Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement
2013, Vol. 45, No. 4, 320–328

0008-400X/13/$12.00 DOI: 10.1037/a0033357

Evaluating the Equivalence of, or Difference Between, Psychological Treatments: An Exploration of Recent Intervention Studies

Teresa A. Allan and Robert A. Cribbie
York University

In behavioral science research there is often the need to determine if an outcome variable differs, or is equivalent, across groups. Significance tests are the most prevalently applied data analysis method for this type of question. The purpose of this study was to examine how statistical tests for equivalence and difference have been applied to compare clinical interventions. Peer-reviewed journal articles that made treatment comparisons were examined. For each study, the primary hypothesis, statistical test usage, and the stated conclusion were recorded. Of the 270 studies investigated, 54.4% inappropriately made equivalence-based conclusions from difference-based test statistics (e.g., *t* test, ANOVA). Significance tests are often applied as a matter of course regardless of the research question. We have found that difference tests are similarly favored and have been applied to examine difference and inappropriately applied to examine equivalence. We discuss our findings and provide resources for researchers who want to statistically evaluate between-groups equivalence.

Keywords: equivalence tests, statistical equivalence, clinical interventions, treatment comparisons

In the behavioral sciences, researchers often need to compare two or more psychological treatments. Some researchers explore their questions qualitatively, and others explore them quantitatively. Within the quantitative domain, there are many methodologies and schools of thought on which mathematical process is the best procedure to apply in order to answer a given research question. Among those many methods are significance tests, Bayesian methods, and confidence intervals. Instead of entering into the debate of which method is the best, the present study will only focus on what is occurring in our field when researchers apply the most controversial, yet most used method: statistical tests of significance (Kline, 2004). We will, however, elaborate briefly on alternatives to significance testing and provide resources for further reading.

In some investigations, the researchers' objective is to determine whether the observed differences across treatment conditions are significant. Thus they would select a null-hypothesis of "no difference," hoping that the result of their test would be a rejection of the null, which would support the presence of the differential effect that they hypothesized they would find. In others, the aim is to determine if the treatments of interest are equivalent, in which case a null hypothesis of nonequivalence would be selected, and a rejection of that null would support treatment equivalence (within the predefined range for clinical equivalence). If researchers are using significance testing, a test can be applied to compare their data to an established distribution to see if they can reject the null, and in a perfect world, support their hypothesis. As predictions relating to difference and those relating to equivalence are distinct hypotheses, the tests of statistical significance that appropriately address them also differ. Thus, we have significance tests for difference (e.g., Student's *t* test and ANOVA) and significance tests for equivalence (e.g., Schuirmann's Two-One-Sided Test [TOST] procedure and Wellek's noncentral *F* procedure; Schuirmann, 1987; Wellek, 2010.

Methods of statistically testing for difference are familiar to most of us because these tests are presented in nearly every introductory statistics course, textbook, and statistical software package. Researchers learn during their undergraduate years to apply these tests in an attempt to answer the question: *Are these groups (or treatments) different?* Alternately, equivalence tests, can be applied when the research question is: *Are these groups (or treatments) equivalent?* (Cribbie, Gruman, & Arpin-Cribbie, 2004; Rogers, Howard & Vessey, 1993; Wellek, 2010). Equivalence tests, however, are not included in introductory statistics courses, textbooks, and most statistical software packages and thus are less popular.

What Difference Tests Can and Cannot Do

Because difference tests can only provide a yes/no answer to the question: *Is there a statistically significant difference between these groups of data?* These tests are neither able to detect, nor to provide any useful information about equivalence. Failing to detect a statistically significant difference between groups is not the same as establishing that the treatments or groups are equivalent on that given measure (Cribbie, Arpin-Cribbie & Gruman, 2009; Gordon, 1985; Kline, 2004; Schuirmann, 1987; Stegner, Bostrom, & Greenfield, 1996; Tryon & Lewis, 2008; Westlake, 1976). To conclude equivalence or nonequivalence from a difference test is a logical error that equates to saying verbally, *I failed to find a difference in my set of data; therefore, a difference does not exist, and these treatments must be equivalent* (See Cohen, 1994, for a more detailed discussion on hypothesis testing logic). If we look

Teresa A. Allan and Robert A. Cribbie, Quantitative Methods Program, Department of Psychology, York University, Toronto, Ontario, Canada.

Correspondence concerning this article should be addressed to Teresa A. Allan. E-mail: TeresaAllan@rogers.com

for an effect, and fail to find it, it is not appropriate to conclude that the effect does not exist. A real-world analogy that we might apply here to clarify this idea is to think about hypothesis testing logic in the context of a courtroom. Failing to find guilt is not the same as proving innocence, even though guilt and innocence seem to be semantically opposed in our everyday vocabulary. Like guilt and innocence, difference and equivalence are separate questions: failing to find one of these is not sufficient evidence to conclude the presence of the other.

