

# The Same or Not the Same: Equivalence as an Issue in Educational Research

Scott E. Lewis and Jennifer E. Lewis\*

Department of Chemistry, University of South Florida, Tampa, FL 33620; \*jlewis@cas.usf.edu

In a recent submission to this *Journal*, we described the results of a hybrid reform method, peer-led guided inquiry (PLGI), aimed at improving students' chemistry understanding (1). For the evaluation of the effectiveness of PLGI we employed a quasi-experimental design (2, p 341), comparing an experimental section with a control section, to discern whether the technique provided any noticeable effects. As part of the design, we attempted to control for as many variables as possible between the experimental section and the control section. For example, both sections were taught the same semester by the same instructor and took the same exams at the same time. In addition, we compared the average SAT scores for both sections to determine whether there was any difference between the two sections with respect to this cognitive measure. Unfortunately, the method we employed for this comparison, though commonly used, was not suitable for establishing equivalence.

## The Conventional Method

The method selected for the comparison was an independent samples  $t$  test, which was used to examine whether the difference in average SAT scores between the two sections was statistically significant. The use of such a test is based on the null hypothesis ( $H_0$ ) that the two means ( $\mu_{\text{exp}}$  and  $\mu_{\text{cont}}$ ) are equal, and the alternative hypothesis ( $H_1$ ) that they are not:

$$H_0: \mu_{\text{exp}} = \mu_{\text{cont}} \quad \text{and} \quad H_1: \mu_{\text{exp}} \neq \mu_{\text{cont}}$$

From a  $t$  test, a  $p$  value is obtained so that it can be compared with a predetermined alpha value ( $\alpha$ ), typically 0.05. If the  $p$  value is below 0.05, then the criterion for rejecting the null hypothesis is met, and the alternative hypothesis is accepted; however, for our purposes we were interested in determining whether two groups (two sections of a particular course) were equivalent. Running the  $t$  test on our cognitive measure and receiving a value above 0.05, we concluded that there was no statistically significant difference between the two groups. In other words, we decided to *accept* the null hypothesis on the basis of a  $p$  value obtained from a  $t$  test. The acceptance of the null hypothesis, then, was our basis for the statement that the two sections in our study were equivalent according to a standard cognitive measure.

This method for investigating equivalence has also been utilized in several other recent studies in this *Journal* (3–7), but it is not robust. While the  $t$  tests provided no evidence to contradict equivalence in our study and in these other studies, the  $t$  test itself was not designed for *establishing* equivalence. First, although the  $t$  test does allow rejection of the

alternative hypothesis, this rejection does not constitute, statistically speaking, an acceptance of the null hypothesis. Second, even the rejection of the alternative hypothesis is probabilistic.

For example, if a  $t$  test is used to compare two means, and the result from the test indicates that  $p$  is less than or equal to 0.05 (the preset  $\alpha$  for comparison in most cases), the researcher may conclude that there is a significant difference between the two means. This result carries with it at most a 5% chance of error (this number arises from the use of  $\alpha = 0.05$ ). In other words, the researcher can be 95% certain that there is a significant difference between the two means. The remaining 5% chance that this conclusion is in error is termed type I error, and it refers to the likelihood that the null hypothesis was in fact true even though the  $p$  value recommended rejection. The chance of an incorrect rejection occurring (type I error) is known before the test is performed and corresponds exactly with the pre-chosen  $\alpha$ .

Similarly, if the same  $t$  test is done but the results show that  $p$  is greater than 0.05, the researcher may conclude that the two samples have the same mean, since the  $p$  value did not permit the acceptance of the alternative hypothesis. This conclusion is also probabilistic, but in this case type II error, which applies to the decision to accept the null hypothesis when in fact it should be rejected, is of prime importance. Unfortunately, unlike the case of the preset  $\alpha$  that informs type I error considerations, there is no preset control in place to limit the chance of type II error. While type I errors are generally considered more serious, in the quest to determine equivalence of two groups (as in the scenario described initially, in which the desire is to show that two sections of the same class are equivalent in terms of a cognitive measure), type II error and the probability of type II error ( $\beta$ ) is pivotal. Discussions of equivalence informed by conventional  $t$  tests are therefore incomplete without an estimate of  $\beta$ .

