Practical Assessment, Research, and Evaluation

Volume 16 *Volume 16, 2011* Article 7

2011

Applying Tests of Equivalence for Multiple Group Comparisons: Demonstration of the Confidence Interval Approach

Shayna A. Rusticus

Chris Y. Lovato

Recommended Citation

Rusticus, Shayna A. and Lovato, Chris Y. (2011) "Applying Tests of Equivalence for Multiple Group Comparisons: Demonstration of the Confidence Interval Approach," *Practical Assessment, Research, and Evaluation*: Vol. 16 , Article 7.
DOI: https://doi.org/10.7275/d5wf-5p77
Available at: https://scholarworks.umass.edu/pare/vol16/iss1/7

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 16, Number 7, April 2011 ISSN 1531-7714

Applying Tests of Equivalence for Multiple Group Comparisons: Demonstration of the Confidence Interval Approach

Shayna A. Rusticus and Chris Y. Lovato
University of British Columbia

Assessing the comparability of different groups is an issue facing many researchers and evaluators in a variety of settings. Commonly, null hypothesis significance testing (NHST) is incorrectly used to demonstrate comparability when a non-significant result is found. This is problematic because a failure to find a difference between groups is not equivalent to showing that the groups are comparable. This paper provides a comparison of the confidence interval approach to equivalency testing and the more traditional analysis of variance (ANOVA) method using both continuous and rating scale data from three geographically separate medical education teaching sites. Equivalency testing is recommended as a better alternative to demonstrating comparability through its examination of whether mean differences between two groups are small enough that these differences can be considered practically unimportant and thus, the groups can be treated as equivalent.

The challenge of assessing the comparability of different groups is an issue facing many researchers and evaluators. Occasionally the question of interest is not one of whether two or more groups (or treatments or methods) are different from one another, but rather one of whether the groups can be considered the same. A prime example of this is the work of William Blackwelder which has established the importance of examining the equivalence of clinical trials (e.g., Blackwelder 1982, 2004). The purpose of this paper is to practically demonstrate a method of assessing comparability among two or more groups. Recognizing that it is more effective to illustrate a statistical technique in the context of an example rather than describe it in abstract terms, this paper will focus on a medical education example to facilitate the demonstration.

Determining the comparability of teaching methods and different geographical sites is an issue facing many evaluators working in educational settings. With the expansion and development of geographically separated medical education programs and the increased use of technology for curriculum delivery, medical schools have the challenge of ensuring the comparability of students' educational experiences across program sites and/or methods of instruction. The Liaison Committee on Medical Education (LCME) accreditation standards state, "There must be comparable educational experiences and equivalent methods of evaluation across all alternative instructional sites within a given discipline" (ED-8).

It has been common practice to use analysis of variance (ANOVA) methods (or t-test methods in some two group cases) to demonstrate the equivalence of alternative instructional sites or modes of instruction. These statistical methods (classified under the umbrella of null hypothesis significance testing; NHST) test the hypothesis that groups are statistically different on a particular outcome measure, with the null hypothesis stating that the groups are not statistically different. For instance, Bianchi, Stobbe, and Eva (2008) were interested in comparing the academic performance of students studying at rural versus urban settings. These researchers used ANOVAs to examine group differences on multiple types of assessment scores. The non-significant findings were interpreted as showing that "academic performance among students was at least comparable across all learning sites" (p. 67). Waters, Hughes, Forbes and Wilkinson (2006) also made academic comparisons across students in rural and urban clinical settings using ANOVA methods. Again, non-significant findings were used to conclude that "academic performance among students studying in rural and urban settings is comparable" (p. 117). Hatala, Issenberg, Kassan, Cole, Bacchus and Scalese (2008) used ANOVA to address their research question assessing the "comparability of clinical competence using [real patients] compared with that using simulation technology" (p. 629). Other examples of research studies using the same procedures/logic include Fydryszewski, Scanlan, Guiles, and Tucker (2010), Lovato and Murphy (2008), McFall and Freddolino (2000), and McKendry, Busing, Dauphinee, Braiovsky, and Boulais (2000). The point here is not to criticize the work of these researchers, but to highlight that this has been common practice for examining whether distributed sites or instructional methods are comparable.

