HYPOTHESES OF EQUIVALENCE AND THEIR TESTING

Lawrence L. Garber Jr., Ünal Ö. Boya, and Eva M. Hyatt

Hypothesis of no difference are null hypotheses for studies to show that populations differ. To show that populations are essentially the same, the appropriate null is that substantial differences do exist. We propose that there is a pent-up conceptual need for equivalence hypothesizing in all of marketing— e.g, for the testing of core marketing concepts including the marketing concept, optimization of the marketing mix, product differentiation, market segmentation, the building of brand loyalty, product positioning, test marketing, as well as marketing pedagogy. We present two statistical tests appropriate for Equivalence Hypothesis Testing (EHT). Usefulness of the method to marketing is discussed.

It is the convention in many disciplines that the only proper null hypothesis formulation is one that predicts no significant difference between populations (Dolado, Otero and Brown 2014; Tryon 2001). Logic and the precepts of the scientific method would seemingly support the notion that there could also be times when one would observe and therefore would want to predict that some similarity exists between populations (Barker et al. 2001; Castelloe and Tobias 2006). That is to say, there would be times when primary interest would rest in verifying rather than rejecting the null hypothesis (Wellek 2003). In which case, the appropriate null hypothesis formulation would actually be the converse of standard practice, namely, the prediction of substantial differences. Why, then, is it deemed proper in many disciplines to only hypothesize differences as the alternative hypothesis (Dolado, Otero and Harman 2014)? Why can there be no hypotheses of equivalence (Ennis and Ennis 2010), by which one is seeking to prove that two treatments are essentially the same, that any difference is of no practical consequence (Motulsky 2007)?

Though equivalence testing was originally largely confined to pharmacological studies, there are those who are discovering that it has wider application, and who are currently introducing it into other disciplines—as, for example, Dolado, Otero, and Harman (2014) have recently done for the empirical software engineering field, Ennis and Ennis (2010) for the food sciences, Robinson and Froese (2004) for forestry, Seaman and Serlin (1998) for psychology, Rogers, Howard and Vessey (1993) for behavioral and educational research, and Barker et al. (2001) for statistical areas beyond biostatistics. Meyners (2012) reports it entering the sensory literature in 2008.

Within business and marketing, issues of equivalence have been raised since the 1970s with respect to making cross-cultural comparisons. However, to our knowledge these inquiries have largely concerned qualitative marketing research, or have dealt with finding methods for establishing construct (Green and White 1976; van Raaij 1978), data (Reynolds, Simitiras and Diamantopoulos 2003; Salzberger and Sinkovics 2006; van Herk, Poortinga and Verhallen 2005), functional (Green and Alden 1988; Green and White 1976), linguistic (Bhalla and Lin 1987), and measurement equivalences, (Dadzie, et al. 2002; Myers et al. 2000) among others, to enable valid cross-cultural comparisons. For a review, see Polsa (2007).

In this article, it is our purpose to introduce equivalence hypotheses and Equivalence Hypothesis Testing (EHT) to the business disciplines, specifically marketing. We propose that there is a pent-up conceptual need for equivalence hypothesizing in other parts of the marketing discipline—for example, for the testing of core marketing concepts, including the optimization of the marketing mix, product differentiation, market segmentation, the building of brand loyalty, product positioning, test marketing, and as well for marketing pedagogy.

Given that testing for equivalence is not simply the opposite of testing for significant differences, we go on to present two statistical tests appropriate for EHT. Usefulness of the method for marketing is discussed.

Lawrence L. Garber, Jr., Associate Professor of Marketing, Elon University, Elon, North Carolina, lgarber@elon.edu

Ünal Ö. Boya, Professor of Marketing, Appalachian State University, Boone, North Carolina, boyauo@appstate.edu

Eva M. Hyatt, Professor of Marketing, Appalachian State University, Boone, North Carolina, hyattem@appstate.edu

Journal of Marketing Theory and Practice, vol. 26, no. 3 (Summer 2018), pp. 280–288.

