The Institute for Perception

For technical reports from The Institute for Perception or for information about short courses, please visit www.ifpress.com or email us at mail@ifpress.com

Reprinted from IFPress (2011) 14(4) 3,4

Interpreting Applicability Scores

Daniel M. Ennis and John M. Ennis

Background: "Check-all-that-apply" (CATA) lists are a popular tool in product tests[1,2,3]. In a typical test, consumers respond to a series of statements and mark those statements that apply to the product of interest. An advantage of CATA testing is that it provides an opportunity to obtain information from consumers that would be difficult in some cases to extract using either a rating or 2-AFC format. A related method, explored by Loh and Ennis[4] in 1982, is called applicability scoring. In applicability scoring, consumers mark statements that are applicable but also mark statements that are not applicable. In a CATA list, an unmarked item may imply that the consumer does not think that the item applies, but could also mean that the consumer merely missed that item – applicability scoring avoids this ambiguity.

In this report the topic of how to analyze and interpret applicability scores will be discussed. This report will provide guidance on the analysis of applicability counts to test a null hypothesis of no difference and will also discuss the scaling of applicability data using a Thurstonian model. One application of particular interest will be the comparative evaluation of two products on liking.

Read the phrases and, for each phrase, mark the box to the left if the phrase describes the product tested. Mark the box on the right if the phrase does not describe the product tested.

	Does Apply	Does Not Apply
I like this product	☐	☐
Has a lasting aftertaste	☐	☐
Causes salivation	☐	☐
Tastes spicy	☐	☐
Tastes very salty	☐	☐

Fiqure 1. Sample items from a typical applicability questionnaire.

Scenario: In a consumer test you have obtained applicability data in a sequential design on two seasoned pretzels in a blind format. Figure 1 shows a sample of the applicability items. The applicability variables were randomized for each respondent using computerized data entry and the products were tested in a balanced order of presentation. The consumers were recruited from an homogenous group of users regarding taste preferences and are loyal users of your main brand. The two products consist of your main brand and a low calorie prototype repacked in a blind format.

McNemar's Test: In Table 1, there are four cells and only two of them reflect differences between the products. These cells are the off-diagonal elements that count the number of consumers who found that the item "I like this product" applied to one product but not the other. If the products were identical the expected value of these cells would be equal. A chi-square test with one degree of freedom using these cells provides a basis for testing a null hypothesis of no difference. This test is conducted by checking whether $(n_{12}-n_{21})^2/(n_{12}+n_{21})$ is greater than 3.84 – the cutoff for a central chi-square with one degree of freedom at $\alpha = 0.05$. Note that since different respondents populate all four of the cells in Table 1, the independence assumption made by McNemar's test is satisfied.

		Low Calorie Prototype: DOES APPLY	Low Calorie Prototype: DOES NOT APPLY	
Main Brand	DOES APPLY	88 n_{11}	104 n_{12}	192
Main Brand	DOES NOT APPLY	50 n_{21}	58 n_{22}	108
		138	162	300

Table 1. Applicability counts for the main brand and the low calorie prototype based on the statement "I like this product."

Loh and Ennis[4] compared McNemar's test on applicability data with binomial tests on a forced choice preference question using the same pair of products tested separately. They found that McNemar's test on the off-diagonal of the applicability matrix was at least as sensitive in finding differences as a binomial test on the preference data. In addition, they used factor analysis to show that the applicability items provided a richer source of information than 7-point rating scales on similar attributes.

Applies to Both/Neither: In a typical paired test with a no difference/preference option[5], we obtain preference counts for each product as well as a count of the no differences/preferences. However, we do not know from the preference counts if the consumers are satisfied with either product. Applicability scores provide this additional information. For example, you know from Table 1 that 88 people liked both products and 58 liked neither of them. If this latter count had been relatively much larger than the former you would have inferred that whatever preferences existed, neither of the products were generally acceptable. Further research to account for scoring bias could then have been conducted with identical products among loyal users to substantiate this conclusion in this case.

It is important to note that consumers falling into the like both/like neither categories may still prefer one product over the other. In the case of the "like both" category, they may not dislike the less preferred product enough to

provide a "does not apply" score to that product on liking. The purpose of the next section on Thurstonian scaling is to estimate where that cut-off lies.

Thurstonian Scaling: In a Thurstonian model, we assume that each percept is randomly drawn from a distribution with a certain mean and unit variance. One product is assumed to have a mean of zero and the other a mean of δ. A criterion, c, is placed on the attribute axis. If the percept exceeds c, the subject responds "does apply," otherwise the subject responds "does not apply."

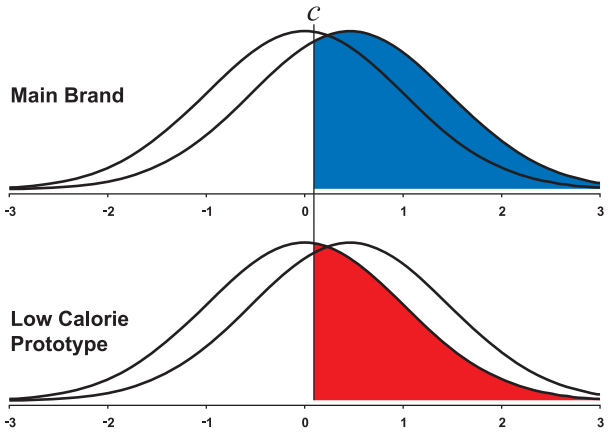

Figure 2. Underlying distributions that give rise to the four possible response patterns. Shaded areas correspond to the probability that a product is liked.

Figure 2 illustrates the distributions of the percepts for two products in the case of "I like this product." The criterion, c, that determines the point above which liking would be declared for either product. To obtain the probability of a joint event such as "I like Product A/I do not like Product B," we multiply the probabilities of the independent events "I like Product A" and "I do not like Product B." The probabilities of these individual events are found by computing the areas under the curves for the respective products that are either above or below the c value. For example, to find the probability that a respondent likes the main brand, we determine the area of the blue region shown in Figure 2. To find the probability that a respondent does not like the low calorie prototype, we compute one minus the area of the red region shown in Figure 2.

Interpretation of the Applicability Counts: Applying McNemar's test to the results in Table 1, you reject the null hypothesis that the products performed identically[6]. You then perform a Thurstonian analysis as described above using the *IFPrograms*™ software package. From this analysis you find that d', the estimate of δ, is 0.46. The criterion c is estimated to be 0.1. Table 2 shows the actual and predicted values for the data in Table 1 using the two-parameter Thurstonian model involving δ and c. In a similar manner, you perform this analysis for every attribute. If needed, this approach could be generalized to include more than two products.

Cell	Observed	Predicted
Applies to Main Brand but not Prototype	104	103.74
Applies to Prototype but not Main Brand	50	49.62
Applies to Both	88	88.43
Applies to Neither	58	58.2

Table 2. Actual applicability score frequencies for the item "I like this product" and predicted results from a Thurstonian model with $c = 0.1$ and $\delta = 0.46$.

Having completed your analysis, you can predict the results of other methods for which a Thurstonian model has been developed[7,8,9,10]. For instance, for $\delta = 0.46$, you predict a forced-choice preference result to be 63% in favor of the main brand. Using the Thurstonian approach, you can also investigate the power of your applicability scoring relative to other testing methods[11].

Conclusion: Applicability scores are frequently obtained in a number of consumer testing applications. For two products, hypothesis testing for data of this type can be accomplished using McNemar's test. These data are also valuable for scaling differences between products within a Thurstonian perspective.

References and Notes

1. Ares, G., Barreiro, C., