

New Developments in Equivalence Testing
Daniel M. Ennis and John M. Ennis

Background: *Equivalence* is a term used to describe a condition in which two items differ by less than some predetermined amount. Similar terms include *bioequivalence*, *parity*, *equality* and *similarity*. Bioequivalence is used in the context of drug effect differences while the term parity is often used in the context of consumer product differences. Equality has been used in a variety of contexts, including the recently reissued American Society for Testing and Materials (ASTM) claims guide,¹ while it seems preferable to avoid using the term similarity as it has a different meaning altogether in the quantitative psychology literature.^{2,3,4} The ASTM claims guide distinguishes between two types of parity advertising claims, respectively termed equality and unsurpassed. An equality claim can be made when two products are essentially equivalent within some predetermined bounds on an attribute of interest. This means that the term equality could be misleading as it could be misunderstood to imply that there are no bounds involved in its definition. This is in spite of the fact that such bounds are required. In what follows we simplify and use the single term equivalence in place of all of the above application-specific terms.

The domain of applications for equivalence testing is quite broad. This domain includes false advertising cases brought under the provisions of the Lanham Act (Title 15, §1125 of the United States Code), changes in an ingredient or process made to reduce cost, improve healthfulness or comply with government regulations, and changes made to food products to address a public health issue such as childhood obesity. In this last case, the objective may be to produce a more healthful product that has a similar taste or odor profile to an existing product. Another application arises when a pharmaceutical company wishes to market a drug that is more efficacious than an existing drug and wishes to show that it is as safe to use as the existing drug.

In the context of drug testing, the method of the “Two One-Sided Tests” (TOST) due to Westlake⁵ and Schuirmann⁶ has received much notoriety. In a recent paper⁷ we identified a fundamental shortcoming of the TOST and proposed a theory of equivalence testing that is more consistent with the classical theory of hypothesis testing. In this report we review this new theory of equivalence testing and apply the new theory to a clinical trial scenario.

Scenario: Your company produces nicotine patches for use by smokers as an aid in smoking cessation. Although there are important differences in nicotine

delivery between patches and cigarettes, such as the ‘bolus’ effect of nicotine due to smoke inhalation rather than continuous infusion, patches have nevertheless been found to be a useful cessation aid among smokers who wish to quit smoking. Psychopharmacological effects of nicotine in humans include effects on information processing and memory as well as adverse side effects involving the cardiovascular and digestive systems. Your company has an active research program on the development of nicotine analogs to find new compounds that offer the psychopharmacological effects of nicotine without many of the negative side effects. In fact, your company has discovered an alternative to nicotine that is as psychopharmacologically effective as nicotine when concentration levels are comparable. The question remains whether your patch system will deliver this new drug as effectively as it does nicotine. To answer this question you conduct a small scale experiment to test for equivalence. Before we discuss the results of your experiment we review some background regarding equivalence hypothesis testing.

Equivalence Hypotheses: Equivalence hypothesis testing, like all hypothesis testing, requires a clear statement of both the null and alternative hypotheses. In the case of equivalence testing, the null hypothesis is that the items tested differ by more than some predetermined amount. Correspondingly the alternative hypothesis, which is the hypothesis of equivalence, is that the items tested differ by less than that amount. For instance, it has been common practice to define two drugs to be equivalent when their relative performances on a clinical measure of efficacy, usually measured on a logarithmic scale, falls within $\log(0.8)$ and $\log(1.25)$. The $\log(0.8)$ and $\log(1.25)$ bounds imply that one of the drugs is between 80% and 125% as efficacious as the other

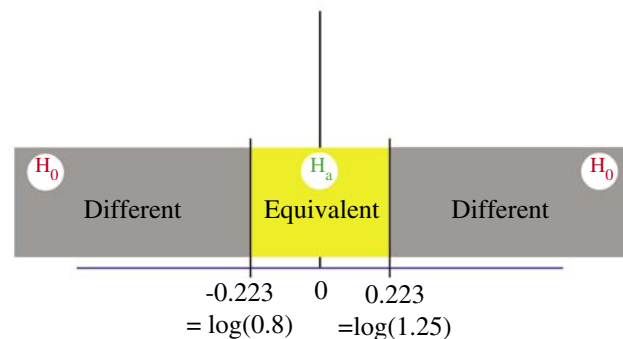


Figure 1. Predetermined Bounds for Equivalence Testing

Figure 1 illustrates the null and alternative hypotheses in equivalence testing using these boundaries. Note that other predetermined bounds could be appropriate in other applications. For instance in a preference test one could consider two products to be equivalent if the true preference proportion lies between 0.45 and 0.55. These bounds are recommended in the ASTM claims guide for the definition of equivalence in a preference test. Also note that for our present purposes we assume that the bounds are symmetric about the point of equality.