ISSN: 1069–6679 (print) / ISSN 1944–7175 (online)
DOI: https://doi.org/10.1080/10696679.2018.1450634

HISTORY OF EQUIVALENCE TESTING

There would appear to be historical and theoretical bases for this bias toward a primary interest in predicting differences rather than equivalences, stemming from Fisher's (1995) 1925 introduction of null hypothesis statistical testing (NHST, referred to henceforth as Fisher's approach) (Tryon 2001), and from the early work of Neyman and Pearson (1933) (Dolado, Otero and Harman 2014)—a bias which has apparently been controversial almost from its beginning (Meyners 2012; Pearce 1992). In spite of its controversy, Fisher's approach has become the basis for statistical analysis in social science research (Rogers, Howard and Vessey 1993; Tryon, 2001). The irony of this bias toward Fisher's approach, which Dolado, Otero and Harman (2014, p. 216) characterize as having been "... followed recently with increasing frequency, and perhaps with a certain degree of ritual," is cannily pointed out by Wellek (2003, p. xi) as follows:

> "A particularly striking phenomenon which demonstrates the real need for such methods [i.e., Equivalence Hypothesis Testing—EHT] is the adherence of generations of authors to using the term 'goodness-of-fit' tests for methods which are actually tailored for solving the reverse problem of establishing absence or lack of fit."

Welleck (2003, p, xi) goes on to briefly recount these origins as follows:

> "From a 'historical' perspective (the first journal article on an equivalence test appeared as late as in the sixties of the twentieth century), the interest of statistical researchers in equivalence assessment was almost exclusively triggered by the introduction of special approved regulations for so-called generic drugs by the Food and Drug Administration (FDA) of the U.S. as well as the drug regulation authorities of many other industrialized countries. Essentially, these regulations provide that the positive result of a test, which enables one to demonstrate with the data obtained from a so-called comparative bioavailability trial the equivalence of the new generic version of a drug to the primary manufacturer's formulation, shall be accepted as a sufficient condition for approval of the generic formulation to the market."

As Ennis and Ennis (2010) have found for the food sciences, we find that the range of applications of EHT for business and marketing is quite broad. Examples may include all aspects of the marketing mix—for instance, with respect to product, one might hypothesize that Kraft's new, healthier formulation for macaroni and cheese looks and tastes the same as the original, that "dropped call rates are equivalent among cell phone providers, that an artificial sweetener is equivalent to a natural sweetener, that one tooth whitening product is as effective as another" (Ennis and Ennis 2010, p. 253). With respect to price, applications of EHT might be that the performance of lower-priced generic products are equal to those of premium brands, or that small discounts achieve equal response to large discounts. With respect to place, applications of EHT might be whether one channel of distribution is equal to another, or whether sales territories are equal. For promotion, it might be that a ten second commercial is as effective as a fifteen second commercial, or that a Facebook page is as effective as a website, or that a free carafe of wine to encourage patrons to stay longer for a restaurant dinner has the same effect whether it is red or white (Ennis and Ennis 2010).

Bioequivalence studies are deemed particularly useful when an untreated control would be considered unethical (Gøtzsche 2006). There are analogous circumstances in marketing—when an untreated control is infeasible for reasons of equity, or unavailable within natural experiments or quasi-experimental studies. Such can be the case when, for example, a firm is loath to manipulate compensation packages across sales reps, stop advertising altogether for several years, post a blank website or empty Facebook page, manipulate prices in circumstances where price discrimination would be seen as clearly in violation of the Robinson-Patman Act or simply seen as unfair, or manipulate classroom learning tools across sections of a principles of marketing class to see which are more effective.

Another area for which the logic of equivalence testing would seem better suited than Fisher's approach is model validation, "... central to the application of models to scientific and managerial problems ... [whose intent is] to seek to match a sample of observations from some target population against a sample of predictions taken from a model" (Robinson and Froese 2004, p. 349). When primary interest is in finding equivalence between these two samples, rather than difference, the latter indicating that the model is invalid, equivalence testing would therefore be the more appropriate approach (Robinson and Froese 2004).

And another potential avenue for the efficacy of equivalence testing, vis-à-vis its usefulness in

circumstances where there is no untreated control group, is as a variation to A/B testing (Kohavi and Longbotham 2015). Commonly used in online settings such as web design to identify changes to web pages that increase or maximize an outcome of interest (e.g., click-through rate), A/B testing is a term for a randomized experiment with two treatments, A and B, which are the control (e.g., a currently used version of a website) and an alternative that is a variation on that which is the control (the treatment). However, there may be circumstances where primary interest may be in determining whether some variable is equivalent in effect to some original variable.

