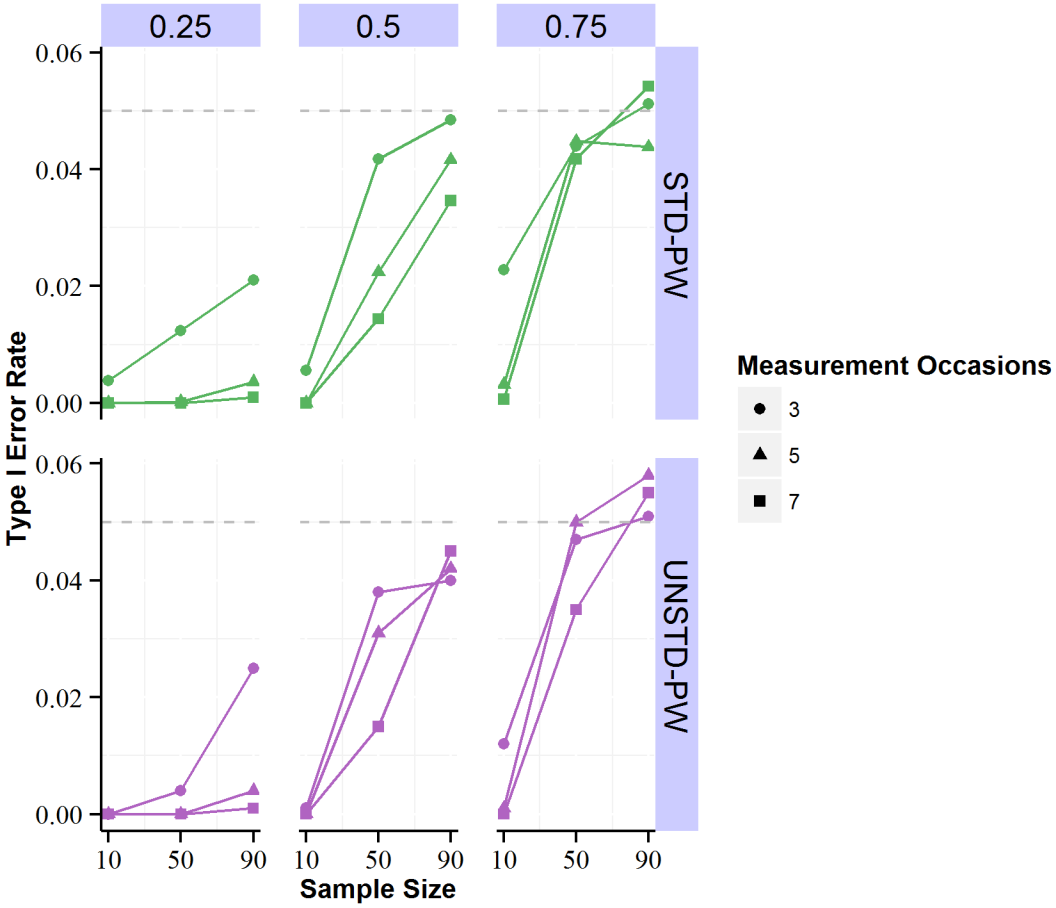

Figure 3: Type I error rates for the pairwise procedures. Vertical panels show conditions for the equivalence intervals of 0.25, 0.50, and 0.75. Data points belong to conditions with equal standard deviations of one and the equicorrelated structure with .50 off-diagonals. STD-PW: Standardized pairwise test. UNSTD-PW: Unstandardized pairwise test.

ized pairwise and unstandardized pairwise.

The tests that had power rates that were invariant to correlation structure and standard deviations were the Hotelling T^2 and standardized pairwise tests. Figure 4 shows that with increased σ_0, power for the unstandardized pairwise dropped dramatically; only with the lowest standard deviation did the unstandardized pairwise procedure yield comparable power to the Hotelling T^2 and standardized pairwise tests. The unstandardized pairwise was also sensitive to correlation structure. Among the equicorrelated structures, the starkest differences in power for the unstandardized test are equicorrelated at .25 and at .90. For non-equicorrelated structures, power rates also tended to be higher for Mixed B, which generally had stronger associations between the measurement occasions.

Notable cases included the equicorrelated structure with off-diagonals of .90 and cases in which measurement occasions had σ_0 of 0.70. With the equicorrelated 0.90 structure, the unstandardized pairwise test had optimal power compared to the others. As well, though narrower equivalence intervals tended to yield low power for all other tests, the unstandardized pairwise test maintained adequate power and outperformed the others when applied under this particular correlation structure. With low standard deviations of 0.7, the unstandardized pairwise test has better power relative to the other tests. In these cases, the test maintained high power across the tested sample sizes, number of measurement occasions, standard deviations, and equivalence interval width — even in spite of the test's conservatism.

