Qual Quant (2014) 48:755–765
DOI 10.1007/s11135-012-9800-6

Meta-analytic methods to test relative efficacy

Bruce E. Wampold · Ronald C. Serlin

Published online: 7 November 2012

Abstract In the context of multiple treatments for a particular problem or disorder, it is important theoretically and clinically to investigate whether any one treatment is more effective than another. Typically researchers report the results of the comparison of two treatments, and the meta-analytic problem is to synthesize the various comparisons of two treatments to test the omnibus null hypothesis that the true differences of all particular pairs of treatments are zero versus the alternative that there is at least one true nonzero difference. Two tests, one proposed by Wampold et al. (Psychol. Bull. 122:203–215, 1997) based on the homogeneity of effects, and one proposed here based on the distribution of the absolute value of the effects, were investigated. Based on a Monte Carlo simulation, both tests adequately maintained nominal error rates, and both demonstrated adequate power, although the Wampold test was slightly more powerful for non-uniform alternatives. The error rates and power were essentially unchanged in the presence of random effects. The tests were illustrated with a reanalysis of two published meta-analyses (psychotherapy and antidepressants). It is concluded that both tests are viable for testing the omnibus null hypothesis of no treatment differences.

Keywords Meta-analysis · Relative efficacy · Treatment effects

1 Introduction

Meta-analytic methods have become the standard procedures for synthesizing research findings in medicine, education, psychology, and other fields (Cooper et al. 2009; Hunt 1997; Mann 1994). As originally proposed, meta-analysis primarily addressed research questions about the effectiveness of some treatment by aggregating effect sizes that indexed the degree

B. E. Wampold (✉) · R. C. Serlin
University of Wisconsin, Madison, WI, USA
e-mail: wampold@education.wisc.edu

B. E. Wampold
Modum Bad Psychiatric Center, Vikersund, Norway

to which the treatment was superior to some type of control group (Cooper and Hedges 1994; Glass 1976; Hedges and Olkin 1985). That is, an effect size that gauged the difference between the treatment group and the control group is calculated for each study and aggregated over the studies that compared the treatment to a control group. Typically, the effects are scaled so that a positive effect indicates that the treatment is superior to the control group. The characteristics of the distributions of the effects and their aggregate are well known (Cooper and Hedges 1994; Cooper et al. 2009; Hedges and Olkin 1985), providing a variety of tests, depending on the null hypothesis being tested.

Not infrequently, researchers in a particular area wish to identify the most effective treatment among a set of treatments. For example, there are numerous psychological treatments for depression, and it is important to know whether one of these treatments is more effective than any other (See, e.g., Cuijpers et al. 2008; Gloaguen et al. 1998; Robinson et al. 1990; Wampold et al. 2002). When there exists only two treatments for a particular disorder, standard meta-analytic methods are easily adapted to compare the relative efficacy of two treatments, say Treatment A and Treatment B, in which case an effect would be calculated and scaled in such a way that a positive effect would indicate that Treatment A was superior to Treatment B. Rejection of the null hypothesis of no differences between the two treatments leads to a conclusion that one of the treatments is superior to the other. However, it is often the case that the number of treatments for a disorder is relatively large and the number of studies comparing any two particular treatments is relatively small. That is, researchers choose to compare two treatments sampled from a larger set of treatments because they are interested in these two treatments; other researchers choose other pairs of treatments. In this instance, the null hypothesis is that all treatments are equally effective—the true difference of any two treatments is zero. This is a critically important question, clinically as well as theoretically. In psychotherapy research, there is much debate about the answer to this question—are treatments, in general or for particular disorders, equally effective, or are some treatments more effective than others (Benish et al. 2008; Crits-Christoph 1997; Howard et al. 1997; Imel et al. 2008; Miller et al. 2008; Siev et al. 2009; Wampold 2001; Wampold et al. 2009, 1997)? Similarly, there is a debate about the relative efficacy of new-generation antidepressants (Cipriani et al. 2009).

Two meta-analytic strategies have been used to test the relative efficacy of many treatments. The first strategy is to aggregate effects for studies comparing two particular treatments (i.e., Tx A and Tx B), which creates m(m − 1)/2 tests, where m is the number of treatments studied. This strategy is problematic, because most pairwise comparisons involve few studies (i.e., there are few comparisons of the particular Treatments A and B) and because this strategy uses multiple tests to examine a single hypothesis. A variant of this strategy, aimed at reducing the number of statistical tests, involves creating classes of treatments and examining pairwise comparisons of classes of treatments (e.g., Gloaguen et al. 1998; Shapiro and Shapiro 1982a,b). This latter strategy addresses questions related to relative efficacy of the classes of treatments but does not address directly the hypothesis that all treatments are equally effective. Moreover, classification of treatments into classes typically is controversial and unreliable (Wampold 2001; Wampold et al. 2010) and, as well, obviates examination of treatments within classes, which are often the focus of the primary studies (Wampold et al. 1997).

