Equivalence Testing in Demonstrating Substantial Equivalence for New Tobacco Products

Kimberly Frost-Pineda[1], Leanne R. Campbell[1], Michael Polster[2], Geoffrey M. Curtin[1]
[1]RAI Services Company, Winston Salem, NC; [2]Naxion, Philadelphia, PA

Abstract

Under Section 905(j) of the Tobacco Control Act, manufacturers must submit reports to US Food and Drug Administration (FDA) to demonstrate that a new tobacco product is substantially equivalent (SE) to a predicate product. Demonstrating equivalence with traditional statistical significance testing is, however, challenging for two reasons. First, the tests are specifically designed to demonstrate that two samples are different (i.e., to reject the null hypothesis that the samples are the same), and additional information may be necessary to demonstrate substantial equivalence when the null hypothesis cannot be rejected. Second, with very large samples, small differences can rise to statistical significance, but not actually be meaningful. As a result, there may be value in using equivalence testing, which examines whether or not the difference between means is smaller than a smallest effect size of interest by testing two null hypotheses (i.e., larger than the upper bound and smaller than the lower bound). If both hypotheses are rejected, the two means are deemed to be equivalent. Here we report on results of consumer testing in which likelihood of using a tobacco product was rated following presentation of an image of the product in a large sample (n=4,720) of current, former, and never tobacco users. The findings demonstrate that traditional statistical significance t-tests and equivalence tests often yield similar results (i.e., significant differences are not equivalent), but that the two types of tests yield divergent results when: (a) sample sizes are very large and effect sizes are very small (i.e., significant differences that are statistically equivalent); or (b) sample sizes are small and effect sizes are large (i.e., differences are neither significant nor statistically equivalent). Consideration is given to how to interpret discrepant findings in the context of consumer testing for SE product applications.

Introduction

Section 905(j) of the Tobacco Control Act requires manufacturers to demonstrate that a new tobacco product is SE to a predicate product. It is, of course, challenging to demonstrate equivalence because traditional statistical tests are designed to identify *differences* (i.e., to reject the null hypothesis that the samples are the same). Moreover, with large samples, it is possible for small differences to be statistically significant, but not actually be meaningful. There may, therefore, be value in using equivalence testing, which examines whether or not the difference between means is smaller than a smallest effect size of interest by testing two null hypotheses (i.e., larger than the upper bound and smaller than the lower bound). If both hypotheses are rejected, the two means are deemed to be equivalent. This research compares the results of traditional significance testing and equivalence testing in the context of data collected for an SE application, and provides suggestions about how to interpret discrepant findings.