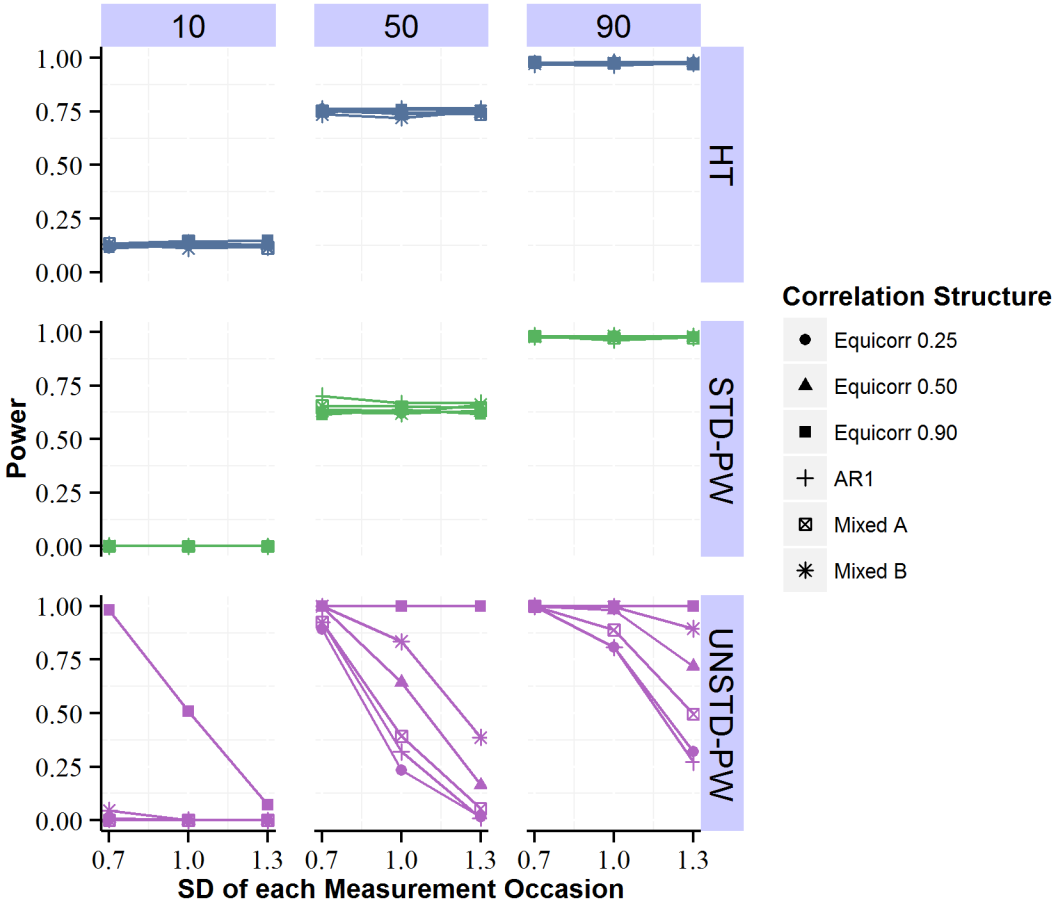

Figure 4: Effects of σ_0, sample size, and correlation structure on power rates. Vertical panels show the conditions for sample sizes of 10, 50, and 90. Data points belong to conditions with five measurement occasions and an equivalence interval of 0.50. HT: Hotelling T^2. STD-PW: standardized pairwise test. UNSTD-PW: unstandardized pairwise test.