The problem with using ANOVA methods is that a statistically non-significant value (failure to find a group difference) is used to imply that the groups are comparable. To be precise, however, a statistically non-significant finding only indicates that there is not enough evidence to support that two (or more) groups are statistically different. It does ***not*** show evidence for the null hypothesis being true; that is, it does not show any evidence for the groups being comparable. It is possible that the two groups are comparable, but it is also possible that the study did not have enough power to detect a statistical difference, there was high variability in the sample, and/or that the study was poorly designed. Concluding equivalence based on a lack of a statistically significant difference has been identified as one of the most common misuses of NHST (Tryon, 2001). Thus, using this method does not properly address the question of comparable educational experiences, including results from student assessments.

While NHST is appropriate to answer questions about whether group differences exist, it is not appropriate to provide evidence for comparability (whether this intention is explicitly stated or covertly implied). To correctly address questions about comparability, the real question to be answered is whether two (or more) groups are equivalent. Note here that the key word is equivalent, not equal. One does not, and should not, expect two groups to be exactly equal – that is virtually impossible to do. Rather, the goal is to demonstrate that the differences that do exist between the groups are small enough that, for practical purposes, the groups can be treated as equivalent. Equivalency testing can be used to accomplish this purpose.

Equivalency testing assesses whether mean differences between two groups are small enough that the groups can be considered equivalent/similar (i.e., differences found are considered practically unimportant; Blackwelder, 2004; Rogers, Howard, & Vessey, 1993). As noted by Rogers et al. (1993), there are three general categories of equivalency tests: the confidence interval approach (also known as the two one-sided tests procedure), the nonequivalence null hypothesis approach and Bayesian methods. The approach that we have chosen to use is the confidence interval approach (Rogers et al., 1993; Schuirmann, 1987; Westlake, 1976) because of its popularity as an equivalency testing method and its ease of use and interpretation. Briefly stated, this approach calculates a confidence interval around the mean difference between two groups. If this confidence interval is within a specified range (the equivalence interval) then the groups are said to be equivalent. Thus, the first, and most important, step in conducting equivalency testing is to operationalize equivalency prior to statistical testing. Equivalency is described by Rogers et al. (1993) as "the minimum difference between two groups that would be important enough to make the groups nonequivalent" (p. 554). As the difference between two groups could be in either a positive or negative direction, there is both a positive and a negative value used to define equivalence, forming an equivalence interval. Lewis, Watson, and White (2009) recently noted that there are no set standards for equivalence intervals; although ±20% appears to be the most commonly used in the areas of bioequivalence and social science. However, both Lewis and colleagues (2009) and Rogers and colleagues (1993) caution against the thoughtless use of rules of thumb and advise that the equivalence interval selected should be relevant for its particular use and based on a strong rationale. The second step in conducting tests of equivalence using the confidence interval approach is to construct a 90% confidence interval around the mean group difference on the outcome measure (Rogers et al., 2003). Equivalence can be concluded if the confidence interval is contained within the equivalence interval.

This article will demonstrate the confidence interval approach to equivalency testing. Traditional ANOVAs are also presented to provide a comparison to the equivalence test results, as the former method has been the most widely used approach. The results will be of particular interest to evaluators and researchers working in settings in which demonstrating the comparability of groups is of concern.

METHODS

Data

Assessment data. Second year assessment data from four cohorts of students enrolled at the University of British Columbia (UBC) were included in this analysis (n = 884). See Table 1 for the distribution of students by year and campus site. These students completed their second year courses between 2006 and 2009. Assessment data was collected from students at all three sites of the UBC distributed program (herein referred to as Site 1, Site 2 and Site 3) and included exam scores for Gastroenterology, Blood and Lymphatics,

DOI: https://doi.org/10.7275/d5wf-5p77

Musculoskeletal and Locomotor, Endocrine and Metabolism, Integument, Brain and Behaviour, Reproduction, Growth and Development, Doctor, Patient and Society, Family Practice, and Clinical Skills.