A second strategy, which avoids multiple statistical tests and classification, tests the null hypothesis directly. This strategy involves calculating the effect size for each comparison and then examining the distribution of effects to test the omnibus null hypothesis of no differences among pairs of treatments. Wampold et al. (1997) used a variant of this strategy to test the null hypothesis of no differences among psychotherapies (See also Benish et al. 2008; Imel et al. 2008; Miller et al. 2008). Unfortunately, the statistical properties of the test proposed by Wampold et al. (1997) have not been thoroughly investigated and, consequently, the validity of the conclusions reached using this test is questionable.

The purpose of this article is to investigate the properties of the test proposed by Wampold et al. (1997), as well as those of an alternative that will be proposed here. Both tests ultimately involve the same null hypothesis but are based on different statistical models.

2 General problem

Suppose that there is a set, potentially infinite in size, of treatments for a particular disorder or problem. Researchers typically choose to compare two particular treatments. The effect size for each study is calculated in the usual fashion:

$$g_i = (\bar{Y}_{A_i} - \bar{Y}_{B_i})/S_i$$

where, for the ith study, $\bar{Y}_{A_i}$ is the sample mean for treatment A on the variable of interest, $\bar{Y}_{B_i}$ is the sample mean for treatment B, and S_i is the sample pooled standard deviation (Hedges and Olkin 1985). An arbitrariness issue becomes apparent, because designation of Treatment $\mathrm{A_i}$ or Treatment $\mathrm{B_i}$ is ambiguous—which is $\mathrm{A_i}$ and which is $\mathrm{B_i}$? The designation of one of the two treatments as Treatment $\mathrm{A_i}$ will determine the sign of $\mathrm{g_i}$. As it turns out, the choice is irrelevant, and the sign of g_i can be determined randomly or even arbitrarily without affecting the statistical tests proposed here.

We first consider a fixed effects model, in which $g_i = \delta_i + e_i$, where $\delta_i = (\mu_{A_i} - \mu_{B_i})/\sigma_i$ is the true standardized mean difference and e_i are the errors of estimation (Raudenbush 2009). The errors of estimation are assumed to be independent and normally distributed with mean zero and variance v_i [i.e., $e_i \sim N(0, v_i)$]. The sample standardized mean difference is biased, with expected value $E(g_i) = \delta_i / C_{m_i}$, where$C_{m_i} = \frac{\Gamma(\frac{m_i}{2})}{\sqrt{\frac{m_i}{2}}\Gamma(\frac{m_i-1}{2})}$, $\Gamma(\cdot)$ is the gamma function, $\mathrm{m_i = N_i} - 2$ is the number of degrees of freedom associated with the pooled within-group variance, and $\mathrm{N_i}$ is sum of sample sizes for Treatment $\mathrm{A_i}$ (n_{A_i}) and Treatment $\mathrm{B_i}$ (n_{B_i}). The bias in $\mathrm{g_i}$ is corrected in the usual way (Hedges 1981; Hedges and Olkin 1985), creating an unbiased estimator of δ_i, namely $\mathrm{d_i} = C_{m_i}\ \mathrm{g_i}$. The variance of $\mathrm{d_i}$is equal to $\sigma^2(d_i) = \frac{C^2_{m_i} m_i}{(m_i-2)\tilde{n}_i}\left[1 + \tilde{n}_i \delta_i^2\right] - \delta_i^2$, where $\tilde{n}_i = \frac{n_{A_i} n_{B_i}}{N_i}$.

Given k independent studies, the null hypothesis is $\mathrm{H_0} : [\delta] = [0]$, where $[\delta]$ is a $k \times 1$ vector of the k true standardized mean differences for the particular pairs of treatments, δ_i, [0] is a vector of zeros, and the alternative hypothesis states that there exists at least one nonzero effect size. Under the Wampold model, it is assumed that the k sample effect sizes have been sampled at random and independently from k populations of studies in which the particular pairs of treatments are compared and whose parameters are μ_{A_i}, μ_{B_i}, and σ_i. The aim is to utilize the $\mathrm{d_i}$s and their distributions to test the null hypothesis.

3 Statistical tests of null hypothesis

Two statistical tests of the null hypothesis are described here: Wampold et al.'s (1997) test and a new one based on a modification of Geary's (1935) test of normality. We then investigate the statistical properties of the two tests.