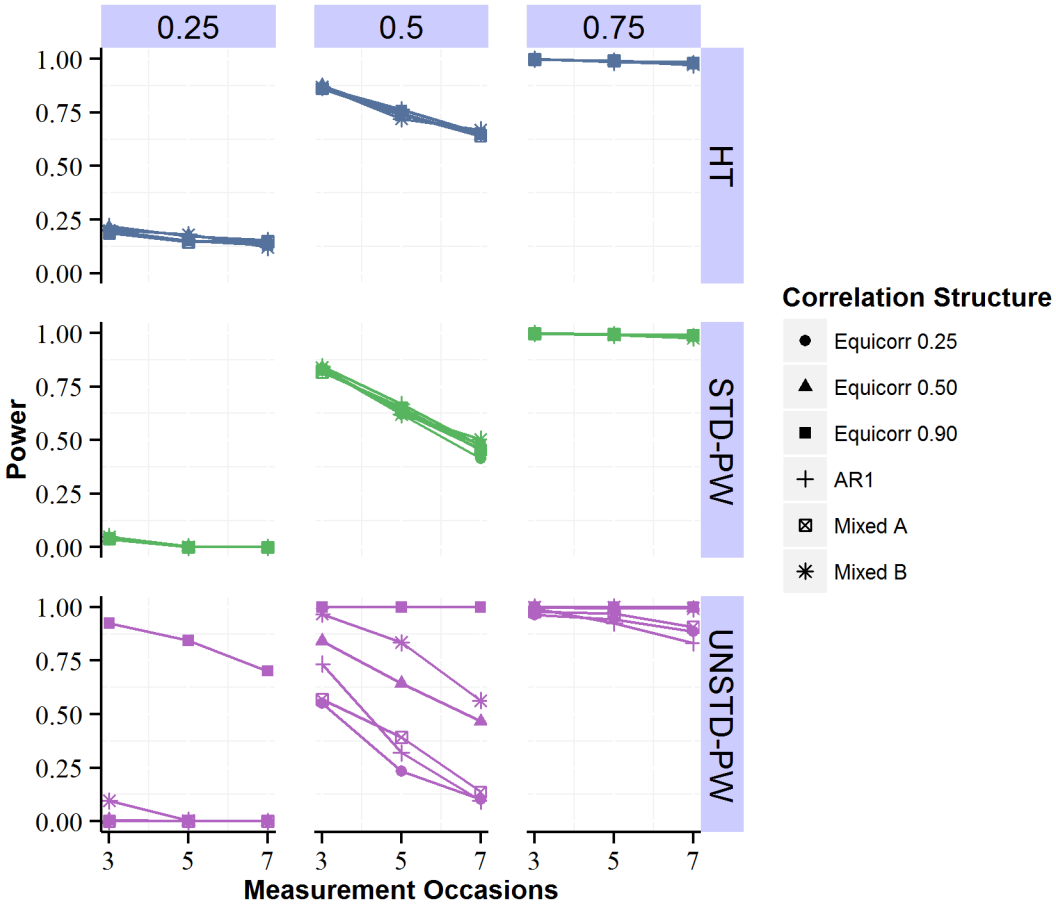

Figure 5: Effects of measurement occasions, equivalence interval width, and correlation structure on power rates. Vertical panels show conditions for the equivalence intervals of 0.25, 0.50, and 0.75. Data points belong to conditions with sample size of 50 and equal standard deviations of 1. HT: Hotelling T^2. STD-PW: standardized pairwise test. UNSTD-PW: unstandardized pairwise test.

TOST for Negligible Trend

Empirical Type I Error Rates

The TOST for negligible trend had Type I error rates that were approximately accurate. For n = 10, Type I error reached a maximum of .031; for n = 50, .056; for n = 90, .063, all of which are below the acceptable upper bound (Figure 6). Although Figure 6 seems to suggest that rates are higher with a wider equivalence interval and a higher sample size, an ad hoc simulation using n = 200, an equivalence width of 0.75, and k = 3 did not show inflated rates.

Empirical Power Rates

Similar to the tests for mean equivalence, the test of negligible trend has higher power with higher sample sizes and wider equivalence interval widths. Further, with more measurement occasions, the TOST for trend had increased power (Figure 7).

Correlation of Difference Scores

For each of the 27 tested conditions of correlation structures, an empirical sampling distribution of the intraindividual correlations, $|r_{diff}|$, was obtained. To summarize these empirical distributions, proportions of $|r_{diff}|$ were calculated for the following ranges of correlational magnitude: $\left(|r_{diff}| \leq .30\right)$, $\left(.30 < |r_{diff}| \leq .50\right)$, $\left(.50 < |r_{diff}| \leq .70\right)$, and $\left(|r_{diff}| > .70\right)$. It was noted earlier that those intraindividual correlations with high magnitude could be considered problematic.