Looking at the above examples, particularly with respect to A/B testing, it may at first appear that the distinction between Fisher's approach and EHT may be conceptual only, that the results of statistical testing from one of these would only mirror the other. But, this turns out not to be the case. Testing for equivalence is not simply the opposite of testing for significant differences—that is, it is simply untrue that the answer to the question "Are the samples significantly different?" is the same as the answer to the question "Are the samples significantly similar?" (Castelloe and Tobias 2006). If Fisher's approach reaches the conclusion that a difference is not statistically significant, one has no information to indicate that the difference is zero. Therefore, one cannot conclude that the two treatments are functionally equivalent (Motulsky 2007). (Neyman-Pearson's approach will allow for concluding that two treatments are functionally equivalent, but only if the test has reasonably high power.)

TESTING FOR EQUIVALENCE

This article advocates for the use of equivalence hypotheses in the business and marketing disciplines. In the following sections we present two of the more popular tests for statistical tests of equivalence (Ennis and Ennis 2010): testing for equivalence with confidence intervals (CI), and the two one-sided *t* tests (TOST). There are others (for reviews, see Barker et al. 2001; Kruschke 2013; Meyners 2012; Wellek 2003), but many turn out to be variations on these two, or more complicated when "the simpler methods will usually do as good a job as the more demanding ones" (Meyners 2012, p. 243).

Testing for Equivalence Using Confidence Intervals (CI)

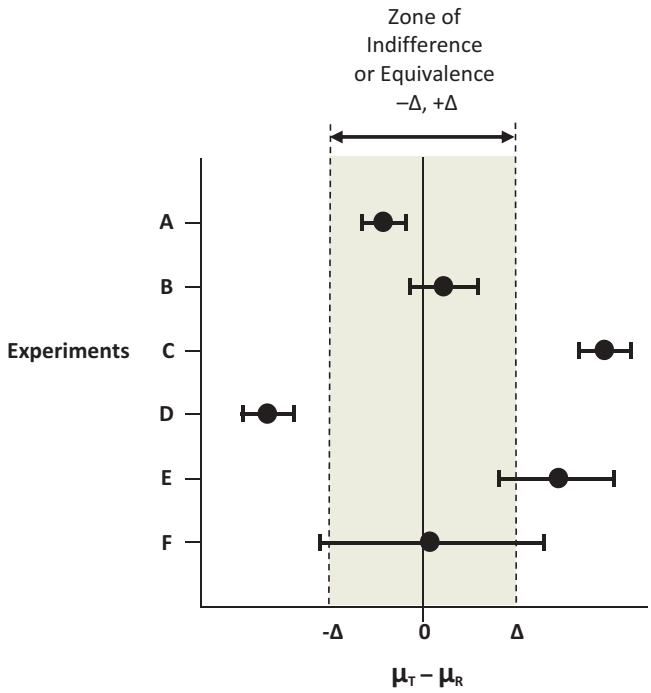
Tests beyond Fisher's approach are necessary for equivalence because it cannot accept the null value, only reject it (Kruschke 2013). But an issue that comes with accepting the null value is that there will always be some difference in outcomes between two treatments. So, the question with equivalence testing is not whether two outcomes are identical, but whether the differences in outcome are sufficiently small to be considered scientifically inconsequential (Robinson and Froese 2004). Therefore, to ask questions about equivalence one must first define a range of treatment effects that one considers to be scientifically (*not* statistically) equivalent (Motulksy 2007). The boundary that divides inconsequential treatment differences from the consequential is commonly denoted as Δ (Wellek 2010), such that, for two treatments T and R, the null hypothesis is

$$H_0: \ |\mu_T - \mu_R| \geq \Delta$$

Following Motulsky (2007), Figure 1 shows the confidence intervals resulting from six contrived experiments, A to F, hypothesizing equivalence. The result of each experiment (i.e., the treatment effect) is indicated by a dot showing the mean difference between treatments T and R with a 95 percent confidence interval bracketing each mean difference. For a resulting mean difference between treatments T and R to be declared equivalent, the entire range of the confidence interval surrounding it must fall completely within the "Zone of Indifference," or equivalence, indicated by the gray area shown in Figure 1. Therefore, the results of experiments A and B show support for the contention that treatments T and R are equivalent. The results of experiments C, D, E, and F show support for the contention that treatments T and R are not equivalent, since the range of their confidence intervals is not completely within the Zone of Indifference.