At the level of the sample, difference tests can only provide binary answers to the question: *In the sets of data that I have collected, is there a significant difference between treatment outcomes for these two (or more) groups?* To determine this, a difference significance test, such as a Student's *t* test, can be run on the researcher's data. The results of that test are binary: significance or nonsignificance. To verbalize these ideas more plainly, a finding of significance is a statement: *Yes, there is a significant difference between these groups of data, where H_0, the null hypothesis of "no difference," has been rejected.* Conversely, *No, based upon my data, there is not a statistically significant difference between these groups on this measure*, could be a simplified verbal representation for the case when the test yields a "failure to reject the null hypothesis." When difference tests are adequately powered, most are effective tools for determining if even very small differences between groups are statistically significant.

What Equivalence Tests Can and Cannot Do

Because difference tests are an appropriate avenue within significance testing to detect statistically significant differences between treatments, equivalence tests are similarly specific to the task of detecting clinical equivalence (Stegner et al., 1996). Let us first distinguish equality from clinical equivalence (which from here on, we will simply refer to as equivalence). Equal is, as it sounds, exactly identical central tendencies. Therefore, in reality, what we might want to know is: Are these two treatments similar enough that I can recommend the shorter, less costly, or less invasive treatment to my client and have my client get the same beneficial result as the lengthier, more expensive, or more invasive treatment?

In order to examine equivalence, we must identify a range that represents inconsequential difference: an equivalence interval (Rogers et al., 1993). Imagine a situation where we might consider treatments to be equivalent if posttreatment test scores were within ± 5 points on our valid and reliable outcome measure. This would mean that we would consider a score of 80 points to be clinically equivalent to a score of 85. Providing that this difference is small enough to be considered inconsequential, that is, within a predefined range of acceptable difference, we would set our equivalence interval to $[-5, 5]$ to represent that a difference of this magnitude in either direction would be considered meaningless. It is important to point out that equivalence intervals are established as part of the research design a priori and factors such as the attributes of the participant population, reliability, the scale of the outcome variable, and the underlying nature of the study will all affect the size of the equivalence interval (Greene, Concato, & Feinstein, 2000; Rogers et al., 1993). The range of values that will be considered to be clinically equivalent will vary greatly from study to study; thus, it is up to the authors of each study to determine that range and provide evidence to support their selection. Because defining this value is a complex procedure, the interested reader is referred to Rogers et al. (1993) and Wellek (2010) for a more thorough discussion.

Like difference tests, the results of equivalence tests conducted at the level of the sample also provide us only binary answers based on our data. When the hypothetical question is: *Based on the range I have defined, are these groups of data significantly equivalent on this outcome measure?* there are only two logically correct answers that may be gleaned from the results of a significance test for equivalence: *Yes, these groups of data (treatments outcome scores) that I have collected are significantly equivalent* or *No, these groups of data that I have collected are not significantly equivalent.* Given the logic behind a binary (yes/no) question and what can be inferred from it, it is logically incorrect to say: I have failed to find these groups to be significantly equivalent based on my data, therefore a significant difference between them exists. The reason that this statement is incorrect is that a test for equivalence can neither confirm nor refute the presence of difference because difference is not being evaluated.

Present Study

The significance testing for equivalence literature cites both examples of correctly and incorrectly applied procedures (Kline, 2004; Wellek, 2010). However, the prevalence of both correct and incorrect conclusions of equivalence, and the tests used to support them, is something that has not been previously assessed in psychological treatment comparisons. The value that the present study brings to the behavioral science field is a description of the prevalence of correct and incorrect implementations of, and conclusions from, difference- and equivalence-based significance test statistics. We also provide several equivalence testing resources for those who want to use significance testing to examine equivalence, and we also present some other (also underutilized) alternatives and associated resources.

The purpose of the present study was to examine years 2000 to 2010 of the psychological literature and provide a detailed description of how significance testing statistics are being used in the behavioral sciences for cases where researchers have compared two or more treatments. Of interest, more specifically, was the prevalence of both appropriate and inappropriate conclusions of clinical difference or equivalence between treatments and how statistical tests for difference and equivalence have been used to support these statements. In addition to forming a simple accounting of how statistical tests have been used in relation to findings of equivalence, we wanted to examine the hypothesis–test–conclusion process as a whole from start to finish, per study, to determine the congruence of the overall process. To accomplish this, hypotheses, statistical tests used, and the subsequent conclusions stated were categorized in accordance with the definitions below, and these are further elaborated on in the Methods section.