Across the 27 tested correlation structures, the mean difference score correlations ranged from

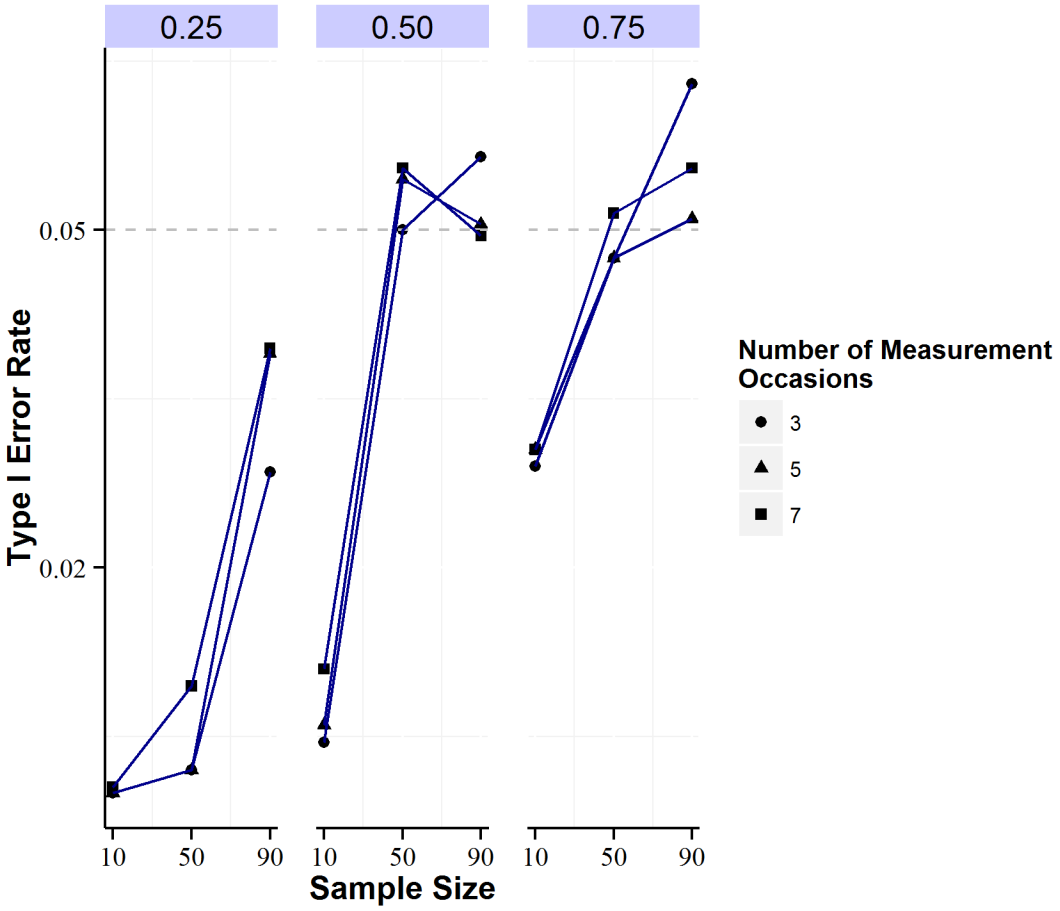

Figure 6: Type I error rates of TOST for negligible trend. Vertical panels show conditions for the equivalence intervals of 0.25, 0.50, and 0.75.