3.1 Wampold et al's homogeneity test

Consider the vector $[d]$ containing the $k \quad d_{is}$. The distribution of $[d]$ is well known, having expected value $[\delta]$, a variance–covariance matrix Σ_d whose diagonal elements are $\sigma^2(d_i)$ and whose off-diagonal elements are zero, as the $\mathrm{d_i}$ are independent, and that is approximately multivariate normal as the sample sizes increase. If the null hypothesis is true (i.e., $\delta_\mathrm{i} = 0$), then the effects are said to be homogenously distributed, the expected value of $[d]$ is [0], the variance of $\mathrm{d_i}$ is equal to $\sigma_0^2(d_i) = \frac{C_{m_i}^2 m_i}{(m_i-2)\tilde{n}_i}$, and the diagonal elements of Σ_d are equal to $\Sigma_{d_0} = \sigma_0^2(d_i)$. If the effects are not uniformly zero (i.e., at least one $\delta_\mathrm{i} \neq 0$), then the sample effects will appear to be heterogeneously distributed about zero. Hedges and Olkin (1985) describe a test of the null hypothesis of homogeneity based on Cramér's (1946) modified minimum χ^2 test. In the present case, and assuming that the null hypothesis is true and that the normal approximation to the distribution of d_i is adequate, Pearson (1900) showed that the form

$$W = [d]^{'} \sum_{d_0}^{-1} [d] = \sum_{i=1}^{k} \frac{d_i^2}{\sigma_0^2(d_i)}$$

is distributed as a central χ^2 variate with k degrees for freedom. If W is sufficiently large, the null hypothesis is rejected.[1]

3.2 Geary's test of normality

Geary (1935) derived the moments of the sample mean of absolute values, $\overline{|y|} = \frac{1}{k}\sum_{i=1}^{k}|y_i|$ of observations drawn from a standard normal distribution. He proposed that the first and second moments of $\overline{|y|}$ could be used as a test of normality and showed that the mean and variance of $\overline{|y|}$ are equal to $\sqrt{2/\pi}$ and $(1 - 2/\pi)/k$, respectively.[2] In addition, he showed that the convergence of $\overline{|y|}$ to normality is fairly rapid, with the skewness and kurtosis of $\overline{|y|}$ equal to 0.1 and 0.009, respectively, at $k = 20$. Therefore, Geary suggested that $Z_G = (\overline{|y|} - \sqrt{2/\pi})/\sqrt{(1-(2/\pi))/k}$ could be used as a normally distributed test statistic to examine the hypothesis that the y_i were sampled from a standard normal population. In the present meta-analytic context, under the null hypothesis of no treatment differences, effects can be standardized given known variances, so that $|y_i| = |d_i|/\sigma_0(d_i)$ can serve as the absolute values of k observations y_i drawn from a standard normally distributed population, and Z_G can be used to test the null hypothesis regarding treatment effectiveness.

4 Monte Carlo study of Wampold et al. and absolute value tests: fixed effects

In both the Wampold et. al. test and the extension of Geary's method based on absolute values of the effects, the tests are one-tailed, in that treatment effectiveness would be revealed by the d_i being further from zero than expected or by the $|\mathrm{d_i}|$ being larger than expected. The issue is whether the Wampold et al. test and the extension of the Geary test, to be referred to as the absolute value test, adequately protect Type I error rates and are reasonably powered for use

[1] Wampold et al. (1997) performed the homogeneity test with k – 1 degrees of freedom rather than k, ignoring the fact that the mean effect size was stipulated rather than estimated.

[2] For the sake of historical accuracy, Geary denoted the mean of the absolute values as $\overline{|y|}$ whereas we denoted it by $\overline{|y|}$.

in the typical meta-analytic context, and if so, to determine which of the tests is preferred. We conducted a Monte Carlo study to make these determinations. When specifying conditions for the Monte Carlo study, it would be informative to note that the Wampold test statistic equals $W = \sum_{i=1}^{k} |y_i|^2 = k\overline{|y|}^2 + \sum_{i=1}^{k} (|y_i| - \overline{|y|})^2$. The first term in the sum on the right is proportional to the square of the difference from zero of the mean of the absolute standardized effect sizes, whereas the second term in the sum reflects the heterogeneity among the absolute standardized effect sizes. Therefore, if the effect sizes are about equally nonzero, so that the rightmost term is small, we would expect that the Wampold and extended Geary tests might perform about equally well, whereas for a given nonzero average, the Wampold test may perform better when there is heterogeneity among the effect sizes. We will examine conditions in which the effect sizes are equally nonzero and others in which there is some heterogeneity among the effect sizes.

The two tests were examined in simulations whose conditions were intended to cover a reasonable range of likely sample sizes, numbers of studies included in meta-analyses, and effect sizes. Specifically, the sizes of the treatment groups were chosen to be $n_{\mathrm{A}} = n_{\mathrm{B}} = 10$, 20, 40, and 80, and the numbers of studies were set equal to $k = 10$, 20, 50, and 100.