Comparison Tests for Difference

We defined comparison testing for difference as those comparisons in which the researcher had hypothesized that a difference would be found between two or more treatments. A simple in-

stance of this is when a researcher hypothesized that a given treatment would be more effective than a placebo. Difference tests might also be used when there is a need to determine the effectiveness of a treatment as usual (TAU) compared with the TAU plus an additional feature—essentially to see if the new feature generates a noticeable difference on the outcome measure. A practical example of this is when the TAU for depression was administered to one group—a type of selective serotonin reuptake inhibitor (SSRI), for example, and results of that drug-only intervention were compared with another group of study participants that received both the SSRI and a new feature: weekly sessions of cognitive behavioral therapy (CBT). Presume that the study used random assignment and a complete data set was obtained for all participants across all time points. If a statistically significant difference were found, favoring the drug plus CBT condition, that finding would suggest that additional therapy sessions may have been clinically useful, provided the difference found represents a meaningful clinical effect. Although a finding such as this indicates that there is a difference between the two treatment groups, it is important to highlight that this conclusion does not provide definitive evidence of the presence of a difference. It is still possible that the statistically significant outcome represents a Type I error, even though the probability of a Type I error is typically set a comfortably low level (e.g., 5%).[1]

Comparison Tests for Equivalence

In addition to searching for ways to augment existing treatments, it may be of value to researchers and treatment providers to determine if the effects of two differing treatments can be considered to be clinically equivalent. An everyday example of a need to determine if two groups are equivalent on a posttreatment measure is when two formats of administering counseling, either face-to-face or via the Internet, are being evaluated. The research goal, in this case, would usually be to find out if the Internet therapy can provide results equivalent to face-to-face therapy. Additionally, a researcher may want to evaluate treatments for interchangeability. It may also be of value, in some cases, to determine if equivalence exists between a new drug that might be less expensive, or have fewer side effects, than an existing treatment for the same condition (e.g., does it reduce anxiety as well as the earlier formulation?) (Anderson & Hauck, 1983; Howland, 2009). A third instance of when an equivalence test might be a useful tool is when two administration periods of psychotherapy are being evaluated for efficacy and cost-effectiveness. In a study such as this, 12 weeks of therapy might be compared with both an eight- and 10-week program to see if similar benefits can be obtained in a shorter amount of time.

An important distinction between difference-based and equivalence-based tests is how sample size and effect size affect the power of these two types of tests. A traditional difference-based test gains power for detecting differences as the sample size and mean difference increase (e.g., power for detecting differences goes up when we increase the sample size of each group from 50 to 100 participants, and power also goes up when the differences in the means of the groups is increased from 5 to 10 points). On the other hand, for equivalence tests, power for detecting equivalence increases as sample sizes increase and mean differences decrease. It is important for the power of equivalence tests to increase with sample size in order for the tests to be consistent with the principles of null hypothesis testing and for the power to detect equivalence to increase as the effect size decreases, because the tests are designed to detect a situation in which there is very little difference in the means.

Method

Studies Examined

The studies examined in this study were collected from the PsycINFO database using a keyword searching procedure. The Boolean search phrase: ti("vs." OR "vs." AND "treatment*" or "therapy") was used in the PsycINFO advanced search field. (The "ti" outside of the parentheses is a shortcut to tell PsycINFO to apply this Boolean phrase to the publication title field.) The search was limited to peer-reviewed journal articles published between January 1, 2000, and December 31, 2010, and yielded over 1,000 results.

In addition to the requirement that the study must be published in English, two selection criteria were used. First, the study compared treatments for a diagnosable behavioral condition (e.g., anxiety, depression, phobias) or compared two or more treatments for a measurable psychological factor (e.g., interventions to improve one's perceived quality of life). Thus, studies that examined purely medical treatments (e.g., cardiac medication comparisons, or comparisons of treatments for broken bones) were excluded. To determine if a study examined "a measurable psychological factor," we operationally defined a measurable psychological factor as one that was obtained through the use of a psychological inventory (e.g., Minnesota Multiphasic Personality Inventory, Beck Depression Inventory, or a novel psychological measurement scale). Studies that were not selected because they did not meet this criterion included those that measured bone density as an evaluation of treatments for osteoporosis, those that used electrocardiographs to evaluate heart medications, and similar purely medical investigations. This criterion was also used to exclude single case studies. Second, The study used original data. This criterion excluded meta-analyses, which prevented the collection of redundant data. To determine if a study used "original data," we operationally defined original data studies as those that were not meta-analyses or compilations/reviews of previous works. Also, to meet this criterion, a study needed to indicate how data were collected. Studies that did not clearly state that the collection of new/unique data took place were not included in our sample.

This selection process yielded 270 current, peer-reviewed psychological comparison studies from 106 journals. Of these, there were 139 therapy studies (therapy vs. therapy, $n = 97$; therapy administration methods compared, $n = 31$; and therapy vs. placebo, waitlist, or TAU, $n = 11$), 113 pharmaceutical studies (drug vs. drug, $n = 95$; drug vs. placebo, $n = 12$; and six were additional

[1] Further, depending on individual research practices, the Type I error rate may be higher than the nominal alpha (John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011). For example, Simmons et al. (2011) point out that the number of false positives is much higher when researchers engage in questionable research practices, such as interim analysis and stopping the data collection upon achieving significance.

variants of this, some including an additional treatment of interest or waitlist control group), and 18 studies that combined differing variants of drug versus psychotherapy. An average of 24.5 studies per publication year were collected (n = 17, 20, 18, 29, 15, 29, 40, 19, 26, 29, and 28, respectively, from the years 2000 to 2010).