## Power Analysis and Equivalence

A conventional method for estimating  $\beta$  is found in power analysis. Power analysis estimates the power of a statistical test, which is the probability that a test will correctly lead to the rejection of the null hypothesis (8, p 4). A detailed discussion of how power depends on sample size, type I error and effect size can be found in Stevens' *Intermediate Statistics* (9, pp 121–124). The complement of power is the error rate of failing to reject the null hypothesis, or  $\beta$ . When power is estimated, type II error can be accounted for in any decision regarding failing to reject the null hypothesis. In the equivalence example given above, if the data fail to reject the

null hypothesis ( $p$  value greater than 0.05) and the power of the test is found to be 0.9, then the conclusion may be made that the two samples are equivalent, since the probability of a type II error is 10%. However, if in the same example, the power is found to be only 0.5, claims of equivalence are much more suspect, since there is a 50% chance of type II error. Traditionally, a power of 0.8 (20% probability of type II error) has been used as a cutoff in this decision-making process (10, p 17). The power for a variety of statistical tests can be estimated using Cohen's *Statistical Power Analysis for the Behavioral Sciences* (8).

To perform a power analysis, effect size must be estimated. The effect size is designed to measure "the degree to which the phenomenon is present in the population" (8, p 9). For the comparison of two means, effect size is characterized by Cohen's  $d$

$$d = \frac{\mu_{\text{exp}} - \mu_{\text{cont}}}{\sigma} \quad (1)$$

where  $\sigma$  is equal to the standard deviation for either particular group.<sup>1</sup> Cohen describes a  $d$  of 0.2 as a small effect size, one in which "the influence of uncontrollable extraneous variables (noise) makes the size of the effect small relative to these" (8, p 25). Since any difference in means that can be attributed to the noise of the sample should not be grounds for declaring groups nonequivalent, this effect size should be appropriate for the estimation of power. Using our own data with this effect size and Cohen's tables, the power of the  $t$  test we used to determine whether the two sections in our study had different mean SAT scores can be estimated to be 27%. Thus there is a 27% chance that the two groups were equivalent, and a 73% chance that the groups are different but the  $t$  test was unable to find this difference (due to insufficient sample size). Therefore, by the aforementioned criteria, there was a failure to establish equivalent SAT scores between the two groups in our study. Fortunately, attempts to establish equivalence need not end with power analysis.

### A Better Way To Examine Equivalence

In the pharmaceutical industry there has been a sustained interest in developing tests of equivalence to determine whether, for example, a generic substitute has equivalent effects as compared to the original product. In a comparison of procedures for testing equivalence, Schuirmann (11) indicates that power analysis becomes a very conservative test when equivalence is sought, especially when the variance is small or the sample size is large. This can eventually lead to occasions where there is little or no chance of concluding that the means are equivalent when they are. Lest other chemical education researchers founder, as we did, on the basis of power analysis and conclude that they may not be working with equivalent groups, we propose a re-visioning of the null hypothesis for education research based on Schuirmann's work in which equivalence of two groups is being sought. This approach is both more focused and more robust than the conventional consideration of type I and type II error discussed above.

Schuirmann evaluated the use of the "two one-sided  $t$  tests" procedure. In this procedure, an interval is established

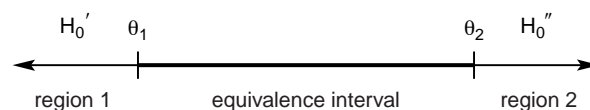


Figure 1. The relative positioning of the two null hypotheses and the equivalence interval.