Table 1: *Participants by Site and Year*

	Site			
Year	Site 1	Site 2	Site 3	Total
2006	22	24	141	187
2007	24	23	166	213
2008	23	23	177	223
2009	30	33	198	261
Total	99	103	682	884

Rating scale data. Self-reported student course evaluation data was also selected for analysis to demonstrate equivalency testing for rating scale data. These data included anonymous responses from 270 second year medical students in 2008 or 2009 (29 in Site 1, 36 in Site 2, 205 in Site 3). Only two years of data were selected because it was only these two years in which the items asked were identical. The surveys were distributed electronically to all students across sites at the end of the course.

The student course evaluation data examined students' educational experiences in the Blood and Lymphatics course and consisted of the following dimensions: direction, learning support, level of engagement, confidence level, and overall satisfaction. The direction dimension (4 items) measures students' satisfaction with direction provided in the course to focus learning. The learning support dimension (9 items) measures students' satisfaction with three aspects of learning support: course content, learning materials and instructional strategies. The level of engagement dimension (2 items) measures students' level of engagement with learning activities based on relevance and interest in course content and activities. The confidence level dimension (2 items) measures students' perception of performance confidence within each course and compared with other courses. The overall satisfaction single item measures a global rating of students' overall satisfaction with the course. With the exception of the confidence level dimension, which is rated along a 7-point scale, all other dimensions are assessed using a 5-point response format.

Analyses

Statistical equivalency between sites was tested using the confidence interval approach outlined by Rogers and colleagues (1993). To conduct tests of equivalence, a critical a priori decision must be made regarding an equivalence interval that is relevant and appropriate to the particular context. This represents the boundaries of the difference between the means of two groups (positively or negatively) that would indicate a meaningful difference (i.e., at this difference or greater the groups are *not* equivalent; Rogers et al.,1993). Any group mean difference found that is within the equivalence interval indicates that the difference is not practically meaningful and the groups can be treated as equivalent/considered comparable.

Based on an examination of the literature, internal studies, and discussions with relevant stakeholders, the following equivalence intervals were used for the present study: (1) ±5% between groups (Site 1-Site 2, Site 1-Site 3, Site 2-Site 3) for the assessment data, (2) ±1.0 points between groups for the 5-point rating scale items, and (3) ±1.4 points[1] between groups for the 7-point rating scale item.

A series of one-way ANOVAs were conducted to investigate traditional group differences, as well as to calculate the 90% confidence intervals on the pair wise mean group differences using a Games-Howell post-hoc test[2]. It is these 90% confidence intervals, taken from the post-hoc comparisons, which were used to test whether the mean group differences were within the equivalence intervals. If the confidence intervals were within the equivalence intervals, equivalency was concluded.

RESULTS

Assessment Data

Table 2 presents the means and standard deviations for the assessment data for each of the second year medical school courses. Table 3 presents the equivalency test results for each of the courses. Equivalency testing showed that: (1) Site 1 and Site 2 were statistically equivalent for 5 out of 11 courses. (2) Site 1 and Site 3 were statistically equivalent for all 11 courses, and (3) Site 2 and Site 3 were statistically equivalent for 7 out of 11 courses.

Table 3 also presents the ANOVA results for each of the second year medical school courses. Using a non-significant ANOVA result as an indicator of group equivalence showed that: (1) Site 1 and Site 2 were equivalent for 2 out of 11 courses). (2) Site 1 and Site 3 were equivalent for 10 out of 11 courses, and (3) Site 2 and Site 3 were equivalent for 2 out of 11 courses.

[1] This criterion was calculated in reference to the 5-point scale: 1/5*(point value) = 1/5*7 = 1.4.