Procedure

Categorization of significance tests. Common examples of difference tests are Student's t test, ANOVA, chi-square, and Fisher's exact test, and examples of equivalence tests are Schuirmann's TOST, Wellek's Test, and tests that use confidence interval approaches alongside significance tests to detect equivalence (Schuirmann, 1987; Seaman & Serlin, 1998; Tryon & Lewis, 2008; Wellek, 2010; Westlake, 1976). The statistical tests that were used in each study were categorized as Difference Tests and/or Equivalence Tests based on the null hypothesis (i.e., null hypotheses of no difference were categorized as difference tests and null hypotheses of nonequivalence were categorized as equivalence tests). Studies that did not state the use of significance tests or provide the results of any mathematical comparisons were categorized accordingly.

Categorization of hypotheses and conclusions. For our purposes, studies that specifically used verbiage indicating that the study objective was to determine if two treatments or groups were "equivalent," "equal," "comparable," "similar," "as effective as" (or some combination of these terms) in their hypotheses or "purpose of the study" statements were categorized as intending to examine equivalence. Concluding phrases stating that the treatment(s) of interest "are both effective," "is as effective as," "is an alternative to," "are comparable," "are similar," the "new treatment is equal to old treatment," "are equal"/"equally effective," and "are equivalent" were categorized as a conclusion relating to the state of equivalence between the treatments or therapies (see Figure 1).

Hypotheses that specifically predicted difference, either directional or nondirectional, were classified as intending to examine difference. Conclusions that only stated that there was "no significant difference" between treatments, or conclusions that only stated that "a difference" or "significant difference" was found were coded as conclusions relating to the state of difference between groups. If hypotheses were unstated, exploratory, or minimal and nonpredictive (e.g., "our study *evaluated* Drug A vs. Drug B for the treatment of . . ."), these were classified as "intending to conduct comparisons." Conclusions that did not specifically conclude difference or equivalence as a result of the study were few and were categorized as descriptive or inconclusive.

Congruence

Three types of congruence were examined: Hypothesis-Test Congruence, Test-Conclusion Congruence, and Overall Congruence (Hypothesis-Test-Conclusion). When hypothesis types matched test types (e.g., difference was hypothesized, and a difference test was used) this was coded as Hypothesis-Test congruence. Conversely, when the hypotheses predicted equivalence and evaluated that prediction with a statistical difference test, this was coded as Hypothesis-Test incongruence. Test-Conclusion Congruence was evaluated similarly: for example, those that conducted difference tests and stated a conclusion relating to the state of difference that was found (or not found) were coded as congruent, and those that conducted difference tests and used them to formulate conclusions regarding a state of equivalence were coded as incongruent. Finally, Overall Congruence was defined as the state where all three (hypothesis, test, and conclusion) matched. For example, a difference hypothesis, examined by a difference test, followed by a conclusion regarding the state of difference was congruent (e.g., *we predict that treatment A will be better than treatment B on this given measure*, tested this using a t test, and concluded that a significant difference was found). Studies were only coded as having overall congruence if the hypothesis, test, and conclusion were consistent (difference-difference-difference, or equivalence-equivalence-equivalence). (Further details about this coding process can be found in the Appendix).

For each article, the above variables were recorded by the first author, and subjective decisions were decided jointly by both of the authors. In a small number of cases where the study was ambiguous to both authors in terms of the type of testing procedure that was used or the study did not state a clear hypothesis or conclusion, categories of "unclear" and "not stated" were used. As noted in the results section, those studies were not included in analyses where these variables were necessary.

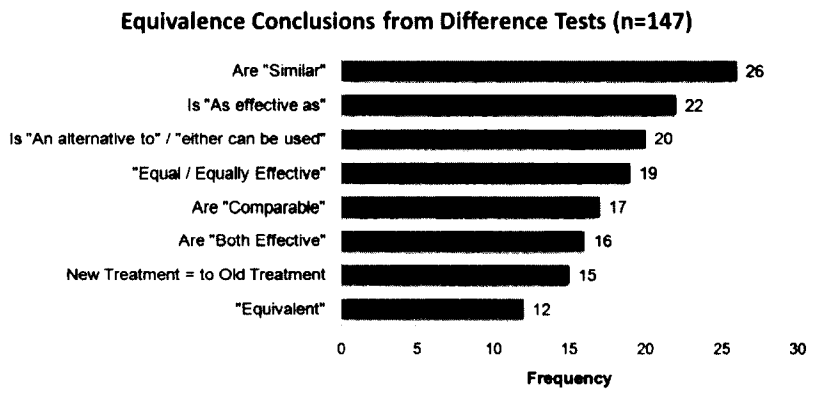

Figure 1. Summary of study conclusion statements of equivalence from difference tests (n = 147).