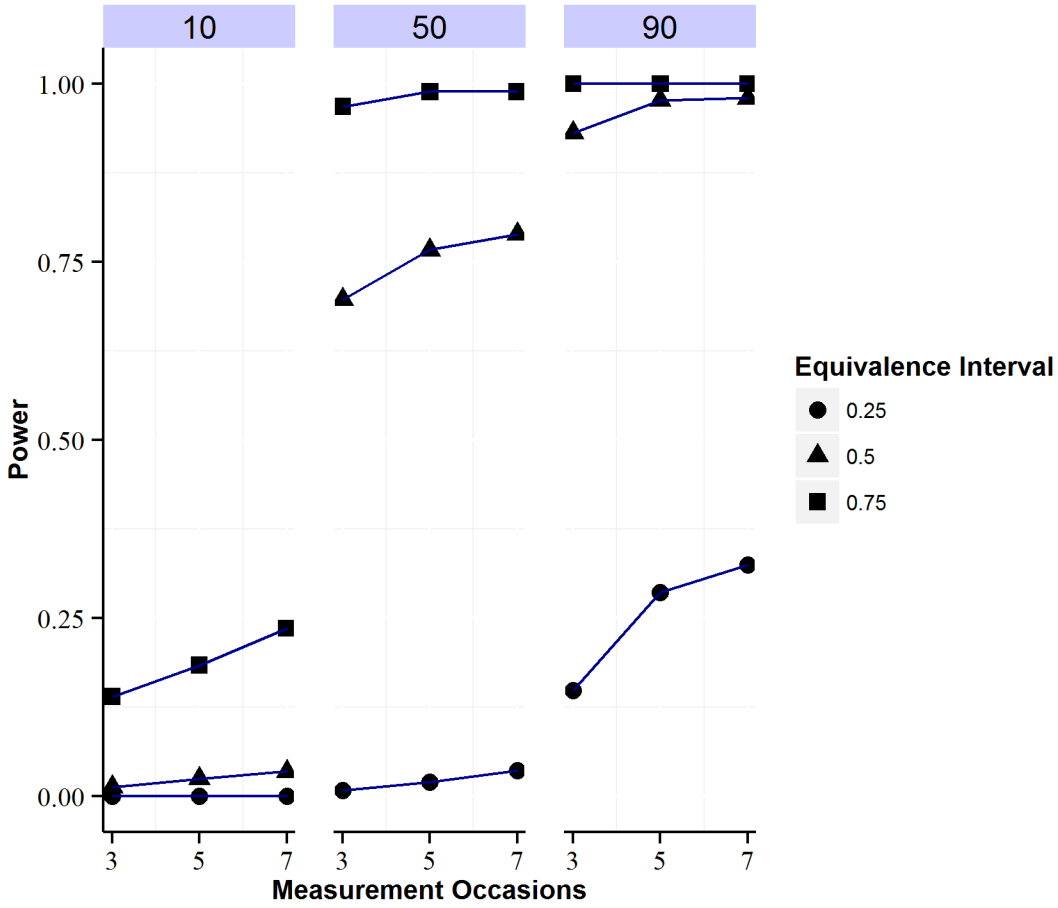

Figure 7: Power rates of TOST for negligible trend. Vertical panels show the conditions for sample sizes of 10, 50, and 90.

.06 to .97. Of the 27 tested conditions, four yielded distributions of intraindividual correlations that centered around relatively high magnitudes, from .50 to .70, and 11 conditions yielded centers greater than .70.

Tables 5 and 6 show the population correlation elements and a descriptive analysis of the resulting sampling distributions for the sample intraindividual correlations, sorted by the maximum and mean values of $|r_{diff}|$.

ρ_{12}	ρ_{13}	ρ_{23}	$0 \leq \lvert r_{diff} \rvert \leq .30$	$.30 < \lvert r_{diff} \rvert \leq 50$	$.50 < \lvert r_{diff} \rvert \leq 70$	$.70 < \lvert r_{diff} \rvert \leq 1.00$	Maximum $\lvert r_{diff} \rvert$	Mean $\lvert r_{diff} \rvert$
.10	.90	.50				1.00	.98	.97
.50	.90	.10				1.00	.98	.97
.10	.90	.10				1.00	.96	.94
.50	.90	.50				1.00	.93	.90
.90	.90	.50				1.00	.89	.86
.50	.90	.90				1.00	.89	.86
.10	.50	.90				1.00	.88	.83
.90	.50	.10				1.00	.88	.83
.10	.90	.90			.12	.88	.80	.72
.10	.50	.10			.16	.84	.80	.72
.90	.90	.10			.13	.87	.79	.72
.50	.50	.10			.89	.11	.76	.67
.90	.10	.50			.89	.11	.75	.67
.10	.50	.50			.89	.11	.75	.67
.50	.10	.90			.89	.11	.75	.67

Table 5: Distribution for correlation of difference scores (Part i). ρs are population correlations. $\lvert r_{diff} \rvert$ is the sample correlation magnitude of the difference scores. Blank space refer to proportions of zero. This table is continued on Table 6.

ρ_{12}	ρ_{13}	ρ_{23}	$0 \leq \lvert r_{diff} \rvert \leq .30$	$.30 < \lvert r_{diff} \rvert \leq 50$	$.50 < \lvert r_{diff} \rvert \leq 70$	$.70 < \lvert r_{diff} \rvert \leq 1.00$	Maximum $\lvert r_{diff} \rvert$	Mean $\lvert r_{diff} \rvert$
.50	.50	.50		.49	.51		.63	.50
.90	.90	.90		.49	.51		.62	.50
.10	.10	.10		.50	.50		.61	.50
.90	.50	.90		.83	.17		.60	.47
.50	.10	.10	.03	.97			.50	.37
.10	.10	.50	.04	.96			.50	.37
.90	.50	.50	.96	.04			.37	.22
.50	.50	.90	.96	.04			.37	.22
.10	.10	.90	1.00				.34	.17
.90	.10	.10	1.00				.30	.17
.50	.10	.50	1.00				.24	.10
.90	.10	.90	1.00				.24	.06

Table 6: Distribution for correlation of difference scores (Part ii). ρs are population correlations. $|r_{diff}|$ is the sample correlation magnitude of the difference scores. Blank space refer to proportions of zero.

