Journal of Modern Applied Statistical Methods

Volume 2 | Issue 2 Article 6

11-1-2003

A Comparison Of Equivalence Testing In Combination With Hypothesis Testing And Effect Sizes

Christopher J. Mecklin
Murray State University, christopher.mecklin@murraystate.edu

DOI: 10.22237/jmasm/1067645160

Journal of Modern Applied Statistical Methods
November, 2003, Vol. 2, No. 2, 329-340

1538 – 9472/03/$30.00

A Comparison Of Equivalence Testing In Combination With Hypothesis Testing And Effect Sizes

Christopher J. Mecklin
Department of Mathematics and Statistics
Murray State University

Equivalence testing, an alternative to testing for statistical significance, is little used in educational research. Equivalence testing is useful in situations where the researcher wishes to show that two means are not significantly different. A simulation study assessed the relationships between effect size, sample size, statistical significance, and statistical equivalence.

Key words: Equivalence testing, statistical significance, effect size

Introduction

The use of statistical inference, particularly via null hypothesis significance testing, is an extremely common but contentious practice in educational research. Both the pros and the cons of hypothesis testing have been argued in the literature for several decades. A recent monograph edited by Harlow, Mulaik, and Steiger (1997) was devoted to these arguments. Some classic references criticizing standard hypothesis testing include Boring (1919), Berkson (1938, 1942), Rozeboom (1960), Meehl (1967, 1978), and Carver (1978). More recently, some support the continued usage of significance testing (Abelson, 1997; Hagan, 1997, 1998; Harris, 1997; McLean & Ernest, 1998), while others desire a greater reliance on alternatives such as confidence intervals or effect sizes (Cohen, 1992, 1994; Knapp, 1998, 2002; Meehl, 1997; Serlin, 2002; Thompson, 1998, 2001; Vacha-Haase, 2001), and still others advocate an outright ban on significance testing (Carver, 1993; Falk, 1998; Hunter, 1997; Nix & Barnette, 1998; Schmidt & Hunter, 1997).

Christopher Mecklin is an Assistant Professor of Mathematics & Statistics. His research interests include goodness-of-fit, educational statistics and statistical ecology. He enjoys working with faculty and students from various disciplines. Email:christopher.mecklin@murraystate.edu.

The references included here are by no means close to an exhaustive list. This debate is not limited to educational research and the social sciences; for instance, it is also being argued in ecology (McBride, 1999; Anderson, Burnham, & Thompson, 2000). Many in the statistical community outside of the niche of educational and psychological research, though, are either unaware of this debate or feel that it is trivial (Krantz, 1999).

The objective of this paper is not to continue this heated argument, but rather to borrow the method of equivalence testing from biostatistics, as suggested by Bartko (1991), and using it in conjunction with standard hypothesis testing in educational research. Lehmann (1959) anticipated the need for interval testing in his classic volume on the theory of hypothesis testing. Many of the currently employed methods of equivalence testing were developed in the 1970's and 1980's to address biostatistical and pharmaceutical problems (Westlake, 1976, 1979; Schuirmann, 1981, 1987; Anderson & Hauck, 1983; Patel & Gupta, 1984). Rogers, Howard, and Vessey (1993) introduced the use of equivalence testing methods to the social sciences. Serlin (1993) essentially suggested equivalence testing when he suggested the use of range, rather than point, null hypotheses.

Methodology

Standard null hypothesis significance testing dates back to the pioneering theoretical work of

Fisher, Neyman, and Pearson. Hypothesis testing can be found in almost every textbook of statistical methods and thus will not be further elaborated on here. Equivalence testing, on the other hand, is a newer technique and one that is unfamiliar to most researchers in education and the social sciences.