$(\theta_1, \theta_2)$  and a test is run on each boundary of the interval to see whether the difference in means falls within this interval. The two null hypotheses are that the difference in means falls outside the interval, as depicted in Figure 1. Written mathematically, the two null hypotheses have to be expressed in two separate equations. For the case in which the difference in means may fall in region 1

$$H_0' : \mu_{\text{exp}} - \mu_{\text{cont}} < \theta_1$$

and for the case in which the difference may fall in region 2

$$H_0'' : \mu_{\text{exp}} - \mu_{\text{cont}} > \theta_2$$

The rejection of both of these hypotheses provides evidence for the alternative hypothesis

$$H_1' : \theta_1 < \mu_{\text{exp}} - \mu_{\text{cont}} < \theta_2$$

To address such issues as equivalence of means, where the attempt is to show that

$$\mu_{\text{exp}} - \mu_{\text{cont}} = 0$$

$\theta_1$  can be set equal to  $-\theta_2$  to build a symmetric interval around 0. In this approach it can be seen that demonstrating equivalence does not rely on the failure to reject the null hypothesis as it did in the power approach, but rather equivalence is shown via the rejection of both null hypotheses. The null hypotheses can be tested by a pair of one-sided  $t$  tests

$$t_1 = \frac{(\bar{X}_{\text{exp}} - \bar{X}_{\text{cont}}) - \theta_1}{s_p \sqrt{\frac{1}{N_{\text{exp}}} + \frac{1}{N_{\text{cont}}}}} \geq t_{1-\alpha(v)} \quad (2a)$$

$$t_2 = \frac{\theta_2 - (\bar{X}_{\text{exp}} - \bar{X}_{\text{cont}})}{s_p \sqrt{\frac{1}{N_{\text{exp}}} + \frac{1}{N_{\text{cont}}}}} \geq t_{1-\alpha(v)} \quad (2b)$$

where  $s_p$  is the pooled standard deviation within means,<sup>2</sup>  $N$  is the total number of subjects in each group, and  $\bar{X}$  is the group mean. Rejecting both null hypotheses, in this context, would then be the same as indicating that the difference in means, augmented by a range that spans  $1 - 2\alpha$  (where  $\alpha$  is the value used for the one-sided  $t$  tests in eq 2), falls completely in the  $(\theta_1, \theta_2)$  interval. We chose an  $\alpha$  level of 0.10 to make the test analogous to the 80% power suggestion described previously. The selection of the  $(\theta_1, \theta_2)$  interval is an important decision that Schuirmann is quick to point out should be made by the experts in the field to which the data

apply, not the statistician. In short, decisions regarding  $\theta_1$  and  $\theta_2$  should be based on what would be an acceptable range within which the two group means could vary but still be said to arise from equivalent groups.

### Interval Construction for SAT Scores

In attempting to perform the two one-sided  $t$  tests procedure with our own data, it is clear that the selection of an appropriate ( $\theta_1$ ,  $\theta_2$ ) interval is the most difficult decision to make. Schuirman proposes the possibility of 0.2 times the mean value of the control group as a suitable value for  $\theta_2$ , with  $\theta_1$  still equal to  $-\theta_2$ . However, when dealing with values such as SAT subscores, which range from 200 to 800, this decision adds 40 points to the interval (in both directions), as a result of the scale not beginning at 0. Even when scaling the subscores down so that they range from 0 to 600, the decision would still produce a range of  $\pm 60$  points (assuming the average was 300 on this scaled-down version). To those accustomed to looking at SAT scores this range still seems rather large when looking for equivalent groups on the basis of group mean scores. Interestingly, one source quotes a representative of the Educational Testing Service, with regards to SAT subscores "...that differences of less than 60 points should not be considered significant" (12, p 72), though this was meant for consideration of individual students, and comparing mean scores is thought to have a much smaller range of variability. In addition, this source makes no mention of how the 60 point rule-of-thumb was derived.