Discussion

Equivalence tests address the logical and statistical issues of the inappropriate application of tests of differences, and the presence of several psychology studies which test for stability or lack of change indicates that potential solutions require explication. With the criteria that the ideal test has nominal Type I error, adequate power, and relative invariance across data situations, the current study empirically evaluated three tests of mean equivalence and one test of negligible trend under the same set of data characteristics. An interpretation of the results provides preliminary suggestions as to how researchers can use the presently available methods.

Comparing the Tests

Out of all the evaluated tests, the TOST for negligible trend satisfies all criteria reasonably well, provided that an appropriate model is fitted. The simulation, using the linear mixed model in tandem with the test, showed that the Type I error rate for the null hypothesis of non-equivalence is conservative for lower sample sizes. This may not be overly concerning, though, because mixed models are asymptotic procedures that do require high sample sizes for sensible inferential tests. Overall, results from the current study support Dixon and Pechmann's (2005) assertion that estimates of slopes, standard errors, and the approximated degrees of freedom from the mixed model,

as well as time-series models, are appropriate for the TOST.

The trend test has, unlike the tests of mean equivalence, higher power with more measurement occasions, though the effect is modest. This aspect aligns with the usual desire for more measurement occasions in longitudinal studies (Singer & Willett, 2003). In modeling change, more repeated measures allows for better reliability in assessing change; if one were interested in a reliable statement of stability over time, then more measures would both allow for higher reliability and greater power for detecting equivalence. Further, there is greater flexibility in the types of longitudinal questions that may be addressed. Because the procedure's tested parameter of the slope may be obtained through the linear mixed model, additional research questions about individual-level variability and stability can be addressed. With a random intercept model, as was presented in the simulation, the TOST for negligible trend addresses the question of whether there is aggregate-level stability of some response variable. However, if one were to implement a random intercept and random slope, as is commonly done in practice, then one could determine both the presence of aggregate-level stability and of individual-level stability. For aggregate-level stability, one could still use the mean slope from the random intercept and random slope model; for individual-level stability, one could examine individual-level coefficients (i.e., individual-level, random slopes) to determine the proportion of the sample that had individual slopes falling within the equivalence interval.

Although linear mixed models are applicable for within-subject variables that are unordered factors, such as experimental conditions, the goal for overall equivalence renders the TOST for negligible trend inapplicable. For k unordered, categorical conditions, there would be $k - 1$ slope

estimates, corresponding to $k - 1$ mean differences between conditions. To establish overall equivalence, all pairwise mean differences would need to be established; if it were to become a pairwise problem, then there would be issues regarding conservatism or multiplicity control. Thus, when one is interested in the equivalence of unordered groups for a factor, an analysis on aggregate means would be more intuitively applied.

Of the tests of mean equivalence, the Hotelling T^2 test is the only one that has nominal Type I error and adequate power rates that are invariant across correlation structures and σ_0. Its general advantage in power over the pairwise procedures is also in line with previous studies about the equivalence of k independent measurements, showing that omnibus procedures have more power over pairwise procedures in equivalence testing settings.

The limitation of the Hotelling T^2's theoretical equivalence region requires discussion. As the magnitudes of intraindividual correlations increase, the equivalence region becomes increasing elongated, making interpretations problematic. From the supplementary simulation for the correlation of difference scores, the observed patterns in even this small subset of conditions indicates that this problem of the elongated equivalence region would likely be prevalent in the applied setting. Further, there appears to be no clear, systematic pattern for which one would observe or predict the pattern of intraindividual correlations from the correlation matrix itself.

Despite its Type I error and power rates, the Hotelling T^2 cannot be wholly recommended as the default solution because its equivalence region depends upon the correlation of the intraindividual differences. This caveat has the consequence that conditions yielding highly correlated differences scores may involve less intuitive interpretations. Still, it is possible that the correlation structures

seen in practice do yield correlations of differences that are unproblematic; indeed, results for the tested subset of conditions do show intraindividual correlations of less than .50. In application, one may consider trade-offs to harness the Type I and power rates of the Hotelling T^2 while also being cautious of intraindividual difference correlations. For example, with an equivalence interval of .50 and three measurement occasions, one may accept intraindividual difference correlations of .50, for its equivalence region is not considered strongly elliptic. Having observed whether the correlations of intraindividual differences from the observed correlation matrix are acceptable, one may then decide to proceed with the Hotelling T^2. For this, it would be possible to calculate the difference scores for adjacent means and to obtain the intraindividual correlation matrix. However, with more than three measurement occasions, the nature of the equivalence region is not obvious.