As can be seen from this example, EHT allows us to explore results in greater detail. It gives us additional insights into the effect size and the range of plausible alternative effects informed by a confidence interval, by examining the equivalence intervals (Dolado, Otero and Harman 2014).

Figure1
Testing for Equivalence with Confidence Intervals



Testing for Equivalence Using the Two One-Sided t Tests (TOST)

The best known and most commonly applied procedure is the TOST—the two one-sided *t* tests—attributable to Schuirmann (1987) and found on many standard statistical packages. Following Dolado, Otero and Harman (2014), it decomposes H_o (i.e, $|\mu_T - \mu_R| \geq \Delta$) into two separate hypotheses and applies *t* tests to each of them individually. That is,

$$H_{01} : \mu_T - \mu_R \leq -\Delta \text{ and}$$

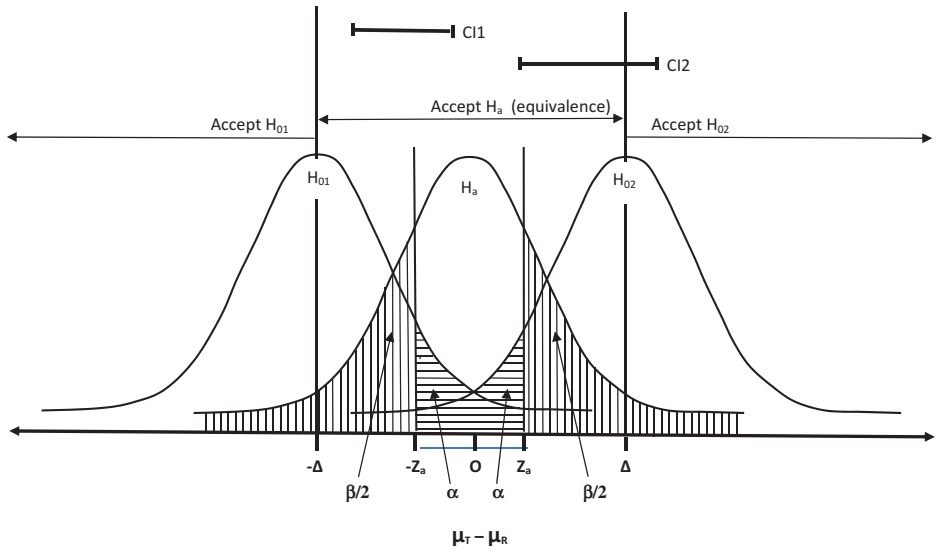
$$H_{02} : \mu_T - \mu_R \geq +\Delta$$

Test 1 seeks to reject H_{01}, asserting that the difference between two means is less than or equal to –Δ. Test 2 seeks to reject H_{02}, asserting that the difference between two means is greater than or equal to Δ. The rejection of both H_{01} and H_{02} infers equivalence (Dolado, Otero and Harman 2014). Figure 2 shows graphically the two null hypotheses H_{01} and H_{02} in TOST versus the alternative Ha (Ha1 and Ha2 together). The logic behind the test is that if μ_T–μ_R is shown to have come from a distribution simultaneously to the right of -Δ and to the left of Δ, the investigator can conclude that the distribution it came from is somewhat in the middle, with the true difference μ_T–μ_R less than the minimum difference of importance determined a priori by the investigator (Rogers, Howard and Vessey 1993).

MARKETING EXAMPLES

In this section, we show the application of EHT to two problems in marketing. For the first, EHT is applied to experimental data, for which EHT was originally intended. For the second, we extend EHT to survey data, common in marketing research, for which EHT is also well suited. Both examples show results from both CI and the TOST.

Figure2
Plot of the TOST and the Confidence Intervals CI1 and CI2 for the Difference of Means



Example 1: EHT Applied to the Size Appearance of Competitor Brands and Their Comparison

In an experiment, Garber, Hyatt and Boya (2014) examined the effect of package shape on size appearance, and found that simple package forms appear larger than complex package forms that are the same size. In empirically demonstrating these results, the authors observed a related phenomenon. Since it was hard to find differently shaped packages that were exactly the same size, the authors instead tested a set of sixteen packages within a certain size range—a range where differences in size among some packages were apparent, but where it was at times also hard to determine which was the larger of subsets of these packages. And, as well, a range of sizes that are not atypical of those found among competitor brands within a category sitting on store shelves.