Results

Hypotheses, Tests, and Conclusions

Of the 270 studies examined, it was found that 116 (43%) provided specific predictive hypotheses. Of those providing specific hypotheses, 91 predicted that a significant difference between treatments would be found, and 25 studies predicted that the treatments of interest would be equivalent (see Table 1). The remaining 57% (n = 154) stated no specific hypothesis, stating only that comparisons were being made, treatments were being evaluated, or the objective of their study was described in terms of Treatment A "versus" Treatment B and made no outcome predictions. Two hundred sixty-five studies used difference tests to evaluate their hypotheses. Two studies used equivalence tests. Five studies did not state that any type of mathematical analysis was used to test their hypotheses and thus provided no statistical data. (It could be that these researchers did not apply statistical tests or applied tests but did not mention them in their publications.) It must also be noted that two of the studies used both an equivalence test and a difference test to evaluate the same set of data.

Forty percent of the studies we examined (n = 109) stated conclusions of difference. Twenty of these used difference tests and correctly stated in the study conclusion that there was "no significant difference" between groups, or that "no differences were found," and the remaining 89 correctly concluded that statistically significant differences were found. Seven studies provided conclusions in the form of a descriptive, where treatments were described through the provision of observed frequencies or in terms of pros and cons rather than in terms of equivalence or difference. Two studies stated that based on their statistical findings, their studies were inconclusive at this time and warranted further research.

One hundred forty-seven (54.4%) stated conclusions of equivalence (see Figure 1). Two studies used equivalence tests to conclude difference between treatments. One of these used both an equivalence test and a difference test on their primary outcome variable first to detect difference and then applied an equivalence test to (inappropriately) confirm it. The other study used an equivalence test to check for differences that an initially applied difference test "might have missed." That study found no significant difference and also failed to find significant equivalence and subsequently concluded that because there was no equivalence, there may be a difference between the treatments that the difference test failed to detect (see Table 1).

Table 1
Distribution of Tests, Hypotheses, and Conclusions

Distributions	n
Statistical tests (n = 270[a])	
Used difference tests to evaluate treatment comparison	265
Used equivalence tests to evaluate treatment comparison	2
Did not use (or did not state the use of) statistics	5
Hypotheses (n = 270)	
Primary hypothesis related to state of difference	91
Primary hypothesis related to a state of equivalence	25
Hypothesis was stated as "vs." or "comparison"	154
Conclusions (n = 265[b])	
Conclusion relating to state of difference	109
Conclusion relating to state of equivalence	147
Conclusion listed as a set of descriptives	7
Conclusion stated as "inconclusive"	2

[a] Two studies used both a difference test and an equivalence test to evaluate their primary hypothesis. [b] The five studies that did not use or did not indicate the use of statistical testing were not included in this analysis.

Congruence

Three types of congruence were examined: Hypothesis-Test Congruence, Test-Conclusion Congruence, and Overall Congruence (Hypothesis-Test-Conclusion) (see the Appendix).

Hypothesis-Test Congruence. Hypothesis-Test Congruence could not be calculated for the 154 studies (57% of the overall sample) that stated their hypotheses as comparisons. Of the 115 studies that provided specific hypotheses and used statistical testing to evaluate their primary hypothesis, 89 of them were congruent with the applied statistical test, and 26 were not. The 89 that were congruent all hypothesized that a difference would be found between the treatments of interest and used difference tests to determine if that difference was statistically significant. Twenty-five that were incongruent hypothesized equivalence and used difference tests to determine if the treatments were equivalent (the study that used an equivalence test to look for differences that the difference test might have missed was also coded as incongruent). Of the cases that used statistical testing to test stated hypotheses, all but two of those that hypothesized difference appropriately tested for difference, and no cases hypothesizing equivalence statistically tested for equivalence. Thus, 97.8% of the studies that stated hypotheses relating to difference correctly adopted a test of difference, and 0% of the studies that stated hypotheses relating to equivalence correctly adopted tests of equivalence.

Test-Conclusion Congruence. Test-Conclusion Congruence could not be calculated for 14 studies—seven of these provided only descriptive conclusions, two were inconclusive, and five did not indicate the use of statistical testing. The remaining 256 used statistical tests and provided either equivalence conclusions (n = 147) or difference conclusions (n = 109). One hundred seven of these were congruent with the statistical tests used, and 149 were not. One hundred forty-seven of the 149 studies that were incongruent were so because they used difference tests to support conclusions of equivalence. We found no applications of equivalence tests that were congruent with stated conclusions. The two equivalence tests that were applied were classified as incongruent because in one of these, the authors used an equivalence test to test for differences that were not found by an initial difference test and the other used an equivalence test to confirm an established difference. The majority of our sample (265/270) applied statistical tests of significance and reported the results of these statistical tests as the supporting evidence for their concluding statements. Of those that used statistical significance testing, 58.2% (149/256) stated conclusions that were logically incompatible with the results of the statistical tests that were used.