Equivalence testing was developed in biostatistics to address the situation where the goal is not to show that the mean of one group is greater than the mean of another group (i.e. the superiority of one treatment to another), but rather to establish that two methods are equal to one another. A common application of this idea in biostatistics is to show that a less expensive "generic" medication is as effective as the more expensive "brand-name" medication. In equivalence testing, the null hypothesis is that the two groups are not equivalent to one another, and hence rejection of the null indicates that the two groups are equivalent. This differs from standard significance testing where the null hypothesis states that the group means are equal and rejection of the null indicates that the two groups are statistically different. A common methodological mistake in research is to conclude that the null hypothesis is true (i.e. two groups have equal means) based on the failure to reject it. This action fails to recognize that the failure to reject the null is often merely a Type II error, especially when the sample sizes are small and the power of the test is low.

An explanation of the theory of equivalence testing can be found in Berger and Hsu (1996); Blair and Cole (2002) give a less technical explanation. Here, we will merely review the most commonly implemented method used for establishing the equivalence of two population means for an additive model, where the difference of means is considered. The multiplicative model, which looks at the ratio of means, will not be considered further in this paper. The commonly used procedure in biostatistics for this problem is to use the "two one-sided tests" procedure, or TOST (Westlake, 1976, 1979; Schuirmann, 1981, 1987). With the TOST, the researcher will consider two groups equivalent if he can show that they differ by less than some constant τ, the equivalence bound, in both directions. The constant τ is often chosen to be a percentage (such as 10% or 20%) of the mean of the control group, although τ can also be chosen to be a constant that is the smallest absolute difference between two means that is large enough to be practically important.

The null hypothesis (i.e. the means are different) for the TOST is $H_0 : |\mu_1 - \mu_2| \geq \tau$. The alternative hypothesis (i.e. the means are equivalent) is $H_1 : |\mu_1 - \mu_2| < \tau$.

The first one-sided test seeks to reject the null hypothesis that the difference between two means is less than or equal to $-\tau$; similarly, the second one-sided test seeks to reject the null hypothesis that the difference in the means is greater than or equal to τ. If the one-sided test with the larger p-value leads to rejection, then the two groups are considered to be equivalent.

For the first one-sided test, we compute the test statistic

$$t_1 = \frac{x_1 - x_2 + \tau x_2}{s_p \sqrt{1/n_1 + 1/n_2}}$$

where s_p is the pooled standard deviation of the two samples and compute the p-value as

$$p_1 = P(t_\nu > t_1)$$

where t_ν is a random variable from the t-distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom.

The second one-sided test is similar to the first. The test statistic is

$$t_2 = \frac{x_1 - x_2 - \tau x_2}{s_p \sqrt{1/n_1 + 1/n_2}}$$

and the p-value is

$$p_2 = P(t_\nu < t_2).$$

If we let $p = \max(p_1, p_2)$, then the null hypothesis of nonequivalence is rejected if $p < \alpha$.

The choice of τ is a difficult choice that is up to the researcher. This choice is analogous to the selection of an appropriate alpha level in

standard significance testing, an appropriate level of confidence in interval estimation, or a sufficiently large effect size, and should be made carefully. Knowledge of the situation at hand should be used to specify the maximum difference between population means that would be considered clinically trivial. Researchers in biostatistics typically have the choice made for them by government regulation.

As in standard hypothesis testing, an equivalency confidence interval can also be constructed. If the entire confidence interval is within $(-\tau, \tau)$, then equivalence between the groups is indicated. If the entire confidence interval is within either $(-\tau, 0)$ or $(0, \tau)$ (i.e. zero is not in the interval), then we would reject the null hypotheses of both a significance and an equivalence test. In that case, we could make the somewhat discomforting conclusion that the difference of means was both statistically significant and equivalent.

It is important to note that the equivalency confidence interval is expressed at the $100(1-2\alpha)\%$ level of confidence. Rogers et al. (1993) noted that if one performs both a standard significance test and an equivalence test on the same data set, making either a "reject" or "fail to reject" decision, that there are four possibilities. These four conditions are given in Table 1.