[2] Games-Howell was selected because this method takes unequal group sizes into account, as well as violations of homogeneity of variance (which occurs more often with unequal group sizes). Additionally, this method has been shown to perform well when groups are homogenous (Dunnett, 1980). The only course to violate the variance assumption was Endocrine and Metabolism.

Table 2: *Means and Standard Deviations of Exam Scores for Second Year Medical School Courses Grouped by Site*

	Site 1		Site 2		Site 3	
Course	*M*	*SD*	*M*	*SD*	*M*	*SD*
Gastroenterology	77.08	8.30	77.55	8.38	74.10	9.28
Blood and Lymphatics	83.76	7.60	84.75	7.50	81.57	7.88
Musculoskeletal and Locomotor	84.15	7.85	83.19	7.29	80.08	7.77
Endocrine and Metabolism	85.29	6.97	85.45	7.84	80.66	9.94
Integument	83.57	10.15	83.85	9.12	79.37	9.85
Brain and Behaviour	81.08	6.97	80.63	6.81	76.69	7.90
Reproduction	82.40	7.90	82.10	7.39	79.83	7.38
Growth and Development	79.48	8.01	78.84	7.89	76.45	6.55
Doctor, Patient, and Society	86.54	3.32	85.79	3.59	85.90	3.02
Family Practice	87.47	3.76	85.19	4.27	84.89	4.94
Clinical Skills	77.41	4.81	77.02	5.22	74.03	5.07

Table 3: *Equivalence and ANOVA Test Results for Second Year Exam Scores*

	90% CI						ANOVA		
	Site 1-Site 2		Site 1-Site 3		Site 2-Site 3				
	Lower	Upper	Lower	Upper	Lower	Upper	*p*	η^2	Group Difference
Gastroenterology	0.43	5.54[a]	-2.31	1.38	-5.46[a]	-1.44	.001*	.01	Site 1-2, Site 2-3
Blood and Lymphatics	-0.06	4.44	-2.68	0.70	-4.90	-1.47	<.001*	.02	Site 2-3
Musculoskeletal and Locomotor	1.80	6.34[a]	-0.77	2.69	-4.79	-1.42	<.001*	.02	Site 1-2, Site 2-3
Endocrine and Metabolism	2.15	7.12[a]	-1.74	1.42	-6.92[a]	-2.67	<.001*	.04	Site 1-2, Site 2-3
Integument	1.29	7.10[a]	-2.52	1.95	-6.62[a]	-2.34	<.001*	.02	Site 1-2, Site 2-3
Brain and Behaviour	2.23	6.56[a]	-1.09	2.00	-5.64[a]	-2.24	<.001*	.03	Site 1-2, Site 2-3
Reproduction	0.34	4.79	-1.45	2.04	-3.89	-0.66	.012*	.01	Site 1-2, Site 2-3
Growth and Development	0.89	5.16[a]	-1.15	2.42	-3.86	-0.92	.008*	.01	Site 1-2, Site 2-3
Doctor, Patient, and Society	-0.36	1.64	-0.06	1.57	-0.64	-0.87	.184	.01	--
Family Practice	1.30	3.85	1.42	3.14	-1.36	-0.77	<.001*	.03	Site 1-2, Site 1-3
Clinical Skills	1.95	4.82	-0.69	1.48	-4.10	1.88	<.001*	.04	Site 1-2, Site 2-3

Note. Equivalence interval = ±5.00 [a] Groups are not equivalent $*p < .05$

Rating Scale Data

Table 4 presents the means and standard deviations and Table 5 presents the equivalency test results for each dimension of the student course evaluation data. Equivalency testing showed that criterion for equivalency was met across all sites and dimensions; that is, all groups are comparable. Table 5 also presents the ANOVA results for each of the student course evaluation dimensions. Using a non-significant result as an indicator of group equivalence showed that all sites are comparable for each of the assessed dimensions.