Overall (Hypothesis-Test-Conclusion) Congruence. We defined Overall Congruence as a complete match between the hypothesis, statistical test, and conclusion. Therefore, the studies that could be evaluated for this variable were only those that presented

all three items (n = 113). Studies that were coded as being congruent overall either had a difference hypothesis, used a difference test, and stated a conclusion relating to the state of difference or stated an equivalence hypothesis, used an equivalence test, and presented a conclusion relating to the observed state of equivalence or nonequivalence. There were 55 studies that met these criteria. All had difference hypotheses, used difference tests, and stated conclusions relating to an observed state of significant or nonsignificant difference. All studies (n = 25) that hypothesized equivalence used difference tests to test for equivalence. Although 22 of these studies concluded that the psychological treatments of interest were equivalent, the test statistics that were used in all of these cases were not appropriate to test for equivalence.

Discussion

Inappropriately Applied Difference Tests and Testing Logic

Kline (2004) and Wellek (2010) have both noted that significance tests seem to be selected and applied regardless of the research question or objective. We have found that within significance testing a similar preference seems to exist for difference tests. Difference tests seem to be selected and applied in treatment comparisons whether the objective is to examine these treatments for equivalence or for difference and also seem to be applied when the research is exploratory. Although this is not an incorrect application of these tests in cases where researchers want to use significance testing as a method of examining difference, it is inappropriate to form conclusions of equivalence from these tests. In line with the result we have obtained here, erroneous conclusions of equivalence from the application of significance tests for difference have been well noted in both the psychological and medical literature (Cribbie et al., 2004, 2009; Greene et al., 2000; Rusticus & Lovato, 2011; Tryon, 2001).

There seems to be a clear need for a better understanding of what information can be gleaned from the application of a statistical procedure and for further guidelines for researchers who want to examine equivalence. This is evinced by finding 25 studies that specifically hypothesized that equivalence would be found, and 147 studies that concluded (based on statistical tests for difference) that the treatments of interest produced clinically equivalent results. All of these studies opted to use significance testing, and within significance testing, none selected equivalence tests. Instead, all of them used traditional difference tests, primarily Student's t and ANOVA, to provide support for concluded states of equivalence. As noted earlier, this is logically incorrect, because difference test statistics cannot be used to support conclusions of equivalence. In contrast, 90 of the 91 studies that hypothesized a difference would be found or stated they were looking for a difference used difference statistics to evaluate their data. Over half of these (n = 55) stated conclusions that were appropriately supported by the test statistics that were used. Thirty-four, however, concluded equivalence—which again leads us to believe that although there is a need for processes to test for equivalence, there is some confusion about how to mathematically address this task if a researcher wants to evaluate between-groups equivalence within the framework of significance testing.

It may simply be that researchers are unaware that this option is available. Most of us are familiar with significance tests in their more common forms (Student's t test, ANOVA) because these tests are presented in nearly every introductory statistics course and every introductory statistics textbook. Some of us are also familiar with these tests because most, if not all, statistical software packages include them in a relatively easy-to-use format (e.g., R, SPSS, SAS, and even Microsoft Excel). Perhaps it is simply due to this ease of access and early exposure that difference-based significance testing remains the most highly prevalent data analysis testing method.

For those wanting to read further about significance testing for equivalence, we briefly list the following resources. Please note, that the equivalence analogues to the difference tests listed here are subject to the same limitations and assumptions as their null hypothesis of no difference counterparts. One equivalence test that is analogous to Student's t test is Schuirmann's TOST. In this test, data are presumed to be normal and homoscedastic, and two hypotheses representing equivalence are generated and tested. Both of these hypotheses must be rejected in order to support that groups are equivalent (Schuirmann, 1987). When data are still normal, but have unequal variances, the analogue to the Welch t test, is the Schuirmann-Welch Test. The interested reader is referred to Gruman, Cribbie, and Arpin-Cribbie (2007) for a demonstration and study of this method. One test that has been proposed to be analogous to the ANOVA is the Wellek Test of Equivalence (Koh & Cribbie, 2012; Wellek, 2010).

Lack of Hypotheses

Another important observation was that of the full sample of 270 studies, 154 stated no hypotheses or specific purpose for the study other than to conduct comparisons. Although in many instances it is of importance to conduct nonpredictive exploratory comparisons between treatments, it seems unlikely that over half of the sample was engaging in exploratory research. If researchers want to use significance testing to examine their nonexploratory questions, it is integral to developing and following a statistical significance testing plan that a hypothesis is generated. If it is possible to establish a concrete study purpose, it follows that it is relatively simple to select a type of statistical test that is able to provide meaningful information about that hypothesis so that a logically appropriate conclusion (that is supported by the data) can be formulated.