To examine Type I error rates, all standardized mean differences δi were set equal to zero. The examination of power involved a number of non-null scenarios. First, the δ_{i} were set uniformly to 0.1, 0.2, 0.3, 0.4 or 0.5. Second, various non-uniform differences were modeled by considering three scenarios. In Scenario A, 25 % of the δ_{i} were set to 0.0, 25 % of the δ_{i} were set to 0.1, 25 % of the δ_{i} were set to 0.2, and 25 % of the δ_{i} were set to 0.3. In Scenario B, 25 % of the δi were set to 0.0 and 75 % of the δ_{i} were set to 0.2. Finally, in Scenario C, 50 % of the δ_{i} were set to 0.0 and 50 % of the δ_{i} were set to 0.3. In all three scenarios the mean δ was 0.15.

In each replication, uniform random numbers were generated using RAN2 (Press et al. 1992) and transformed to random standard normal deviates using the method of Box and Muller (1958). To provide a set of treatment means sampled at random from a population whose variance was unity, k of the random deviates were divided by $\sqrt{n_A}$ and another k were divided by $\sqrt{n_B}$. A particular value of δ was then added to each simulated Treatment mean. Similarly, uniform random numbers were used as percentiles to generate χ^2-distributed variates using the inverse cdf method. To simulate the treatment variances, k of these values

Table 1 Simulated meta-analyses error rates for W and Z_G under null hypothesis and nominal $\alpha = 0.05$

k	$n_A = n_B$	W	Z_G
10	10	0.0669	0.0610
	20	0.0564	0.0551
	40	0.0520	0.0613
	80	0.0520	0.0588
20	10	0.0652	0.0525
	20	0.0574	0.0539
	40	0.0574	0.0601
	80	0.0507	0.0539
50	10	0.0721	0.0484
	20	0.0547	0.0491
	40	0.0546	0.0512
	80	0.0488	0.0516
100	10	0.0701	0.0399
	20	0.0552	0.0455
	40	0.0569	0.0485
	80	0.0501	0.0496

were divided by $(n_\mathrm{A} - 1)$ and k were divided by $(n_\mathrm{B} - 1)$. These variances were pooled an "experiment" at a time, and for each the mean difference was divided by the pooled within-group standard deviation to yield k effect sizes, the absolute values of which were then taken. Finally, the k effect sizes were used to yield W and Z_G, and this process was repeated 10,000 times for each condition.

The Type I error rates for the 80 conditions are shown in Table 1. Examination of the error rates suggests that both tests perform adequately, maintaining Type I error rates within the range of 0.0455 and 0.0613 when the sample sizes of the treatment groups are greater than 20. When the samples sizes are equal to 10, Wappears to be more liberal than Z_G, with error rates ranging up to 0.0721.

The results of the power analyses are shown in Table 2 and do not include the conditions for which $\delta = 0.4$ or 0.5 because the power is so high for these cases as to be uninformative. Given the small effects modeled in Table 2, these analyses demonstrate that the power of both tests is adequate in most instances in which meta-analyses are applied. In several instances, particularly the non-uniform alternative hypothesis, W is slightly more powerful than Z_G, as was suspected.

5 Extension to random effects model

In the meta-analytic context, the random effects model considers the random sampling of studies from a population of studies and the random sample of subjects in each study from a population of subjects. Accordingly, $g_i = \delta_i + u_i + e_i$, where the error u_i is the random effect of study i and represents the deviation of study $i$from the expected value (Raudenbush 2009). In this model, $u_i + e_i \sim N(0, v_i^*)$ where $v_i^* = \sigma_{ui}^2 + v_i$ and is the total variance in the observed effects g_i. We extended the random effects meta-analysis by considering different values of the intraclass correlation coefficient ρ, where ρ is defined as $\sigma_{\mu i}^2/(\sigma_{ui}^2 + v_i)$.

Simulations were run for models in which ρ was equal to 0.05, 0.10, and 0.20 and δ was equal to 0.0, 0.10, and 0.20, and replicated 10,000 times. The proportion of rejections for these various values are presented in Table 3. The simulation clearly shows that the nominal error rates are maintained in the random effects models and that the power of the two tests in the random model was similar to the power of the fixed effects model across various combination of ρ and δ

6 Examples

We illustrate the two procedures by applying to them to existing data sets, one that examined the relative efficacy of new-generation antidepressants (Cipriani et al. 2009) and one that examined the relative efficacy of various psychological treatments (Wampold et al. 1997).