If one decides that the advantages of Type I error and power of the Hotelling T^2 do not supercede the interpretative limitations of its equivalence region, then researchers could consider the pairwise procedures. The choice between these two pairwise procedures may be made on the basis of comparing their Type I error rates, power rates, and choice of distributional parameter.

The extent of conservatism differed little for the two pairwise procedures, as both have Type I error rates that do approach nominal levels at higher sample sizes. However, it should be noted that this occurs only in the best case scenario, in which only one pairwise mean difference falls at the equivalence boundary (i.e., all other pairwise mean differences are, by themselves, power conditions); if more than one pairwise mean difference falls outside the boundary (which also yield Type I error conditions), then conservatism increases dramatically.

Power rates differ markedly between the standardized and unstandardized forms. Generally

higher powered than the unstandardized, the standardized test also has the additional advantage that power rates are invariant to the standard deviations of each measurement occasion and correlation structures. In contrast, the sensitivity of the unstandardized test to standard deviations and correlation structures makes for less predictable power estimation.

However, the unstandardized test does have some advantage for certain conditions. Comparisons within the sets of equicorrelated conditions (particularly those with diagonals of .25 and .90) and non-equicorrelated conditions (AR, Mixed A, Mixed B) suggest that conditions for highly related measurement occasions allow for higher power. This is even more pronounced when occasions are measured to have low spread in which the standard deviation is small. In these conditions, the unstandardized test has much higher power and is comparable to the Hotelling T^2 — even the diminishing effects of low sample sizes, narrow equivalence intervals, more measurement occasions, and high standard deviations no longer pose limitations. With highly correlated measurement occasions, the unstandardized test is highly powered in almost all scenarios.

The effects of the magnitude of correlation can be explained. With a strong pairwise association, the standard error decreases, which makes it easier for each pairwise test statistic to fall in the appropriate critical region. Indeed, Equation 19 indicates that an increase in r corresponds to a larger rejection region. This finding for the unstandardized pairwise test is thus a generalization of Mara and Cribbie's (2012) finding that the TOST-based paired samples test holds advantage over the standardized pairwise test when scores are highly correlated. In practice, however, it is highly unlikely to observe equicorrelated structures. Mixed correlation structures, whose dimensions and elements also vary with the number of measurement occasions, do not have the same predictability

in power as compared to the equicorrelated structures. Nevertheless, it is useful to be aware of the general situation in which the unstandardized would be most advantageous.

The stark differences between the two pairwise procedures can be related back to the choice of test parameter. From a theoretical perspective, Wellek (2010) has argued that the standardized difference is a more useful metric in equivalence testing. For some large, raw mean difference between two population distributions, the distributions are well distinguished when the variance of these distributions decrease; but, when the variances of these distribution increases, the two populations become increasingly indistinguishable (Lehmann & Romano, 2005; Wellek, 2010). On the other hand, it would seem more intuitive, particularly in the behavioural sciences, to choose an equivalence interval based on raw mean differences, especially when such a metric may be more interpretable in a practical context (Wilkinson, 1999). The choice, then, relates to the parameter that one chooses to make inference to and to the statistical properties one prefers to gain when choosing one test over the other.

Generally, when considering power for detecting equivalence, researchers should consider the factors of sample size, measurement occasions, precision, correlation structure, and equivalence interval width. A cautionary note must be made about the equivalence interval. Unlike the other factors studied here, the equivalence interval width is a characteristic not of the data nor the research design but is rather a decision to be made by the researcher and the context of his field. Equivalence intervals are often more well-defined (e.g., the difference in drug efficacy that is deemed negligible) outside the social sciences but less so in psychology; psychologists must be careful to examine previous literature for effect sizes particular to their scales in use. Because there is some

subjectivity inherent in the process of choosing an equivalence interval in psychology, one must be careful not to let the promise of higher power influence the choice of equivalence interval.

Recommendations

Based on empirical results, suggestions may be made about test selection. Overall, the researcher should consider the nature of the data (particularly regarding whether the predictor is ordered), the data's spread, and the correlation structure. When the nature of the within-subject variable is ordered (such as time), the TOST for negligible trend would be best, particularly because it can be used in conjunction with other flexible modeling procedures (at least with linear mixed models and time-series models). When the within-subject variable is unordered, such as with experimental conditions, tests of mean equivalence should be applied.