Conclusion

Our primary purpose for conducting this study was to describe how, in the behavioral science field, statistical tests are being applied in the comparison of two or more psychological treatments. As mentioned above, researchers use significance tests more than any other testing method. Researchers seem to apply tests of significance almost automatically without considering whether a significance test is the best test to help them answer their research questions. In the present sample, we have found that difference tests are the most prevalent significance testing choice. Difference tests seem to be similarly applied across the board whether the researcher is examining treatment outcome data for difference or for equivalence. Even in cases where the research is

exploratory, difference tests are the most prevalent. Equivalence tests, in this domain, are underutilized. We have noted that many researchers inappropriately make equivalence-based conclusions when the result of a difference test is not statistically significant. This is inappropriate mathematical support for an equivalence conclusion because the test statistic that was selected did not test for equivalence.

Although many statistical errors can be avoided within the domain of significance testing by selecting a test that is suited to examining the stated hypothesis and presenting only conclusions that are appropriate to the results of these tests, many statisticians and psychological researchers do not believe null hypothesis significance testing is useful.[2] To remain neutral in this debate, we would like to clarify that we have focused our discussion on significance testing because, for better or worse, it is the most utilized means of data analysis in the behavioral sciences (Cohen, 1994). We recognize that significance testing, and specifically, equivalence testing, is not a one-size-fits-all statistical solution to describe the relationship between sets of data and groups of study participants. Data analysis methods should be selected on a case-by-case basis, very carefully, keeping in mind what can and cannot be inferred from these analyses. By better understanding the information a statistical test can provide and by selecting the most appropriate analytical method that we can to examine our research questions, the quality and accuracy of psychological treatment research can be improved.

[2] Whereas proponents of significance testing point out that statistical significance tests are an easy to use, readily accessible means to objectively test hypotheses, it has long been questioned how well significance testing works to accurately examine our data due to the logic underlying the test and whether or not researchers have an accurate understanding of what information the test provides (Daniel, 1998; Kline, 2004). As the debate has been thoroughly documented, this is a brief overview, and we refer the interested reader to the cited resources for methodological demonstrations of alternatives to significance testing and a more complete picture of the debate.

It has been pointed out that the p values that we come to know in our earliest introductory statistics courses in the form of the probability associated with values obtained by running a Student's t test and in the analysis of variance (ANOVA), are potentially problematic because, alone, p values do not quantify the statistical evidence, nor do they provide details of the magnitude of the observed effect (Cohen, 1994; Wagenmakers, 2007). One method of solving this problem is to augment reports of significance test statistics with effect sizes and confidence intervals, and the APA recommends that researchers provide this additional information whenever possible (American Psychological Association, 2009; Thompson, 1999; Wilkinson & APA Task Force on Statistical Inference, 1999).

Confidence intervals have also been presented both as a stand-alone method of analysis and as a way to augment other statistical tests by illustrating the margin of error surrounding the observed data that are being mathematically tested (American Psychological Association, 2009; Thompson, 1999; Wilkinson & APA Task Force on Statistical Inference, 1999). One advantage of using confidence intervals is that visually these can be easier to interpret because the interval contains both information about the magnitude of the observed effect and the measure of uncertainty associated with observed values (Hoekstra, Kiers, & Johnson, 2010). Mau (1998) presents a discussion and formulas that can be used to examine the overlapping regions of confidence intervals as a means of assessing equivalence. The computation of a Bayes Factor has been presented as an alternative to significance testing (p values) (García-Pérez, 2012; Wagenmakers, 2007; Wetzels et al., 2011).

Résumé

Dans la recherche en science du comportement, il est souvent nécessaire de déterminer si une variable de résultat diffère ou est équivalente parmi des groupes. Des tests de signification constituent la méthode d'analyse des données la plus utilisée pour répondre à cette question. Cette étude avait pour but d'examiner comment les analyses statistiques visant à déterminer l'équivalence ou la différence ont été appliquées pour comparer des interventions cliniques. Des articles de revues ayant été évalués par des pairs et comportant des comparaisons de traitements ont été examinés. Pour chaque étude, on a consigné l'hypothèse principale, l'usage d'une méthode statistique et la conclusion énoncée. Parmi les 270 études de l'échantillon, 54,4 % arrivaient à des conclusions erronées d'équivalence au moyen de méthodes statistiques reposant sur la variance (par ex., test t, ANOVA). On a souvent recours à des tests de signification, de façon routinière, peu importe le sujet de la recherche. Nous avons conclu que les tests de variance sont aussi privilégiés et qu'ils ont été utilisés pour examiner des différences et mal appliqués pour analyser l'équivalence. Nous discutons de nos résultats et citons des ressources à l'intention de chercheurs qui veulent évaluer d'un point de vue statistique l'équivalence entre des groupes.

Mots-clés : tests d'équivalence, équivalence statistique, interventions cliniques, comparaison de traitements.