6.1 Comparative efficacy of new-generation antidepressants.

Cipriani et al. (2009) were interested in identifying which of 12 new-generation antidepressants were more effective than others. There were 117 randomized clinical trials that compared two or more of these antidepressants. Clearly, not all antidepressants were compared to each of the other. A Bayesian analysis using Markov chain Monte Carlo methods were used to estimate direct and indirect paths from each of the nodes in a network analysis.

Table 2 Proportion of rejections in simulated meta-analyses (i.e., power) for various non-null scenarios (α = 0.05) for W and Z_G

		$\delta = 0.1$		$\delta = 0.2$		$\delta = 0.3$		Scenario A		Scenario B		Scenario C	
k	$n_A = n_B$	W	Z_G	W	Z_G	W	Z_G	W	Z_G	W	Z_G	W	Z_G
10	10	0.0775	0.0691	0.1293	0.1178	0.2453	0.2444	0.1264	0.1154	0.1143	0.1048	0.1410	0.1309
	20	0.0914	0.0930	0.2215	0.2284	0.4829	0.5042	0.2017	0.1983	0.1733	0.1752	0.2401	0.2393
	40	0.1207	0.1303	0.4374	0.4584	0.8491	0.8682	0.3842	0.3782	0.3222	0.3390	0.4767	0.4701
	80	0.2156	0.2287	0.7954	0.8189	0.9949	0.9968	0.6903	0.6846	0.6332	0.6422	0.8011	0.7835
20	10	0.0874	0.0724	0.1747	0.1525	0.3568	0.3372	0.1554	0.1295	0.1421	0.1207	0.1827	0.1601
	20	0.1028	0.0943	0.3129	0.3118	0.7116	0.7212	0.2761	0.2592	0.2327	0.2272	0.3553	0.3379
	40	0.1651	0.1656	0.6396	0.6510	0.9763	0.9817	0.5630	0.5441	0.4825	0.4873	0.7021	0.6865
	80	0.3119	0.3212	0.9597	0.9656	1.0000	1.0000	0.8939	0.8824	0.8554	0.8566	0.9603	0.9497
50	10	0.0981	0.0719	0.2617	0.2135	0.6011	0.5598	0.2365	0.1841	0.1988	0.1562	0.2917	0.2310
	20	0.1296	0.1160	0.5428	0.5223	0.9582	0.9578	0.4656	0.4244	0.4062	0.3758	0.6094	0.5718
	40	0.2469	0.2332	0.9210	0.9218	1.0000	1.0000	0.8601	0.8402	0.7830	0.7734	0.9497	0.9334
	80	0.5472	0.5359	1.0000	1.0000	1.0000	1.0000	0.9968	0.9957	0.9926	0.9924	0.9997	0.9998
100	10	0.1232	0.0789	0.3851	0.2987	0.8280	0.7850	0.3312	0.2525	0.2872	0.2103	0.4381	0.3440
	20	0.1826	0.1504	0.7763	0.7446	0.9988	0.9983	0.6833	0.6335	0.6083	0.5588	0.8377	0.7985
	40	0.3907	0.3623	0.9958	0.9958	1.0000	1.0000	0.9830	0.9747	0.9591	0.9539	1.0000	1.0000
	80	0.7949	0.7777	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Scenario A: 25 % of $\delta = 0.0$, 25 % of $\delta = 0.1$, 25 % of $\delta = 0.2$, 25 % of $\delta = 0.3$; Scenario B: 25 % of $\delta = 0.0$, 75 % of $\delta = 0.2$; Scenario C: 50 % of $\delta = 0.0$, 50 % of $\delta = 0.3$

Table 3 Proportion of rejections in simulated meta-analyses for W and Z_G for random-effects model.