Table 1. Possible Combinations of Significance and Equivalence Testing

Significance Test	Equivalence Test	Term
Fail to reject	Reject	Equivalent
Reject	Reject	Equivalent and Different
Reject	Fail to reject	Different
Fail to reject	Fail to reject	Equivocal

The second condition "equivalent and different", a simultaneous rejection of both inferential procedures, could happen in a situation where large samples provide "too much power", resulting in a trivial difference in means being statistically significant. The equivalence test (and the effect size) should detect the small magnitude of these mean differences. The fourth condition indicates that there is insufficient evidence to conclude that the groups are either equivalent or different. This would most likely occur when the samples are very small and/or the group variances are very large.

The effect size for the difference of means is the standardized difference between the groups (Fan, 2001). We will use the parameter

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$

to represent the effect size of the population, where μ_1 and μ_2 are the population means and σ^2 is the common variance.

Of course, δ is typically unknown and needs to be estimated. Cohen's d (1988) is a statistic often used for this purpose. The effect size (ES) is found with

$$d = \frac{x_1 - x_2}{s_{pooled}}$$

where

$$s_{pooled} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

is the pooled standard deviation of the two samples. We stress that Cohen's d is a sample statistic and has a sampling distribution like other estimates.

Cohen (1988) gave some suggestions for interpreting d. An effect size of d=0.2 is deemed "small", d=0.5 is "medium", and d=0.8 is "large". It is becoming, rather regrettably in our opinion, common for researchers to rigidly apply Cohen's suggestions. Absolute reliance on Cohen's rule of thumb is as misguided as blind adherence to a particular level of significance

(e.g. $\alpha = 0.05$). As Thompson (2001) said, "we would merely be being stupid in another metric."

Results

Rogers et al. (1993) provided empirical examples of the application of equivalence testing on data from the psychological literature. We will do the same with an example from the educational research literature. This will demonstrate that there often exist situations where a statistically significant difference between groups coincides with the groups being statistically equivalent. This is the "equivalent and different" condition that is typically associated with a small to moderate effect size, as opposed to the strong effect sizes that typically occur with the "different" condition and the weak effect sizes that occur with the "equivalent" condition.

Benson (1989), in a study concerning statistical test anxiety, presented means and variances for a sample of 94 males and 123 females on seven variables. Using standard hypothesis testing methods (i.e. t-tests), significant group differences were found for: prior math courses, math self-concept, self-efficacy, and statistical test anxiety. However, after calculating Cohen's d as an effect size (ES) measure and the use of the TOST equivalence test, we see that only prior math courses and statistical test anxiety are "different" between males and females. Not surprisingly, the two largest effect sizes are found for these two variables. Table 2 shows results of both traditional significance and equivalence tests for the Benson data.

Statistical significance was defined as a rejection of H_0 with $\alpha = 0.05$ and equivalence was defined as a rejection of H_0 with $\alpha = 0.10$. The reason for the two different significance levels is because while a traditional significance test at level α corresponds to a $100(1-\alpha)\%$ confidence interval, an equivalence test at level α corresponds to a $100(1-2\alpha)\%$ equivalence interval. We selected $\tau = 0.2$ (i.e. 20% of the mean of the female group). This choice was arbitrary and by no means should be taken as a choice recommended for all equivalence problems. The results could differ with different choices for τ.

Table 2. Comparing Significance and Equivalence Testing for the Benson Data

Descriptive Statistics								
	Males (N=94)		Females (N=123)					
Variable	M	SD	M	SD	Effect Size	Sig. p-value	Equiv. p-value	Category
GPA	3.05	0.44	3.16	0.47	-0.24	0.040	<0.001	Equiv. & Diff.
Prior Math Courses	3.45	2.14	2.20	2.01	0.60	<0.001	0.998	Different
Math Self-Concept	25.77	5.96	23.20	7.05	0.39	0.002	0.012	Equiv. & Diff.
Self-efficacy	12.68	1.77	11.62	2.30	0.51	<0.001	<0.001	Equiv. & Diff.
General Test Anxiety	36.38	0.49	40.62	12.25	-0.37	0.004	0.007	Equiv. & Diff.
Achievement	32.56	5.68	32.26	7.55	0.04	0.374	<0.001	Equivalent
Statistical Test Anxiety	32.65	12.57	41.84	14.83	-0.66	<0.001	0.663	Different

For a test of statistical significance, power is the probability of rejecting the null hypothesis that the population means are equal when they are in fact not equal. The power of an equivalence test is the probability of rejecting that the means are different by at least some equivalence bound τ when the means are in fact equivalent (i.e. differ by less than τ).

Of interest to us is the probability of rejecting both the null hypotheses (of non-significance and non-equivalence) simultaneously. We designed a small simulation study to assess the power of simultaneously concluding that two means are both statistically different and equivalent.

As is always the case with Monte Carlo studies, the choices of simulation parameters are difficult to make and are somewhat arbitrary. We endeavored to simulate situations that were likely to be encountered in actual quantitative data analysis. We also made some simplifying assumptions to keep the number of simulations and associated tables and figures to a reasonable level.

We assumed that both of our populations were always normally distributed with a common variance $\sigma^2 = 1$. Six different sample sizes per group (n=10,20,50,100,200,500) were chosen; only equally sized groups were used in this study. Six different values for the effect size parameter ($\delta = 0, 0.1, 0.2, 0.3, 0.4, 0.5$) were used, reflecting situations from no effect (i.e. equivalent population means) to a "medium" effect size (i.e. population means that differ by one half of a standard deviation). Three different equivalence bounds ($\tau = 0.1, 0.2, 0.4$) were used, defining the minimum difference between means that is practically important (i.e. non-equivalent) to be 10%, 20% or 40% of μ_1.

Hence, we have a fully crossed design with 6 X 6 X 3 = 108 cells. Within each cell (i.e. combination of sample size, effect size, and equivalence bound), 10000 simulations were run. The R statistical computing environment was used to conduct the simulations. Each simulation consisted of generating *n* random normal variates with mean $0 + \delta$ and variance 1 and a second, independent set of *n* random normal variates with mean 0 and variance 1. The independent samples t-test and the TOST with equivalence bound τ was conducted for each simulation, and the number of rejections of each test, along with the number of simultaneous rejections of both procedures and the number of failures to reject either procedure, were noted.

Tables 3 through 8 show the number of rejections of the null hypotheses of the equivalence test, both tests, the significance test, and neither test. Columns involving the equivalence test are in *italics*; columns involving the significance test are in **boldface**. Note that the power of the equivalence test for each situation can be found by dividing the sum of the italicized columns by 10000. Similarly, the power of the significance test is obtained by dividing the sum of the columns in boldface by 10000.

Table 3. Simulated Power of the Tests of Statistical Equivalence and Significance, Effect Size $\delta = 0$

Equivalence Bound τ	Sample Size N	Number of Rejections (10000 Simulations) *Equivalent*	***Both***	**Different**	Neither
0.1	10	*0*	***0***	**506**	9494
	20	*0*	***0***	**500**	9500
	50	*0*	***0***	**476**	9524
	100	*0*	***0***	**535**	9465
	200	*0*	***0***	**504**	9496
	500	*2337*	***0***	**511**	7152
0.2	10	*0*	***0***	**496**	9504
	20	*0*	***0***	**507**	9493
	50	*0*	***0***	**485**	9515
	100	*1063*	***0***	**546**	8391
	200	*5121*	***0***	**514**	4365
	500	*9386*	***3***	**490**	121
0.4	10	*10*	***0***	**486**	9504
	20	*370*	***0***	**469**	9161
	50	*5279*	***0***	**481**	4240
	100	*8757*	***0***	**457**	786
	200	*9493*	***444***	**63**	0
	500	*9483*	***517***	**0**	0