An initial examination of the correlation structure for the tests of mean equivalence would be helpful. For the case of three measurement occasions, the Hotelling T^2 may be used after determining that intraindividual correlations of differences imply relatively little elongation of the equivalence region. For a rough guideline, a low magnitude for the intraindividual difference correlation ($|r_{diff}| < .50$) should be acceptable for the use of the Hotelling T^2, but for magnitudes that exceed .50 (or some higher cut-off), one should be wary of the Hotelling T^2. If the Hotelling T^2 is unsound, then one may consider pairwise procedures. Generally, the standardized test is recommended over the unstandardized test for its general advantage in power and for its invariance to data situations. However, if one observes that values for the measurement occasions are highly correlated, then the unstandardized test would be optimal. In any of these cases, the choice of

equivalence interval should not be influenced by the desire for higher power. Overall, the choice of test depends on the measurement level of the repeated outcome, precision, and correlation structure.

Bibliography

Berger, R. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, *24*(4), 295–300.

Berger, R., & Hsu, J. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, *11*(4), 283–319. doi: doi:10.1214/ss/1032280304

Blair, R. C., & Cole, S. R. (2002). Two-sided equivalence testing of the difference between two means. *Journal of Modern Applied Statistical Methods*, *1*(1), 139–142.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152. doi: 10.1111/j.2044-8317.1978.tb00581.x

Buist, K. L., Reitz, E., & Deković, M. (2004). Stability and changes in problem behavior during adolescence: latent growth analysis. *Journal of Youth and Adolescence*, *33*(1), 1–12. doi: 10.1023/A:1027305312204

Center for Statistical Consultation and Research. (2011). *Simulation of random intercept models in R.* Retrieved from `http://goo.gl/N3IjYY`

Cribbie, R., Arpin-Cribbie, C., & Gruman, J. (2009). Tests of equivalence for one-way independent groups designs. *The Journal of Experimental Education*, *78*(1), 1–13. doi: 10.1080 / 00220970903224552

Davidson, R. J., Kabat-Zinn, J., Schumacher, J., Rosenkranz, M., Muller, D., Santorelli, S. F., ... Sheridan, J. F. (2003). Alterations in brain and immune function produced by mindfulness meditation. *Psychosomatic medicine*, *65*(4), 564–570. doi: 10.1097/01.PSY.0000077505.67574.E3

Dixon, P. M., & Pechmann, J. H. K. (2005). A statistical test to show negligible trend. *Ecology*, *86*(7), 1751–1756. doi: 10.1890/04-1343

Graney, M. J., & Engle, V. F. (2000). Stability of performance of activities of daily living using the MDS. *The Gerontologist*, *40*(5), 582–6. doi: 10.1093/geront/40.5.582

Gruman, J., Cribbie, R., & Arpin-Cribbie, C. (2007). The effects of heteroscedasticity on tests of equivalence. *Journal of Modern Applied Statistical Methods*, *6*(1), 133–140.

Joseph, J. E., Cerullo, M. A., Farley, A. B., Steinmetz, N. A., & Mier, C. R. (2006). fMRI correlates of cortical specialization and generalization for letter processing. *NeuroImage*, *32*(2), 806–820. doi: 10.1016/j.neuroimage.2006.04.175

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 983–997. doi: 10.2307/2533558

Koh, A., & Cribbie, R. (2013). Robust tests of equivalence for k independent groups. *The British Journal of Mathematical and Statistical Psychology*, *66*(3), 426–34. doi: 10.1111/j.2044-8317.2012.02056.x

Laska, E. M., & Meisner, M. J. (1989). Testing whether an identified treatment is best. *Biometrics*, *45*(4), 1139–1151. doi: 10.2307/2531766

Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses*. New York, NY: Springer

Science & Business Media.

Mara, C., & Cribbie, R. (2012). Paired-samples tests of equivalence. *Communications in Statistics - Simulation and Computation*, *41*(10), 1928–1943. doi: 10.1080/03610918.2011.626545

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

Rizopoulos, D. (2007). *[R] Create an AR(1) covariance matrix.* Retrieved from goo.gl/m66ZRe

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), 657–680. doi: 10.1007/BF01068419

Seaman, M. A., & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods*, *3*(4), 403–411. doi: 10.1037/1082-989X.3.4.403

Singer, J. D., & Willett, J. B. (2003). Introducing the multilevel model for change. *Applied longitudinal data analysis: Modeling change and event occurence*, 45–74.

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2nd ed.). Boca Raton, FL: CRC Press.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594. doi: 10.1037/0003-066X.54.8.594

Winston, J. S., Gottfried, J. A., Kilner, J. M., & Dolan, R. J. (2005, September). Integrated neural representations of odor intensity and affective valence in human amygdala. *The Journal of Neuroscience*, *25*(39), 8903–7. doi: 10.1523/JNEUROSCI.1569-05.2005