k	$n_A = n_B$	$\delta = 0$		$\delta = 0.10$		$\delta = 0.20$		$\delta = 0.30$	
		W	Z_G	W	Z_G	W	Z_G	W	Z_G
		$\rho = 0.05$							
10	10	0.0607	0.0578	0.0799	0.0720	0.1286	0.1243	0.2498	0.2475
	20	0.0561	0.0597	0.0916	0.0954	0.2202	0.2221	0.4799	0.5031
	40	0.0520	0.0566	0.1222	0.1309	0.4311	0.4472	0.8381	0.8664
	80	0.0558	0.0636	0.2193	0.2283	0.7898	0.8199	0.9959	0.9969
20	10	0.0652	0.0509	0.0851	0.0720	0.1797	0.1588	0.3586	0.3385
	20	0.0575	0.0534	0.1050	0.0984	0.3206	0.3150	0.7070	0.7151
	40	0.0539	0.0580	0.1602	0.1563	0.6367	0.6563	0.9794	0.9834
	80	0.0505	0.0571	0.3177	0.3216	0.9573	0.9634	1.0000	1.0000
50	10	0.0666	0.0456	0.0948	0.0645	0.2627	0.2121	0.5979	0.5500
	20	0.0559	0.0484	0.1397	0.1266	0.5455	0.5234	0.9548	0.9543
	40	0.0541	0.0527	0.2493	0.2396	0.9221	0.9185	1.0000	1.0000
	80	0.0493	0.0513	0.5469	0.5410	0.9999	0.9998	1.0000	1.0000
100	10	0.0646	0.0390	0.1168	0.0727	0.3863	0.3045	0.8373	0.8373
	20	0.0587	0.0457	0.1834	0.1546	0.7831	0.7514	0.9984	0.9982
	40	0.0530	0.0511	0.3862	0.3603	0.9943	0.9952	1.0000	1.0000
	80	0.0540	0.0490	0.7952	0.7829	1.0000	1.0000	1.0000	1.0000
		$\rho = 0.10$							
10	10	0.0646	0.0581	0.0797	0.0753	0.1305	0.1242	0.2460	0.2381
	20	0.0544	0.0561	0.0914	0.0954	0.2193	0.2300	0.4882	0.5056
	40	0.0507	0.0568	0.1220	0.1286	0.4301	0.4536	0.8402	0.8637
	80	0.0521	0.0598	0.2160	0.2295	0.7843	0.8156	0.9962	0.9983
20	10	0.0679	0.0501	0.0901	0.0714	0.1727	0.1482	0.3562	0.3401
	20	0.0524	0.0535	0.1090	0.1018	0.3189	0.3134	0.6978	0.7122
	40	0.0493	0.0523	0.1498	0.1517	0.6429	0.6585	0.9767	0.9829
	80	0.0534	0.0597	0.3234	0.3249	0.9586	0.9662	1.0000	1.0000
50	10	0.0672	0.0478	0.1018	0.0762	0.2551	0.2062	0.6022	0.5584
	20	0.0551	0.0480	0.1396	0.1247	0.5327	0.5121	0.9549	0.9546
	40	0.0542	0.0498	0.2460	0.2361	0.9249	0.9254	0.9998	0.9999
	80	0.0521	0.0522	0.5435	0.5356	0.9999	1.0000	1.0000	1.0000
100	10	0.0705	0.0408	0.1212	0.0795	0.3812	0.3004	0.8280	0.7838
	20	0.0593	0.0483	0.1770	0.1497	0.7832	0.7531	0.9991	0.9987
	40	0.0513	0.0487	0.3844	0.3564	0.9965	0.9957	1.0000	1.0000
	80	0.0507	0.0502	0.7942	0.7760	1.0000	1.0000	1.0000	1.0000
		$\rho = 0.20$							
10	10	0.0637	0.0589	0.0810	0.0755	0.1243	0.1210	0.2440	0.2403
	20	0.0599	0.0597	0.0864	0.0865	0.2147	0.2227	0.4871	0.5033
	40	0.0526	0.0565	0.1238	0.1314	0.4376	0.4575	0.8367	0.8643
	80	0.0478	0.0560	0.2116	0.2239	0.7946	0.8227	0.9962	0.9970
20	10	0.0634	0.0500	0.0835	0.0686	0.1727	0.1527	0.3576	0.3389
	20	0.0551	0.0569	0.1026	0.0974	0.3107	0.3072	0.7047	0.7181
	40	0.0497	0.0533	0.1599	0.1606	0.6439	0.6573	0.9772	0.9827
	80	0.0504	0.0537	0.3099	0.3151	0.9565	0.9637	0.9999	0.9999
50	10	0.0658	0.0444	0.0970	0.0705	0.2551	0.2123	0.6054	0.5650
	20	0.0568	0.0498	0.1376	0.1177	0.5467	0.5249	0.9522	0.9537
	40	0.0500	0.0525	0.2516	0.2377	0.9305	0.9298	1.0000	1.0000
	80	0.0545	0.0547	0.5416	0.5339	1.0000	1.0000	1.0000	1.0000
100	10	0.0678	0.0391	0.1170	0.0747	0.3827	0.2941	0.8359	0.7934
	20	0.0560	0.0474	0.1875	0.1611	0.7826	0.7537	0.9981	0.9981
	40	0.0558	0.0535	0.3797	0.3592	0.9961	0.9948	1.0000	1.0000
	80	0.0536	0.0522	0.7913	0.7844	1.0000	1.0000	1.0000	1.0000

Outcomes were expressed as odds ratios from the proportion of subjects who responded to treatment. Based on their analysis, the authors concluded that some of the antidepressants were more effective than others.