Upon observing the efforts of test subjects to estimate the sizes of these packages, the authors made a new observation that was corollary to the original. That, within that sort of size range designed into this experiment, the apparent sizes of smaller simple packages would tend to converge on the apparent sizes of larger complex packages—an observation potentially of strategic importance to marketers wanting to present smaller packages that appear as large as larger competitor packages.

The authors could then formulate the following equivalence hypothesis:

> H_a: *Smaller simple packages appear to be the same size as larger complex packages when seen on store shelves, within those size ranges normally found among competitor brands within a category.*

Therefore, the null hypothesis becomes:

> H_o: *Smaller simple packages appear to be a different size than larger complex packages when seen on store shelves, within those size ranges normally found among competitor brands within a category.*

To test H_a, we first apply Schuirman's (1987) two one-sided t tests (TOST) using the TOST statement in SAS PROC TTEST (2015). Specifically, we are testing the equivalence of the size appearance of smaller simple packages to larger complex competitor packages. A part of this method is to a priori choose a criterion for determining equivalence according to the practical circumstances within which it is

Table 1
TOST Results for Example 1: The Size Appearance of Smaller Simple vs. Larger Complex Packages
TOST Level 0.05 Equivalence Analysis *
Variable: Size Appearance in Milliliters

Size and Shape	N	Mean	Std Dev	Std Err	95% CL Mean		95% CL Std Dev		Null Bounds ± Δ	90% CL Mean		Overall P-Value	Assessment
0—Smaller Simple Package Forms	395	332.3	111.9	5.63	321.2	343.4	104.6	120.3					
1—Larger Complex Package Forms	395	367.3	99.3	5.00	357.5	377.1	92.8	106.7					
Diff (0-1)		-35.0	105.8	7.5	-49.8	-20.2	100.8	111.3	± 50	-47.4	-22.6	.0231	Equivalent

***Notes*: ***This output summarizes that provided by the TOST function in SAS PROC TTEST (2015). It also shows results for the CI, in the "90% CL Mean" columns of each table. The results for the TOST and the CI are in agreement in these two examples, though they may not be in other cases.

applied, and what minimum level of difference is meaningful in that context (Castellos and Tobias 2006). The authors' criterion for equivalence in this case is 50 mls, which is roughly 20 percent of the full range of sizes represented by the set of sixteen packages in this experiment (236 mls to 467 mls), a range of 231 mls.

The results of the EHT summarized in Table 1 indicate equivalence between the mean size appearance of small simple packages (332.3 mls, whose actual mean volume is 284 mls) and large complex packages (367.3 mls, whose actual volume is 391.8 mls), whose mean difference is 35.0 mls (whose actual mean difference is 107.8 mls). TOST Level 0.05 Equivalence is indicated by an Overall P-Value of .0231 < α = .05. (Since the TOST consists of two statistical tests applied simultaneously, a multiple comparison correction needs to be made to control the familywise error rate. A Bonferroni correction is a general method that can be used by which the alpha value for each test is lowered to account for the number of comparisons being performed. With such a correction the alpha value for each individual test is calculated by dividing the overall alpha level by the number of tests being made, in this case two, to account for spurious positives. Thus, a TOST alpha of .05 equates to .025 for each test.)

Expressing the alternative hypothesis as an equivalence is not only conceptually in line with the observations of the investigators of this phenomenon, but the EHT that is reported (the TOST in this case) offers advantages over Fisher's approach in terms of its specificity. Additional insights into the effect size, and the confidence with which we may assert differences over ranges of effect size, can be obtained by examining the equivalence intervals across the entire distribution of size estimates shown there, by seeing where and by how much they depart from the normal distribution (Dolado, Otero and Harman 2014).

It happens that, in these two examples, results for CI and the TOST agree, finding for equivalence in all cases. But, by Meyners (2012), these two equivalence tests need not always be in agreement. Thus, we recommend that it would be good practice to apply both tests and report both results for the readers' interpretation.