Table 4. Simulated Power of the Tests of Statistical Equivalence and Significance, Effect Size $\delta = 0.1$

Equivalence Bound τ	Sample Size N	Number of Rejections (10000 Simulations) *Equivalent*	***Both***	**Different**	Neither
0.1	10	*0*	***0***	**535**	9494
	20	*0*	***0***	**606**	9500
	50	*0*	***0***	**817**	9524
	100	*0*	***0***	**1118**	9465
	200	*0*	***0***	**1652**	9496
	500	*709*	***0***	**3366**	7152
0.2	10	*0*	***0***	**521**	9504
	20	*0*	***0***	**605**	9493
	50	*1*	***0***	**786**	9515
	100	*793*	***0***	**1090**	8391
	200	*3452*	***0***	**1687**	4365
	500	*6192*	***15***	**3486**	121
0.4	10	*11*	***0***	**565**	9424
	20	*347*	***0***	**622**	9031
	50	*4759*	***0***	**772**	4469
	100	*7902*	***0***	**1044**	1054
	200	*8361*	***1196***	**443**	0
	500	*6521*	***3475***	**4**	0

Table 5. Simulated Power of the Tests of Statistical Equivalence and Significance, Effect Size $\delta = 0.2$

Equivalence Bound τ	Sample Size N	Number of Rejections (10000 Simulations) *Equivalent*	***Both***	**Different**	Neither
0.1	10	*0*	***0***	**727**	9273
	20	*0*	***0***	**962**	9038
	50	*0*	***0***	**1727**	8273
	100	*0*	***0***	**2865**	7135
	200	*0*	***0***	**5193**	4807
	500	*16*	***0***	**8880**	1104
0.2	10	*0*	***0***	**699**	9301
	20	*0*	***0***	**950**	9050
	50	*0*	***0***	**1678**	8322
	100	*408*	***0***	**2908**	6684
	200	*951*	***0***	**5207**	3842
	500	*915*	***7***	**8924**	154
0.4	10	8	***0***	**734**	9258
	20	*296*	***0***	**967**	8737
	50	*3397*	***0***	**1677**	4926
	100	*5485*	***0***	**2890**	1625
	200	*4886*	***2800***	**2314**	0
	500	*1167*	***8534***	**299**	0

Table 6. Simulated Power of the Tests of Statistical Equivalence and Significance, Effect Size $\delta = 0.3$

Equivalence Bound τ	Sample Size N	Number of Rejections (10000 Simulations) *Equivalent*	***Both***	**Different**	Neither
0.1	10	*0*	***0***	**947**	9053
	20	*0*	***0***	**1540**	8460
	50	*0*	***0***	**3144**	6856
	100	*0*	***0***	**5594**	4406
	200	*0*	***0***	**8482**	1518
	500	*0*	***0***	**9973**	27
0.2	10	*0*	***0***	**985**	9015
	20	*0*	***0***	**1501**	8499
	50	0	***0***	**3203**	6797
	100	*104*	***0***	**5681**	4215
	200	*95*	***0***	**8524**	1381
	500	*19*	***1***	**9973**	7
0.4	10	*11*	***0***	**991**	8998
	20	*225*	***0***	**1563**	8212
	50	*2061*	***0***	**3133**	4806
	100	*2796*	***0***	**5602**	1602
	200	*1516*	***2374***	**6110**	2167
	500	*23*	***6115***	**3862**	0

Table 7. Simulated Power of the Tests of Statistical Equivalence and Significance, Effect Size $\delta = 0.4$