Table 4: *Means and Standard Deviations of the Student Course Evaluation Data Grouped by Site*

	Site 1		Site 2		Site 3	
Dimension	*M*	*SD*	*M*	*SD*	*M*	*SD*
Direction	4.71	0.34	4.69	0.47	4.64	0.43
Learning Support	4.45	0.39	4.53	0.48	4.50	0.45
Engagement	4.64	0.47	4.69	0.50	4.64	0.46
Confidence	5.66	1.22	5.83	1.06	5.76	0.98
Overall	4.14	0.93	4.06	0.95	3.77	0.81

DOI: https://doi.org/10.7275/d5wf-5p77

Table 5: *Equivalence and ANOVA Test Results for the Student Course Evaluation Data*

		90% CI						ANOVA		
	Equivalence Interval	Site 1-Site 2		Site 1-Site 3		Site 2-Site 3				
Dimension		Lower	Upper	Lower	Upper	Lower	Upper	p	η^2	Group Difference
Direction	±1.00	-0.13	0.27	-0.14	0.17	-0.22	0.11	.769	.00	--
Learning Support	±1.00	-0.27	0.17	-0.26	0.09	-0.21	0.14	.650	.00	--
Engagement	±1.00	-0.24	0.25	-0.25	0.16	-0.23	0.13	.789	.00	--
Confidence	±1.40	-0.69	0.48	-0.68	0.33	-0.45	0.31	.687	.00	--
Overall	±1.00	-0.10	0.84	-0.32	0.48	-0.61	0.03	.192	.01	--

Note. Confidence dimension rated on a 7-point scale; all other dimensions rated on a 5-point scale
[a] Groups are not equivalent *p < .05

DISCUSSION

This paper demonstrates tests of equivalence using the confidence interval approach to show comparability of two or more groups. In the field of medical education, a key accreditation standard requires that distributed medical education sites and methods of instruction be comparable in terms of program quality and evaluation. This paper highlights how previous research has tended to incorrectly use null hypothesis significance testing as a means of demonstrating comparability. Instead, tests of equivalence are recommended to demonstrate that any differences found between groups are small and unimportant; thus, the groups can be considered comparable. However, because significance testing has been the approach that has been most commonly used to demonstrate equivalence, ANOVA results were also calculated for each of the outcome measures and compared to the equivalency results.

For the rating scale data, both the equivalence and ANOVA results concluded equivalence for all groups on all dimensions assessed. For the assessment data, there were mixed results. The results of the equivalency tests showed that Site 1 and Site 3 were comparable across all courses. With one exception, the ANOVA results were consistent with these findings (i.e., there was no statistically significant difference). For Site 2 and Site 3, the results of the equivalence tests found statistical equivalence for 7 of the 11 courses. However, the ANOVA results concluded equivalence for only two of the courses. Finally, for Site 1 and Site 2, the equivalence tests found statistical equivalence for five of the courses, while the ANOVA results concluded equivalence for only two of the courses.

For all of the site comparisons (i.e., Site 1 vs. Site 2, Site 2 vs. Site 3, Site 1 vs. Site 3), the equivalency tests were more likely to find equivalence than the ANOVA results. Overall, equivalency testing found that 23 of the 33 group comparisons could be considered equivalent, while significance testing concluded that only 14 of the 33 group comparisons could be considered comparable. There are three possible explanations for these findings. One, the equivalence intervals that were constructed for this research study were based on a rationale that was appropriate for the evaluation goals of the university. These criteria are likely not equal to the criterion of 0.05 that is used to indicate a statistical difference in significance testing. Two, significance testing is known to be sensitive to sample size, such that it is easier to find a statistical difference when sample sizes are large, as in the present study. Thus, while these findings are found to be statistically different, they may not have practical significance. A lack of practical significance is supported by the small effect sizes for the differences found in the ANOVA results. Three, equivalency testing and significance testing are different methodologies that test a different hypothesis. The former tests whether differences between groups are small enough that the groups can be considered equivalent; whereas the latter tests whether differences between groups are large enough that the groups can be considered different.