Example 2: EHT Applied to the Learning Preferences of Female and Male Marketing Game Participants

Several empirical studies have utilized Kolb's Experiential Learning Theory (ELT) to compare the respective learning styles of women and men. Kolb's ELT hypothesizes two dimensions to learning, a grasping experience and a transforming experience. Empirical studies have consistently shown that female versus male learners have differing learning preferences on the grasping dimension, but show no significantly different learning preferences on the transforming dimension. However, none of those studies has utilized EHT to test whether female and male learners are actually equivalent on the transforming dimension. We do so here.

For this example, we again first apply Schuirman's (1987) two one-sided t test (TOST) using the TOST statement in SAS PROC TTEST (2015), in this case to a set of survey data. Our criterion for equivalence (Castellos and Tobias 2006) is that the difference in female and male mean responses for Kolb's (1976) grasping dimension be no greater than two levels out

Table 2
TOST Results for Example 2: The Learning Preferences of Female vs. Male Game Participants
TOST Level 0.05 Equivalence Analysis *
Variable: Scores on Kolb's (1976) Transforming Dimension

Size and Shape	N	Mean	Std Dev	Std Err	95% CL Mean		95% CL Std Dev		Null Bounds ± Δ	90% CL Mean		Overall P-alue	Assessment
0 – Male Game Participants	147	1.88	5.32	0.44	1.02	2.75	4.77	6.01					
1 – Female Game Participants	71	1.51	5.32	0.63	0.25	2.77	4.56	6.27					
Diff (0-1)		0.38	5.32	0.77	-1.14	1.89	4.86	5.87	± 2	-0.89	1.65	.0180	Equivalent

Notes:*This output summarizes that provided by the TOST function in SAS PROC TTEST (2015). It also shows results for the CI, in the "90 percent CL Mean" columns of each table. The results for the TOST and the CI are in agreement in these two examples, though they may not be in other cases.

of a full range of thirty-six levels from Kolb's (1976) Learning Styles Inventory (LSI), or 5 percent of the total range. Findings from the TOST, shown in Table 2, are significant ($p = .0328$, $\alpha < .05$), offering support for our expectation (H_2) that female and male learners share preferences for the transforming experience. In contrast, by that same two level, 5 percent criterion, the TOST for the transforming dimension is not significant ($p = .8925$).

And we may again also find results for CI in the table. The column for the "90% CL Mean" in Table 2 indicates means of –0.89 and 1.65, both of which fall within the Null Bounds, or prespecified equivalence range of ± 2, confirming the TOST's assessment of equivalence.

MANAGERIAL IMPLICATIONS

In summary, equivalency testing is a straightforward process whose basic concepts are already familiar to empiricists. Type I error rates are controlled. There is a null hypothesis, asserting that the difference between groups is at least as large as the Δ specified by the investigator, and an alternative hypothesis, asserting that the difference between groups is smaller than the one that is specified. As with traditional hypothesis testing, the goal of the investigator is to reject the null (Fisher's approach).

However, we believe this exposition goes on to show that EHT answers equivalence questions that Fisher's approach cannot address. There are times, in empirical studies, when investigators would like to explore the confidence with which they can claim that two treatments are equivalent. A current managerial example is Kraft's recent marketing campaign, "It's changed, but it hasn't," for "Blue Box," its highly popular, and virtually ubiquitous, prepared Mac & Cheese, which comes in a signature blue box. Blue Box has a large and loyal following, not simply because it is inexpensive and easy to prepare, but many loyalists are fond of the food's highly identifiable shape, color, and flavor. They simply like "Blue Box," just as it is. It is their brand, and they want no changes, in spite of the fact that healthiness is not a key attribute of "Blue Box."

It is this strong affinity that Blue Box's following feels for it that created a dilemma for Kraft. Heeding the U.S. trend toward eating healthier, Kraft believed that they could attach new "healthy" food segments if they upgraded Blue Box on this attribute, by removing artificial flavors, preservatives, and dyes. But, they wanted to do so by also retaining their traditional following, who wanted Blue Box to stay the same. So, Kraft's strategy was to find ways to remove artificial flavors, preservatives, and dyes without changing Blue Box performance in terms of flavor, shape, and appearance, including color. Ideally, Kraft wanted to effect these changes without its traditional loyalists even knowing, not being able to tell that Blue Box was changed in any way. Kraft accomplished their technical goal, improving the healthy profile of Blue Box while in no way affecting flavor, shape or color, and proceeded to demonstrate this fact to the marketplace with an unusual marketing strategy. Kraft withheld news of the changes it made to Blue Box from the marketplace for six months. Once Kraft had evidence directly from the consumer that the consumer noticed no changes to Blue Box, it then announced its changes

with proof from the consumer that it was still the Blue Box that all loyalists knew and loved.