Equivalence Bound τ	Sample Size N	Number of Rejections (10000 Simulations) *Equivalent*	***Both***	**Different**	Neither
0.1	10	*0*	***0***	**1335**	8665
	20	*0*	***0***	**2333**	7667
	50	*0*	***0***	**5015**	4985
	100	*0*	***0***	**8069**	1931
	200	*0*	***0***	**9769**	231
	500	*0*	***0***	**10000**	0
0.2	10	*0*	***0***	**1344**	8656
	20	*0*	***0***	**2341**	7659
	50	0	***0***	**5077**	4923
	100	*23*	***0***	**8110**	1867
	200	*1*	***0***	**9784**	215
	500	*0*	***0***	**10000**	0
0.4	10	*9*	***0***	**1402**	8589
	20	*164*	***0***	**2346**	7490
	50	*933*	***0***	**5099**	3968
	100	*932*	***0***	**8075**	993
	200	*232*	***806***	**8962**	0
	500	*0*	***1025***	**8975**	0

Table 8. Simulated Power of the Tests of Statistical Equivalence and Significance, Effect Size $\delta = 0.5$

Equivalence Bound τ	Sample Size N	Number of Rejections (10000 Simulations) *Equivalent*	***Both***	**Different**	Neither
0.1	10	*0*	***0***	**1897**	8103
	20	*0*	***0***	**3383**	6617
	50	*0*	***0***	**6981**	3019
	100	*0*	***0***	**9428**	572
	200	*0*	***0***	**9985**	15
	500	*0*	***0***	**10000**	0
0.2	10	*0*	***0***	**1804**	8196
	20	*0*	***0***	**3437**	6563
	50	*0*	***0***	**6905**	3095
	100	*1*	***0***	**9429**	570
	200	*0*	***0***	**9987**	13
	500	*0*	***0***	**10000**	0
0.4	10	*7*	***0***	**1866**	8127
	20	*117*	***0***	**3425**	6458
	50	*370*	***0***	**6936**	2692
	100	*236*	***0***	**9378**	386
	200	*13*	***108***	**9879**	0
	500	*0*	***28***	**9972**	0

Conclusion

The data originally collected and analyzed with traditional significance tests by Benson (1989) showed a statistically significant difference between the means of male and female statistics students on six variables (GPA, number of prior math courses, math self-concept, self-efficacy, general test anxiety, and statistical test anxiety) and failed to find a significance for only one variable (achievement). We computed Cohen's *d* as an effect size. Not surprisingly, the smallest absolute effect size of 0.04 was found for the non-significant variable, while the absolute effect sizes of the six significant variables ranged from 0.24 to 0.66.

We then re-analyzed Benson's data using the TOST procedure for testing for statistical equivalence. This analysis showed that only two variables, number of prior math courses and statistical test anxiety, were "different" (i.e. significant and not equivalent). Not coincidentally, these were the two variables with the strongest absolute effect sizes of 0.60 and 0.66. The non-significant variable (achievement) was found to be statistically equivalent, and the absolute effect size was virtually zero. Four of the variables (GPA, math self-concept, self-efficacy, and general test anxiety) yielded conflicting results of "equivalent and different" since they rejected the null hypotheses of both the statistical and equivalence tests. It is likely that the difference in the means of these four variables, while statistically significant, is trivial. The absolute effect sizes of these four variables ranged from 0.24 to 0.51. This encompasses a range of effect sizes that is often classified as "small" to "medium" (Cohen, 1988), notwithstanding Lenth's (2001) warnings against using "canned" effect sizes.

We noticed that whenever the effect size δ is less than the equivalence bound τ, then the power of the equivalence test was approaching unity as *n* increased. This convergence was slow when δ was nearly equal to τ. Essentially, if the effect size parameter is less than the minimum difference that the researcher considers to be practically important (i.e. the minimum difference between means large enough to matter), we will reject the null of the TOST and conclude equivalence with power increasing to unity with larger sample sizes.

If $\delta > \tau$, the power of the significance test approaches unity and the power of the equivalence test approaches zero as the sample size increases. This is the situation where the effect size parameter exceeds the specified maximum for practical importance; we will reject the t-test and conclude statistical significance with power increasing to unity as the sample size increases.