With the desire to construct a more conservative interval for declaring equivalence, we returned to Cohen's operationalization of effect size. Using eq 1 as described above, and an effect size of 0.2 to characterize any noise in the data, the interval can then be determined by solving for the difference in means. This serves to quantify the specific degree of difference between means that Cohen would describe as noise. Since we are concerned with a difference in means in either direction, the equivalence interval will then be set up to range from the negative of this value to the positive of this value.

For our own data, this provides an equivalence interval of (-18.8, 18.8) for Verbal SAT and (-18.2, 18.2) for Math SAT. Note that both these equivalence intervals are more conservative than the proposed  $\pm 60$  interval. With this interval decided, the two one-sided  $t$  tests can be applied to our data to determine whether the two groups are equivalent. Finally, our own data provide the following results, for Verbal SAT

$$t_1 = \frac{(536.09 - 534.79) - (-18.8)}{90.257 \sqrt{\frac{1}{69} + \frac{1}{142}}} = 1.52$$

$$t_{(\alpha=0.10)} = 1.29 \Rightarrow t_1 \geq t_{(\alpha=0.10)}$$

and

$$t_2 = \frac{18.8 - (536.09 - 534.79)}{90.257 \sqrt{\frac{1}{69} + \frac{1}{142}}} = 1.32$$

$$t_{(\alpha=0.10)} = 1.29 \Rightarrow t_2 \geq t_{(\alpha=0.10)}$$

Since each  $t$  test leads to the rejection of each null hypothesis, it can be concluded that the difference in means falls entirely within the interval of (-18.8, 18.8), with at most a 20% chance of being outside that interval. Using this method of equivalence testing for Math SAT produced similar results, so the same claim can be made with regards to Math SAT scores in the interval of (-18.2, 18.2). Thus a more exact claim regarding equivalence has been made by use of the two one-sided  $t$  tests method than in the conventional method discussed previously. In this case, the use of a robust equivalence test supported our initial assessment of the control and experimental groups as equivalent, but agreement between the initial  $t$  test and a more robust equivalence test may not always exist.

### Implications for JCE Authors

We examined all the articles in the *Journal* with a keyword of chemical education research from January 2000 through August 2003, looking for indications of seeking equivalence between two groups on a scaled measure. Only one of the five articles found in our review (4) provided the information necessary (standard deviations as well as group means and sample sizes) for a reader to perform this equivalence test. The authors should be commended for providing the information necessary for other researchers to examine their work. In addition, this article provides us with a chance to examine ACT scores, which serve as another popular cognitive measure.<sup>3</sup> We decided to examine data from an actual *JCE* research article to illustrate both how this procedure can be performed on data presented in the *Journal*, and how the employment of such a procedure may have implications for research findings.

For our examination of the published data, the two one-sided  $t$  tests were run in a similar manner as above; however, two decisions had to be made on how to treat the data. First, we had to assume that all the students in the study had ACT scores available. Second, in constructing the equivalence interval, as before we used Cohen's effect size with  $d = 0.2$ , but we used the national ACT standard deviation values (13) rather than the standard deviations reported in the article for construction of the equivalence interval. (The reported standard deviations were still used to compute the standard deviation within means for use in eq 2.) Both decisions were made to maximize the likelihood of finding the groups equivalent: the first decision provides the largest possible sample size; the second decision uses the largest standard deviation (all the national deviation values were larger than any values reported on the sample) to construct the largest equivalence interval possible. The results are shown in Table 1.

Based on the results shown in Table 1, there was a failure to establish equivalence on three of the ACT scores. The discussion of results in the article makes clear that the only measure for which experimental and control groups were different was the chemistry conceptual assessment constructed by the authors. With the authors finding no statistical differences between these groups on ACT scores, and our failure to establish equivalence on several of these scores using the above analysis, the best interpretation may simply be that the sample size was insufficient to determine whether there was a difference in ACT scores between groups. We would not question the authors' finding that demonstration assess-

ments were fruitful, but suggest that ACT scores may need to be identified as a potentially confounding variable. This opens another avenue for analysis, and including statistical control of ACT scores may provide additional insights into students' performance on the chemistry conceptual assessment as well as on the other outcome measures (for which no significant difference was found).