Blue Box is an example of how improvements to certain of a brand's attributes may reduce perceptions of others that are not in fact changing, a good example of how framing a research questions in terms of equivalence would be conceptually correct. Because the questions managers often ask are of the form, "If we make this change, will the consumer still perceive ... ?" There are other examples for which the managerial question would also refer to equivalence.

When Pepsi switched its package colors from red, white and blue to principally blue, Pepsi risked reducing its identifiability, brand perceptions, and liking. Testing for equivalence would be prudent in such a case. Consider the cases of Crystal Pepsi and New Coke; they both failed. Perhaps these launches would not have happened, or been modified, if Pepsi and Coke had tested to assure that the perceived qualities of Coke and Pepsi that they wanted to be unchanging were equivalent.

When marketers extend their product lines, they risk diluting the equity of anchor brands. Testing for the equivalence of the perceptions of an anchor brand before and after the launch of an extensions would determine these effects. A similar test would be prudent for brands being differentiated from others in new ways: is obtaining distinctiveness on some dimension going to cause consumers to question delivery of other perceived benefits that the marketer does not want to change? Testing for equivalence could be beneficial in this case. It makes sense to test the equivalence of store brands to premium brands in terms of quality, for example.

Will some newly discovered market segment receive a given brand as well as some established segment? Will some new segment perceive the position of a brand in the marketplace in the same way as others? On some dimensions, but not others?

Will some revised combination of price and quality equal the perceived value of some earlier version of a brand? If we reduce the size of a candy bar, or the amount of cereal in a box, which the producers hope will go unnoticed, will it? Ethical questions aside, these are questions whose testing as equivalence hypotheses are appropriate. More generally, and more properly, the question of equivalence is raised whenever marketers want to make product or marketing mix changes that fall below the level of just-noticeable-differences, which is often the case with package and product updates.

Are perceptions of brand quality retained when distributed via a big box store rather than a boutique? When it is priced lower at the big box store? When it is shopped online? When we double the size of the shelf facings, or reduce them? Put it one shelf down? Show it in groups of three rather than four in specialty displays?

And so on.

In situations like these, it may be tempting for the investigator to use Fisher's approach to explore equivalence. However, once again, it is important to know that failing to reject the null hypothesis in Fisher's approach is not the same as demonstrating equivalence (Dorato, Otero and Harman 2014). Absence of proof is simply not the same as proof of absence (Motulsky 2007).

REFERENCES

Barker, Lawrence, Henry Rolka, Deborah Rolka and Cedric Brown (2001), "Equivalence Testing for Binomial Random Variables: Which Test to Use?" *American Statistician*, 55 (4), 279–87.

Bhalla, Gaurav and Lynn Y.S. Lin (1987), "Cross-Cultural Marketing Research: A Discussion of Equivalence Issues and Measurement Strategies," *Psychology & Marketing*, 4 (4), 275–82.

Castelloe, John and Randy Tobias (2006), "Like Wine, the TTEST Procedure Improves with Age," *SUGI 31 Proceedings*, 208–31.

Dadzie, Kofie Q., Wesley J. Johnston, Boonghee Yoo and Thomas Brashear (2002), "Measurement Equivalence and Applicability of Core Marketing Concepts Across Nigerian, Kenyan, Japanese and US Firms," *Journal of Business and Industrial Marketing*, 17 (6), 430–54.

Dolado, José Javier, Mari Carmen Otero and Mark Harman (2014), "Equivalence Hypothesis Testing in Experimental Software Engineering," *Software Quality Journal*, 22 (2), 215–38.

Ennis, Daniel M. and John M. Ennis (2010), "Equivalence Hypothesis Testing," *Food Quality and Preference*, 21 (3), 253–56.