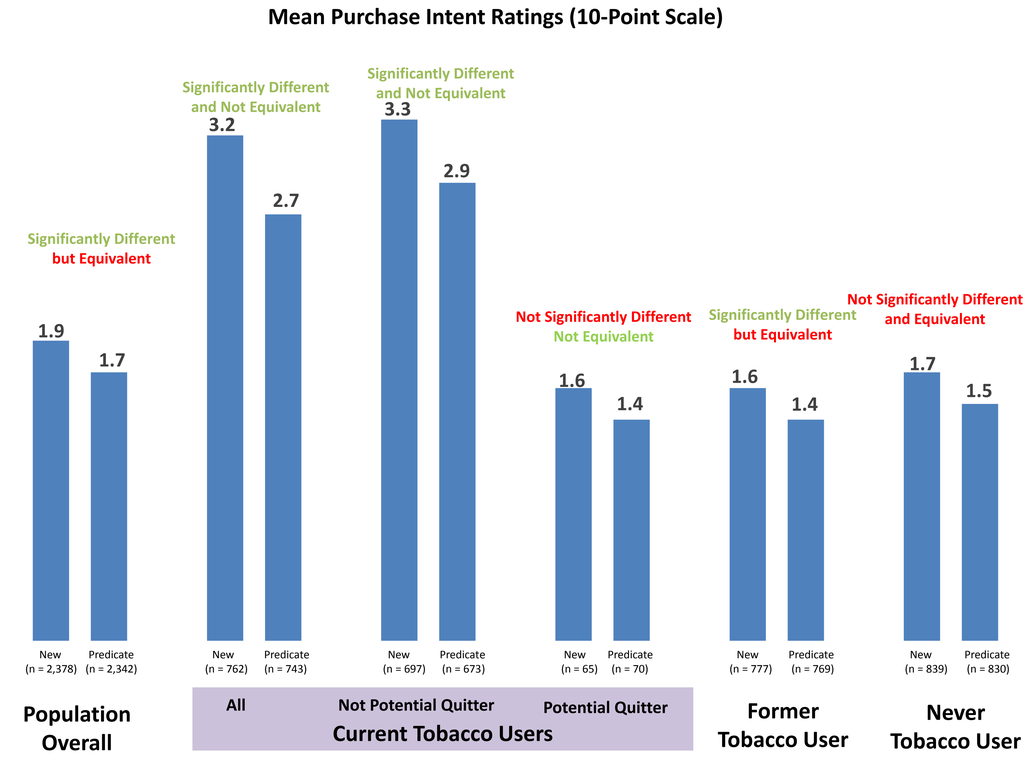
Methods

To provide a robust sample, 4,720 respondents were surveyed across three tobacco user groups (*i.e.,* current, former and never regular tobacco users). This sample size was selected to provide:

- balance on key demographic dimensions within each tobacco user quota group, allowing the sample to be weighted to population counts for all parameters of interest;
- the ability to perform statistical comparisons of new versus predicate products among consumers overall and within tobacco user groups with a high level of statistical sensitivity.

Respondents were shown an image for a tobacco product and then asked to rate the likelihood that they would purchase the product for personal use, on a 10-point scale. They also provided ratings of appeal and perceptions of risk, on 7-point scales. Ratings of purchase intent, appeal and risk perceptions for the population at-large, the three self-defined tobacco user groups (current, former, and never regular tobacco users), and pre-specified sub-groups of interest (e.g., current tobacco users who are and are not potential quitters) were then submitted to unpaired t-tests and equivalence tests using the "two-one sided tests" (TOST) procedure based on the sampling distribution (Schuirmann, 1987; Lakens, 2017). Equivalence testing determines if the difference between the means is smaller than the smallest effect size of interest (defined as Cohen's d = .2) by testing two null hypotheses – one that says the difference between the two products is larger than the upper bound, and one that says the difference between the two products is less than the lower bound.

Results



Results

The figure presents the mean purchase intent ratings for the new and predicate products for the population overall and key sub-groups of interest. As the graphic shows, traditional significance testing and equivalence testing yield consistent results for three of the six analyses. Means for the new and predicate products are significantly different and not equivalent for all current tobacco users and current tobacco users who are not potential quitters, and not significantly different and equivalent for never tobacco users. Results for the other three analyses are inconsistent: For the population overall and former tobacco users, the means are significantly different but equivalent (suggesting that the difference is unlikely to be meaningful), and for current tobacco users who are potential quitters, the means are neither significantly different nor equivalent (likely because of the relatively small unweighted sample size).

Conclusions

These findings illustrate that traditional statistical significance testing and equivalence tests often yield similar results (i.e., significant differences are not equivalent), but that the two types of tests can yield divergent results when: (a) sample sizes are very large and effect sizes are very small (i.e., significant differences that are statistically equivalent); or (b) sample sizes are small and effect sizes are large (i.e., differences are neither significant nor statistically equivalent). In the context of consumer testing for SE product applications, these findings highlight the relevance of sample size to the conclusions that are drawn: Small sample sizes may lead to erroneous conclusions of no difference, whereas large sample sizes may lead to erroneous conclusions that differences are meaningful solely because they are statistically significant. The findings also suggest that equivalence testing may offer a method to determine whether or not small statistically significant differences in mean values are meaningful. Further exploration of this topic is warranted because it could lead to an accepted methodology to demonstrate "substantial equivalence" to FDA.

References

Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J Pharmacokinet Biopharm. 1987 Dec;15(6):657-80.
Lakens D. Equivalence Tests: A practical primer for t tests, correlations, and meta-Analyses. Soc Psychol Personal Sci. 2017 May;8(4):355-362.

Presented at the 72nd Tobacco Science Research Conference, Memphis, TN. September 17, 2018. Poster 12.