### Further Work on Equivalence Intervals

As part of this interpretation, the dependence of the two one-sided  $t$  tests on the construction of the equivalence interval cannot be overemphasized. This article presents data from two standard cognitive measures, SAT and ACT, and constructs equivalence intervals for them. Researchers using these two measures can recompute equivalence intervals based on their own sample means and standard deviations for these two measures and others. Routinely publishing these intervals with the sample size, mean, and standard deviation information used to determine them will allow greater scrutiny on the issue of equivalence in educational research. Since Schuirmann stresses the necessity of input from researchers working in the field regarding the construction of sensible equivalence intervals, the method of interval construction for these two measures is laid out explicitly here so that other researchers can judge the results for themselves. The choice of Cohen's  $d = 0.2$  to compute an acceptable level of noise may result in too conservative an interval, and the development and justification for alternative methods for establishing equivalence intervals would be welcome.

### Beyond Equivalence

Testing equivalence of groups is really about the control of potentially confounding variables. There are multiple methods of controlling variables, and typically the methods are determined during experimental design. The ideal way to implement control of variables is by designing a "true experiment" where participants are randomly assigned to experimental and control groups. This is preferable for its ability to promote equivalence between groups on all variables, including those not explicitly measured in the study. However, this experimental design may not always be possible, especially in situations of educational research on intact classes. Frequently a quasi-experimental design is employed in which efforts are made to approximate a true experiment. In the quasi-experimental scenario, it becomes necessary to identify and control potentially confounding variables. Equivalence tests are typically used to check to what extent

equivalent groups were achieved with regard to a particular variable. What is always certain is that the failure to find a difference does not mean no difference exists, so that a specific test for equivalence with regard to a particular variable is necessary if one needs to show that two groups do not differ in terms of that variable. This article has presented and demonstrated a robust statistical test for the comparison of two groups on a single scaled measure that can be used to support a claim of equivalence.

A failure to establish equivalence does not signal the end of a planned research study. It may be that there was an insufficient sample size to demonstrate equivalence although the samples were equivalent. If this type of statistical verification of equivalence is necessary, there would be a need to redesign the study to enlist more participants. The feasibility of this option depends on other parameters of the study. How difficult would it be to include additional participants? How important is it that the groups are demonstrated to be equivalent with regard to the variable in question? Are they equivalent on other measures? Demonstrating equivalence is helpful, but if it cannot be demonstrated, it may truly be an indication that a confounding variable is present. In this case, insight can be gained by explicitly acknowledging the effects of the potentially confounding variable, in which case the researcher may need to rely on statistical control of the variable (e.g., inclusion of the variable as a covariate, see Williamson's treatment of the TOLT; ref 14) during the examination of the results. This path is preferable to claiming a failure to reject the null hypothesis as demonstrated equivalence, since it provides more information regarding the effects of the variable in question.

### Conclusion

The commonly used method of reliance on failure to reject the conventional null hypothesis as a positive indication of equivalence has been questioned. The alternative method we propose, based on a re-visioning of the null hypothesis, also avoids the uncontrolled type II error present in the traditional method. In addition a procedure for employing this method has been discussed, along with an interpretation of results. Finally, through the use of this method, the need for published sample sizes, standard deviations, and group means in quantitative studies has become evident, for both the employment of the equivalence tests presented in this paper and for any further tests desired by other researchers. If information of this type is not included in the article, this *Journal's* online supplemental features are an appropriate venue.