We sought to test the omnibus hypothesis that there were differences among treatments. To accomplish that goal, we converted the odds ratios to an equivalent d (Borenstein 2009) and applied the Wampold homogeneity test and Geary's test of normality. Both tests yielded particularly small test statistics ($Q = 9.43$, $df = 112$, $p = .99$ for the former and $z = -11.46$, $p = 0.99$ for the latter), indicating that there is no evidence that the 12 antidepressants differed in terms of their effectiveness and therefore, without planned comparisons, it seems to make little sense to examine the data set for particular differences among particular treatments.

6.2 Reanalysis and re-interpretation of Wampold et al. (1997)

In 1997, Wampold et al. (1997) conducted a meta-analysis to test whether legitimate psychotherapeutic treatments were equally effective or not, thus testing the hypothesis that the δ_i found in studies comparing two such treatments are equal to zero. Wampold et al. searched the literature and found 277 such studies. To test that null hypothesis, Wampold et al. used the W statistic discussed above and found that $W = 241.18$, which when compared to the distribution of a variate that is χ^2 with 277 degrees of freedom, is found not to be statistically significant ($p = 0.94$), and the null hypothesis was retained. It was concluded that there is no evidence in these 277 studies to suggest that the any one treatment is more effective than another.

Wampold et al. (1997) then went on calculate the weighted aggregate of the absolute value of the effects, which was found to be equal to 0.19, and concluded that "this estimate of the true effect is an overestimate [of the true difference between treatments] and provides an upper bound only" (Wampold et al. 1997, p. 209). However, the absolute value of the effects was interpreted by many as *an estimate of the true differences among treatments*. For example, Howard et al. (1997) stated, "The .19 mean absolute value effect size reported by Wampold et al. would thus seem to be the value we want, so there is, on the average, a significant difference (according to their data) in outcome in trials of various pairs of psychotherapies" (Wampold et al. 1997, p. 221). An examination of Z_G will clarify the situation. In these data, $\overline{|y|} = \frac{1}{k}\sum_{i=1}^{k} |d_i| / \sigma(d_i)$, (i.e., the average of the standardized absolute values of the effects) was equal to 0.7109, which is less than the expected value $\sqrt{2/\pi} = 0.7979$. In this case $Z_G = -2.40$, which is in the opposite direction of what is needed to reject the null hypothesis that all true effect sizes are equal to zero ($p = 0.99$). Given that there was adequate power in this meta-analysis (See Table 2), it is now clear that Wampold et al.'s statement about the upper bound of the differences among effects was misleading, as the absolute values of the effects provides no evidence whatsoever that the true differences among treatments were anything other than zero.

7 Conclusions

Determining the relative efficacy of multiple treatments is critical theoretically and clinically in any service provision context, such as education, applied psychology, and medicine. When there are multiple treatments, the first question is whether any one treatment is more effective than any other. In this presentation, two tests, based on different statistical models, of the omnibus null hypothesis that there are no treatment differences were investigated. The first one, proposed by Wampold et al. (1997), is based on the homogeneity test discussed first discussed by Pearson (1900) and adapted to the meta-analytic context by Hedges and Olkin (1985). Although this test has been used to test the relative efficacy of treatments, the statistical properties of the test had not been investigated. The second test, proposed here, was adapted from Geary's test of normality using the distribution of the absolute value of the effects.

Although the statistical models were different, both tests adequately maintained nominal alpha levels in a variety of conditions and in a fixed and random effects contexts. Power of the tests depends on the sample sizes of the primary studies, as well as the number of studies in the meta-analysis. Both tests were reasonably powered for relatively moderate effects in most instances in typical situations in which meta-analysis might be applied. Under non-uniform alternative hypotheses, the Wampold test is slightly more powerful than Geary's.

The present analysis yields two viable meta-analytic tests for the null hypothesis that the differences among treatments are zero. Failure to reject the null hypothesis, in instances where power is adequate, suggests that there is insufficient evidence to conclude that any one treatment is superior to another, a particularly informative conclusion.

References

Benish, S., Imel, Z.E., Wampold, B.E.: The relative efficacy of bona fide psychotherapies of post-traumatic stress disorder: a meta-analysis of direct comparisons. Clin. Psychol. Rev. **28**, 746–758 (2008)

Borenstein, M.: Effect sizes for continuous data. In: Cooper, H., Hedges, L.V. & Valentine J.C. (eds.) The handbook of research synthesis and meta-analysis, 2nd ed., pp. 221–235. Russel Sage Foundation, New York (2009)

Box, G.E.P., Muller, M.E.: A note on the generation of random normal deviates. Ann. Math. Stat. **28**, 610–611 (1958)

Cipriani, A., Furukawa, T.A., Salanti, G., Geddes, J.R., Higgins, J.P.T., Churchill, R. et al.: Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. Lancet **373**, 746–758 (2009)

Cooper, H., Hedges, L.V. (eds.): The handbook of research synthesis. Russell Sage Foundation, New York (1994)

Cooper, H., Hedges, L.V., Valentine, J.C.: The handbook of research synthesis and meta-analysis (2nd ed.). Russel Sage Foundation, New York (2009)

Cramér, H.: Mathematical methods of statistics. Princeton University Press, Princeton (1946)

Crits-Christoph, P.: Limitations of the dodo bird verdict and the role of clinical trials in psychotherapy research: comment on Wampold et al. Psychol. Bull. 122, 216–220 (1997)

Cuijpers, P., van Straten, A., Andersson, G., van Oppen, P.: Psychotherapy for depression in adults: a meta-analysis of comparative outcome studies. J. Consult. Clin. Psychol. **76**, 909–922 (2008)

Geary, R.C.: The ratio of the mean deviation to the standard deviation as a test of normality. Biometrika **27**, 310–332 (1935)

Glass, G.V.: Primary, secondary, and meta-analysis of research. Educ. Res. **5**, 3–8 (1976)

Gloaguen, V., Cottraux, J., Cucherat, M., Blackburn, I.: A meta-analysis of the effects of cognitive therapy in depressed patients. J. Affect. Disord. **49**, 59–72 (1998)

Hedges, L.V.: Distribution theory for Glass's estimator of effect size and related estimators. J. Educ. Stat. **6**(2), 107–128 (1981). doi:10.2307/1164588

Hedges, L.V., Olkin I.: Statistical methods for meta-analysis. Academic Press, San Diego (1985)

Howard, K.I., Krause, M.S., Saunders, S.M., Kopta, S.M.: Trials and tribulations in the meta-analysis of treatment differences: comment on Wampold et al. (1997). Psychol. Bull. **122**, 221–225 (1997)

Hunt, M.: How science takes stock: the story of meta-analysis. Russell Sage Foundation, New York (1997)

Imel, Z.E., Wampold, B.E., Miller, S.D., Fleming, R.R.: Distinctions without a difference: direct comparisons of sychotherapies for alcohol use disorders. J. Addict. Behav. 533–543 (2008)

Mann, C.C.: Can meta-analysis make policy?. Science **266**, 960–962 (1994)

Miller, S.D., Wampold, B.E., Varhely, K.: Direct comparisons of treatment modalities for youth disorders: a meta-analysis. Psychother. Res. **18**, 5–14 (2008)

Pearson, K.: On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to hove arisen from random sampling. Philos. Mag. **50**, 157–175 (1900)

Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T.: Numerical recipes in FORTRAN: the art of scientific computing, vol. 1. Cambridge Univeresity Press, Cambridge (1992)

Raudenbush, S.W.: Analyzing effect sizes: random-effects models. In: Cooper, H., Hedges, L.V. & Valentine J.C. (eds.) The handbook of research synthesis and meta-analysis, 2nd ed., pp. 295–316. Russel Sage Foundation, New York (2009)

Robinson, L.A., Berman, R.A., Neimeyer, J.S.: Psychotherapy for the treatment of depression: a comprehensive review of controlled outcome research. Psychol. Bull. **108**, 30–49 (1990)

Shapiro, D.A., Shapiro, D.: Meta-analysis of comparative therapy outcome research: a critical appraisal. Behav. Psychother. **10**, 4–25 (1982)

Shapiro, D.A., Shapiro, D.: Meta-analysis of comparative therapy outcome studies: a replication and refinement. Psychol. Bull. **92**, 581–604 (1982)

Siev, J., Huppert, J., Chambless, D.L.: The dodo bird, treatment technique, and disseminating empirically supported treatments. Behav. Ther. **32**, 69–75 (2009)

Wampold, B.E.: The great psychotherapy debate: model, methods, and findings. Lawrence Erlbaum Associates, Mahwah (2001)

Wampold, B.E., Imel, Z.E., Laska, K.M., Benish, S., Miller, S.D., Flûckiger, C. et al.: Determining what works in the treatment of PTSD. Clin. Psychol. Rev. **30**, 923–933 (2010)

Wampold, B.E., Imel, Z.E., Miller, S.D.: Barriers to the dissemination of empirically supported treatments: matching messages to the evidence. Behav. Ther, **32**, 144–155 (2009)

Wampold, B.E., Minami, T., Baskin, T.W., Tierney, S.C.: A meta-(re)analysis of the effects of cognitive therapy versus "other therapies" for depression. J. Affect. Disord. **68**, 159–165 (2002)

Wampold, B.E., Mondin, G.W., Moody, M., Stich, F., Benson, K., Ahn, H.: A meta-analysis of outcome studies comparing bona fide psychotherapies: empirically, "All must have prizes". Psychol. Bull. **122**, 203–215 (1997)