When $\delta = \tau$, then the power of the equivalence test will approach twice the nominal alpha level. This occurs because the effect size parameter happens to coincide with the specified equivalence bound. Rejecting the TOST (i.e. concluding equivalence) is a type I error, made with probability 2α. The probability is twice the nominal α since an equivalence test at level α corresponds to a $100(1-2\alpha)\%$ equivalence interval.

When $0 < \delta < \tau$, then the power of both the significance and equivalence tests approaches unity (often slowly) as *n* increases. This is the situation where the null hypothesis of a significance test is false (i.e. the difference of means is not equal to zero), but the true difference is too small to be considered practically significant, where τ is the minimum difference between means that is considered important.

It appears to be somewhat common with real data to have situations where the tests of statistical significance and equivalence are simultaneously rejected for reasonable choices of significance level α and equivalence bound τ. Our re-analysis of the Benson (1989) data yielded 4 simultaneous rejections out of 7 variables.

The simulated power of simultaneous rejection showed that the probability of simultaneous rejection was low when the assumptions of the inferential tests (i.e. normality, equal variances, equal sample sizes between groups) were true except when both *n* and τ were large. It is possible that "simultaneous rejection" will be more likely with real data than (at least our) simulated data

because real data will surely violate the normality and homoscedasticity assumptions. We speculate that simultaneous rejection will be more common, and thus potentially more problematic for the researcher using equivalence testing in conjunction with standard hypothesis testing, when the data is non-normal and heteroscedastic.

Sawilowsky and Yoon (2002) demonstrated that large effect sizes could be found in situations where the results of a hypothesis test are 'not significant' (i.e. p>.05). Similarly, we found the magnitude of effect sizes obtained from the statistical re-analysis of typical educational research data to be troubling. Benson's data was of a decent size (groups of 94 and 123 subjects), but an effect size as large as 0.51 yielded both statistical significance (rejecting that the male mean was equal to the female mean) and equivalence (rejecting that the absolute difference of the male and female means were within a constant τ). We make the conjecture that the effect size conventions of Cohen (i.e. 0.2 is small, 0.5 is medium, 0.8 is large) might not be large enough. It is even possible that making any recommendation about the desired magnitude of an effect size independent of the sample sizes and variability of the populations might be futile (Lenth, 2001).

It would be desirable to extend the simulation study to consider several scenarios ignored here. In particular, more attention needs to be given to situations where one or more of the following conditions are true:

1. The populations are non-normal
2. The variances are not equal
3. The sample sizes of the groups are not equal.

It would also be desirable to analytically determine the power function for simultaneous rejection of the significance and equivalence tests, if possible. We will continue to strive for a greater understanding of the link between the effect size and the results of the significance and equivalence tests. It appears that sole reliance on any standard methodology, be it hypothesis testing, confidence intervals, effect sizes, or equivalence testing is ill advised.

References

Abelson, R. (1997). A retrospective on the signficance test ban of 1999 (if there were no significance tests, they would have to be invented). In L. Harlow, S. Mulaik, & J. Steiger, *What if there were no significance tests?* (p. 117-141). Mahwah, NJ: Lawrence Erlbaum Associates.

Anderson, D., Burnham, K., & Thompson, W. (2000). Null hypothesis testing: Problems, prevalance, and an alternative. *Journal of Wildlife Management, 64*(4), 912-923.

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

Bartko, J. (1991). Proving the null hypothesis. *American Psychologist, 46*(10), 801-803.

Benson, J. (1989). Structural components of statistical test anxiety in adults: An exploratory model. *Journal of Experimental Education, 57*(3), 247-261.

Berger, R., & Hsu, J. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science, 11*(4), 283-319.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association, 33*, 526-536.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association, 37*, 325-335.

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.

Boring, E. (1919). Mathematical vs. scientific importance. *Psychological Bulletin, 16*, 335-338.

Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*, 378-399.

Carver, R. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61*(4), 287-292.

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

Cohen, J. (1994). The Earth is round ($p < .05$). *American Psychologist, 49*(12), 997-1003.

Falk, R. (1998). In criticism of the null hypothesis statistical test. *American Psychologist, 53*, 798-799.

Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. *Journal of Educational Research, 94*(5), 275-282.

Hagan, R. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52(*1), 15-24.

Hagan, R. (1998). A further look at wrong reasons to abandon statistical testing. *American Psychologist, 53*(7), 801-803.

Harlow, L., Mulaik, S., & Steiger, J. (1997). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum Associates.

Harris, R. (1997). Reforming significane testing via three-valued logic. In L. Harlow, S. Mulaik, & J. Steiger (Eds.), *What if there were no significance tests?* (p.145-174). Mahwah, NJ: Lawrence Erlbaum Associates.

Hunter, J. (1997). Needed: A ban on the significance test. *Psychological Science, 8*(1), 3-7.

Knapp, T. R. (1998). Comments on the statistical significance testing articles. *Research in the Schools, 5*(2), 39-41.

Knapp, T. R. (2002). Some reflections on significance testing. *Journal of Modern Applied Statistical Methods, 1(*2), 240-242.

Krantz, D. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association, 94*, 1372-1381.

Lehmann, E. (1959). *Testing statistical hypotheses* (1st ed.). New York: Wiley.

Lenth, R. (2001). Some practical guidelines for effective sample size determination. *The American Statistician, 55*(3), 187-193.

McBride, G. (1999). Equivalence tests can enhance environmental science and management. *Australian and New Zealand Journal of Statistics, 41*(1), 19-29.

McLean, J., & Ernest, J. (1998). The role of statistical significane testing in educational research. *Research in the Schools, 5*(2), 15-22.

Meehl, P. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103-115.

Meehl, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806-834.

Meehl, P. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. Harlow, S. Mulaik, & J. Steiger (Eds.), *What if there were no significance tests?* (p. 393-426). Mahwah, NJ: Lawrence Erlbaum Associates.

Nix, T., & Barnette, J. (1998). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. *Research in the Schools, 5*(2), 3-14.

Patel, H., & Gupta, G. (1984). A problem of equivalence in clinical trials. *Biometrical Journal, 26*, 471-474.

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

Rozeboom, W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57*, 416-428.

Sawilowsky, S. S., & Yoon, J. (2002). The trouble with trivials (p>.05). *Journal of Modern Applied Statistical Methods, 1*(1), 143-144.

Schmidt, F., & Hunter, J. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. Harlow, S. Mulaik, & J. Steiger (Eds.), *What if there were no significance tests?* (p. 37-64). Mahwah, NJ: Lawrence Erlbaum Associates.

Schuirmann, D. (1981). On hypothesis testing to determine if the mean of the normal distribution is contained in a known interval. *Biometrics, 37*, 617.

Schuirmann, D. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmokinetics and Biopharmaceutics, 15*, 657-680.

Serlin, R. (1993). Confidence intervals and the scientific method: A case for Holm on the range. *Journal of Experimental Education, 61*, 350-360.

Serlin, R. (2002). Constructive criticism. *Journal of Modern Applied Statistical Methods, 1*(2), 202-227.

Thompson, B. (1998). Statistical significance and effect size reporting: Portrait of a possible future. *Research in the Schools, 5*(2), 33-38.

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education, 70*(1), 80-93.

Vacha-Haase, T. (2001). Statistical significance should not be considered one of life's guarantees: Effect sizes are needed. *Educational and Psychological Measurement, 61*, 219-224.

Westlake, W. (1976). Symmetric confidence intervals for bioequivalence trials. *Biometrics, 32*, 741-744.

Westlake, W. (1979). Statistical aspects of comparative bioequivalence trials. *Biometrics, 35*, 273-280.