**Table 1. Comparison of Published Average ACT Scores**

ACT Area	$\bar{X}_{\text{exp}}$	$\bar{X}_{\text{cont}}$	Interval	$t_1$	$t_2$	$t_{\alpha=0.10}$	Result
Math	28.30	27.75	(-1, 1)	2.562	0.905	1.29	Not within interval
Reading	26.16	26.62	(-1.22, 1.22)	0.983	1.987	1.29	Not within interval
Science	26.66	26.60	(-0.94, 0.94)	1.592	1.434	1.29	Equivalent
English	25.76	26.12	(-1.08, 1.08)	1.235	2.126	1.29	Not within interval
Composite	26.66	26.91	(-0.94, 0.94)	1.629	2.378	1.29	Equivalent

## Notes

1. One of the assumptions of the comparison of means is that both groups will have identical standard deviations:  $\sigma = \sigma_{\text{exp}} = \sigma_{\text{cont}}$ . For cases where the standard deviations of each group are not identical, the square root of the mean of two variances can be used (8, p 44)

$$\sigma = \sqrt{\frac{\sigma_{\text{exp}}^2 + \sigma_{\text{cont}}^2}{2}}$$

2. The pooled standard deviation,  $s_p$ , within means can be found by

$$s_p = \sqrt{\frac{SS_{\text{exp}} + SS_{\text{cont}}}{N_{\text{exp}} + N_{\text{cont}} - 2}}$$

where SS is the sum of squares for each group (15, p 217). Alternatively, this value can be found by running an ANOVA in a statistical software package with the control group and the experimental group as the two values for the categories. In the resulting output the “mean square, within groups” is found. The pooled standard deviation within means is the square root of this value.

3. The authors are indebted to Deese et al. for providing this opportunity to examine ACT scores in this context. In our current data, ACT scores are only available for approximately 50% of our students and are not representative of our sample as a whole.

## Literature Cited

- Lewis, Scott E.; Lewis, Jennifer E. *J. Chem. Educ.* **2005**, *82*, 135–139.
- Babbie, Earl W. In *The Practice of Social Research*, 8th ed.; Wadsworth: Boston, 1998; pp 333–355.
- Bradley, Alexander Z.; Ulrich, Scott M.; Jones, Maitland, Jr.; Jones, Stephanie M. *J. Chem. Educ.* **2002**, *79*, 514–519.
- Deese, William C.; Ramsey, Linda L.; Walczyk, Jeffrey; Eddy, Danny. *J. Chem. Educ.* **2000**, *77*, 1511–1516.
- Mason, Diana; Verdel, Ellen. *J. Chem. Educ.* **2001**, *78*, 252–255.
- Nicoll, Gayle; Francisco, Joseph; Nakhleh, Mary. *J. Chem. Educ.* **2001**, *78*, 1111–1117.
- Rudd, James A.; Greenbowe, Thomas J.; Hand, Brian M.; Legg, Margaret J. *J. Chem. Educ.* **2001**, *78*, 1680–1686.
- Cohen, Jacob. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Lawrence Erlbaum Associates: Hillsdale, NJ, 1988.
- Stevens, James P. *Intermediate Statistics: A Modern Approach*, 2nd ed.; Lawrence Erlbaum Associates: Mahwah, NH, 1999.
- Murphy, Kevin R. *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests*; Lawrence Erlbaum Associates, Inc.: Mahwah, NJ, 1998.
- Schuirman, Donald J. *J. Pharmacokin. Biopharm.* **1987**, *15*, 657–680.
- Owen, David; Doerr, Marilyn. *None of the Above: The Truth Behind the SATs, Revised and Updated*; Rowman & Littlefield Publishers, Inc.: Lanham, MD, 1999.
- The 1997 ACT High School Profile Report — National Normative Data. <http://www.act.org/news/data/97/t1.html> (accessed May 2005).
- Williamson, Vickie. M.; Rowe, Marvin W. *J. Chem. Educ.* **2002**, *79*, 1131–1134.
- Nowaczyk, Ronald H. *Introductory Statistics for Behavioral Research*; Holt, Rinehart and Winston Inc.: New York, 1988.