In our recent paper on this topic we began with an explicit test of a null hypothesis of non-equivalence in the case of binomially distributed data and provided an exact test in the case of normally distributed data with known variance. Along similar lines, tables for testing for equivalence with binomially distributed data have also been published⁸. When the variance is unknown, we have recommended that an adjustment value be used in the calculation of the test statistic. Details concerning how to calculate this adjustment value are given in our paper, a preprint of which is available at www.ifpress.com. The following is the formula for calculating the test probability:

$$p = \Phi\left(\frac{|x| - \sqrt{c}\theta}{s}\right) - \Phi\left(\frac{-(|x| + \sqrt{c}\theta)}{s}\right).$$

In this formula, x is the measure of the difference in efficacy between the two treatments, c is the adjustment value, θ is the upper bound that defines equivalence, s is the sample standard error and $\Phi(y)$ is the area under the standard normal distribution function from $-\infty$ to y . Table 1 contains an excerpt of the adjustment values needed for small experiments when the variance is unknown.

Degrees of Freedom	c
10	0.8204
15	0.8750
20	0.9038
30	0.9344
50	0.9600

Comparison of Nicotine and the Nicotine Analog:

There were twelve participants in your clinical trial in which subjects received one of the treatments and after a washout period received the second treatment in a balanced order. The log of the difference between treatments was 0.03 with a sample standard error of

0.11. Using the tabulated value of 0.8204 for c with 10 degrees of freedom, you reject the null hypothesis using $\theta = \log(1.25)$ at $\alpha = 0.05$ because

$$p = \Phi\left(\frac{0.03 - \frac{\sqrt{0.8204} \log(1.25)}{0.11}}{0.11}\right) - \Phi\left(-\left(\frac{0.03 + \frac{\sqrt{0.8204} \log(1.25)}{0.11}}{0.11}\right)\right) = 0.0414.$$

Your conclusion from conducting this small clinical trial is that the data support the alternative hypothesis that the drugs are equivalent. It is worth noting that in this example the TOST would not reject the null hypothesis. Following this test, you plan a larger clinical trial to broaden the subject pool and to further evaluate the new drug.

Conclusion: A new method for equivalence hypothesis testing has been developed recently. This method is superior to existing methods in that its test statistic has been derived directly from a null hypothesis of non-equivalence. This new method allows for exact testing in the cases of binomial data and normally distributed data with known variance. In the case of normally distributed data with unknown variance, an adjustment is possible that allows for a test that is generally more powerful than the TOST and is as simple to conduct. Based on the results of such a test, one can weigh a null hypothesis of non-equivalence against an alternative hypothesis of equivalence.

References:

- ¹ASTM E 1958. (2006). Standard Guide for Sensory Claim Substantiation, ASTM International.
- ²Ennis, D. M. and Johnson, N. L. (1993). Thurstone-Shepard similarity models as special cases of moment generating functions. *Journal of Mathematical Psychology*, **37**(1), 104-110.
- ³Tversky, A. (1977). Features of similarity. *Psychological Review*, **84**(4), 327-352.
- ⁴Ashby, F.G. and Ennis, D.M. (2007). Similarity measures. *Scholarpedia*, **2**(12):4116.
- ⁵Westlake, W.J. (1981). Response to T.B.L. Kirkwood: bioequivalency testing – a need to rethink. *Biometrics*, **37**, 589-594.
- ⁶Schuurmann, D.J. (1981). On hypothesis testing to determine if the mean of a normal distribution is contained in a known interval. *Biometrics*, **37**, 617.
- ⁷Ennis, D.M. and Ennis, J.M. (2008). Hypothesis testing for equivalence based on symmetric open intervals. *Communications in Statistics*, in press.
- ⁸Ennis, D.M. (2008). Tables for parity testing. *Journal of Sensory Studies*, **23**, 80-91.