Fisher, R.A. (1995), *Statistical Methods, Experimental Design, and Scientific Inference*, Oxford, UK: Oxford University Press.

Garber, Lawrence L., Jr., Eva M. Hyatt, and Ünal Ö. Boya (2014), "The Effects of Package Shape and Presentation Context on Volume Appearance: An Empirical Investigation," *International Journal of Management Practice*, 7 (2), 144–59.

Green, Robert T., and Dana L. Alden, (1988), "Functional Equivalence in Cross-Cultural Consumer Behavior: Gift Giving in Japan and the United States," *Psychology and Marketing*, 5 (2), 155–68.

————, and Dana L. Alden-, and Phillip D. White (1976) "Methodological Considerations in Cross-national

Consumer Research," *Journal of International Business Studies*, 7 (2), 81–87.
Gøtzsche, Peter C. (2006), "Lessons from and Cautions about Noninferiority and Equivalence Randomized Trials," *JAMA*, 295 (10), 1172–74.
Kohavi, Ron and Roger Longbotham (2015), "Online Controlled Experiments and A/B Tests," in *Encyclopedia of Machine Learning and Data Mining*, Claude Sammut and Geoff Webb, eds. New York: Springer Publishing,
Kruschke, John K. (2013), "Bayesian Estimation Supersedes the *t* Test," *Journal of Experimental Psychology, General*, 142 (2), 573–603.
Meyners, Michael (2012), "Equivalence Tests— A Review," *Food Quality and Preference*, 26 (2), 231–45.
Motulsky, Harvey J. (2007), *Prism 5 Statistics Guide*, San Diego, CA: GraphPad Software, Inc.
Myers, Matthew B., Roger J. Calantone, Thomas J. Page. Jr. and Charles R. Taylor (2000), "Academic Insights: An Application of Multiple-Group Causal Models in Assessing Cross-Cultural Measurement Equivalence," *Journal of International Marketing*, 8 (4), 108–21.
Neyman, Jerzy and Ergon S. Pearson (1933), "On the Problem of Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society of London, Series A, Containing Papers of a Mathematical or Physical Character*, 231, 289–337.
Pearce, S.C. (1992), "Introduction to Fisher (1925) Statistical Methods for Research Workers," In *Breakthroughs in Statistics, Vol. II: Methodology and Distributions*. S. Kotz and N.L. Johnson, eds., New York: Springer-Verlag, 59–65.
Polsa, Pia (2007), "Comparability in Cross-Cultural Qualitative Marketing Research: Equivalence in Personal Interviews," *Academy of Marketing Science Review*, 8 (1), 1–18.
Reynolds, N. L., A.C. Simintiras, and A. Diamantopoulos (2003), "'"Theoretical Justification of Sampling Choices in International Marketing Research: Key Issues and Guidelines for Researchers," *Journal of International Business Studies*, 34, 80–89.
Robinson, Andrew P. and Robert E. Froese (2004), "Model Validation Using Equivalence Tests," *Ecological Modelling*, 176, 349–58.
Rogers, James L., Kenneth I. Howard, and John T. Vessey (1993), "Using Significance Tests to Evaluate Equivalence between Two Experimental Groups," *Psychological Bulletin*, 113 (3), 553–65.
Salzberger, Thomas and Rudolf R. Sinkovics (2006), "Reconsidering the Problem of Data Equivalence in International Marketing Research," *International Marketing Review*, 23, (4) 390–417.
Schuirmann, Donald J. (1987), "A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability," *Journal of Pharmacokinetics and Biopharmaceutics*, 15 (6), 657–80.
Seaman, Michael A. and Ronald C. Serlin (1998), "Equivalence Confidence Intervals for Two Group Comparisons of Means," *Psychological Methods*, 3 (4), 403–11.
Tryon, Warren 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 (4), 371–86.
Van Herk, Hester, Ype H. Poortenga and Theo M.M. Verhallen (2005), "Equivalence of Survey Data: Relevance for International Marketing," *European Journal of Marketing*, 39 (3/4), 351–64.
Van Raaij, W. Fred (1978), "Cross-Cultural Research Methodology as a Case of Construct Validity," *Advances in Consumer Research*, 5, 693–701.
Wellek, Stefan (2003), *Testing Statistical Hypotheses of Equivalence*, New York: Chapman & Hall/CRC.