One issue to note is that by using the confidence intervals that are calculated as part of the post-hoc comparisons, the error term that is used in the calculation of the confidence intervals is from the omnibus F test. This error term is used to control type I error (concluding groups are equivalent when they are not equivalent) in the pair-wise comparisons because of the multiple comparisons being conducted. However, there is the possibility that this may result in the confidence intervals being too wide and thus the probability that the confidence interval is contained within the equivalency interval may be low (i.e., type II error). This is an important line of research that should be explored further.

In summary, as demonstrated in this study, and in other studies (e.g., Cribbie, Gruman, Arpin-Cribbe, 2004; Lewis et al., 2009; Rogers at al., 1993) NHST and equivalence testing are not analogous. If the goal is to demonstrate that two groups are equivalent/comparable, then equivalency testing is the recommended procedure to use, with the added recommendation that the equivalence interval selected should be appropriate for the given research and evaluation context.

References

Bianchi, F., Stobbe, K., & Eva, K. (2008). Comparing academic performance of medical students in distributed learning sites: The McMaster experience. *Medical Teacher, 30,* 67-71.

Blackwelder, W. C. (1982). "Proving the null hypothesis" in clinical trials. *Controlled Clinical Trials, 3,* 345-353.

Blackwelder, W. C. (2004). Current issues in clinical equivalence tests. *Journal of Dental Research, 83,* 113-115.

Cribbie, R. A., Gruman, J. A., & Arpin-Cribbie, C. A. (2004). Recommendations for applying tests of equivalence. *Journal of Clinical Psychology, 60,* 1-10.

Dunnett, C. W. (1980). Pairwise multiple comparison in the homogenous variance, unequal sample size case. *Journal of the American Statistical Association, 75,* 789-795.

Fydryszewski, N. A., Scanlan, C., Guiles, H. J., & Tucker, A. (2010). An exploratory study of live vs. web-based delivery of a phlebotomy program. *Clinical Laboratory Science, 23,* 39-45.

Hatala, R., Issengerg, S. B., Kassen, B., Cole, G., Bacchus, C. M., & Scalese, R. J. (2008). Assessing cardiac physical examination skills using simulation technology and real patients: A comparison study. *Medical Education, 42,* 628-636.

Lewis, I., Watson, B., & White, K. M. (2009). Internet versus paper-and-pencil survey methods in psychological experiments: Equivalence testing of participant responses to health-related messages. *Australian Journal of Psychology, 61,* 107-116.

Liaison Committee on Medical Education. Functions and structure of a medical school. 2008. www.lcme.org/functionslist.htm (accessed 24 November 2010).

Lovato, C. Y. & Murphy, C. C. (2008). Comparability of student performance and experiences in UBC's distributed MD undergraduate program: The first 2 years. *BC Medical Journal, 50,* 380-383.

McFall, J. P. & Freddolino, P. P. (2000). Quality and comparability in distance field education: Lessons learned from comparing three program sites. *Journal of Social Work Education, 36,* 293-307.

McKendry, R. J., Busing, N., Dauphinee, D. W., Brailovsky, C. A., & Boulais, A. (2000). Does the site of postgraduate family medicine training predict performance on summative examinations? A comparison of urban and remote programs. *Canadian Medical Association Journal, 163,* 708-711.

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

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

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminancy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis significance tests. *Psychological Methods, 6,* 371-386.

Waters, B., Hughes, J., Forbes, K., & Wilkinson, D. (2006). Comparative academic performance of medical students in rural and urban clinical settings. *Medical Education, 40,* 117-120.

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

Citation:

Rusticus, Shayna A. & Chris Y. Lovato (2011). Applying Tests of Equivalence for Multiple Group Comparisons: Demonstration of the Confidence Interval Approach. *Practical Assessment, Research & Evaluation*, 16(7). Available online: http://pareonline.net/getvn.asp?v=16&n=7.

Authors:

Shayna A. Rusticus
University of British Columbia
Faculty of Medicine
Evaluation Studies Unit
shayna.rusticus [at] ubc.ca

Chris Y. Lovato
University of British Columbia
Faculty of Medicine
School of Population & Public Health

DOI: https://doi.org/10.7275/d5wf-5p77