References

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Anderson, S. A., & Hauck, W. W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics: Theory and Methods, 12,* 2663–2692. doi: 10.1080/03610928308828634

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003. doi:10.1037/0003-066X.49.12.997

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

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

Daniel, L. G. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *Research in the Schools, 5,* 23–32.

García-Pérez, M. A. (2012). Statistical conclusion validity: Some common threats and simple remedies. *Frontiers in Psychology: Quantitative Psychology and Measurement, 3,* 1–11.

Gordon, R. S. (1985). Three current issues: The design and conduct of randomized clinical trials. *IRB A Review of Human Subjects Research, 7,* 1, 3, 12.

Greene, W. L., Concato, J., & Feinstein, A. R. (2000). Claims of equivalence in medical research: Are they supported by the evidence? *Annals of Internal Medicine, 132,* 715–722. doi:10.7326/0003-4819-132-9-200005020-00006

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

Hoekstra, R., Kiers, H. A. L., & Johnson, A. (2010). The influence of presentation on the interpretation of inferential results (Invited Paper). *International Association of Statistical Education.* Available at http://130.203.133.150/viewdoc/summary?doi=10.1.1.205.795

Howland, R. H. (2009). What makes a generic medication generic? *Journal of Psychosocial Nursing and Mental Health Services, 47,* 17–20.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23,* 524–532. doi:10.1177/0956797611430953

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association. doi:10.1037/10693-000

Koh, A. & Cribbie, R. A. (2012). Robust tests of equivalence for *k* independent groups. *British Journal of Mathematical and Statistical Psychology.* Advance online publication. doi:10.1111/j.2044-8317.2012.02056.x

Mau, J. (1988). A statistical assessment of clinical equivalence. *Statistics in Medicine, 7,* 1267–1277. doi:10.1002/sim.4780071207

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113,* 553–565. doi:10.1037/0033-2909.113.3.553

Rusticus, S. A., & Lovato, C. Y. (2011). Applying tests of equivalence for multiple group comparisons: Demonstration of the confidence interval approach. *Practical Assessment, Research & Evaluation, 16,* 1–6.

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

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. doi:10.1177/0956797611417632

Stegner, B. L., Bostrom, A. G., & Greenfield, T. K. (1996). Equivalence testing for use in psychosocial and services research: An introduction with examples. *Evaluation and Program Planning, 19,* 193–198. doi:10.1016/0149-7189(96)00011-0

Thompson, B. (1999). If statistical significance tests are broken/misused, what practices should replace them? *Theory & Psychology, 9.* 165–181.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods, 6,* 371–386. doi:10.1037/1082-989X.6.4.371

Tryon, W. W., & Lewis, C. (2008). An inferential confidence interval method of establishing statistical equivalence that corrects Tryon's (2001). reduction factor. *Psychological Methods, 13,* 272–277. doi:10.1037/a0013158

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14,* 779–804. doi:10.3758/BF03194105

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority,* (2nd ed.). New York, NY: CRC Press. doi:10.1201/EBK1439808184

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

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science, 6,* 291–298. doi:10.1177/1745691611406923

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604. doi:10.1037/0003-066X.54.8.594

(Appendix follows)

Appendix

Congruence Coding

Table A1 shows the distribution of statistical tests, hypotheses, and conclusions.

Table A1
Overall Congruence Coding Conceptualization

Hypothesis	Test	Conclusion	Congruence
Stated a hypothesis	Used a test appropriate to test the hypothesis	Stated a conclusion logically corresponding to the test that was conducted	Yes
Stated a hypothesis	Used a test inappropriate to test the hypothesis	Stated a conclusion logically corresponding to the test that was conducted	No
Stated a hypothesis	Used a test appropriate to the hypothesis	Stated a conclusion not logically corresponding to the test that was conducted	No

Table A2 shows the overall congruence coding.

Table A2
Overall Congruence Coding in Greater Detail

Hypothesis	Test type	Conclusion statement	Congruent
Predicted difference	Difference	Equivalent	No
Predicted no difference	Difference	Equivalent	No
Predicted difference	Difference	Not equivalent	No
Predicted no difference	Difference	Not equivalent	No
Predicted difference	Difference	Found no significant difference	Yes
Predicted no difference	Difference	Found no significant difference	Yes
Predicted difference	Difference	Found significant difference	Yes
Predicted no difference	Difference	Found significant difference	Yes
Predicted equivalent	Equivalence	Equivalent	Yes
Predicted not equivalent	Equivalence	Equivalent	Yes
Predicted equivalent	Equivalence	Not equivalent	Yes
Predicted not equivalent	Equivalence	Not equivalent	Yes
Predicted equivalent	Equivalence	Found no significant difference	No
Predicted not equivalent	Equivalence	Found no significant difference	No
Predicted equivalent	Equivalence	Found significant difference	No
Predicted not equivalent	Equivalence	Found significant difference	No

Received September 14, 2012
Revision received May 7, 2013
Accepted May 9, 2013 ■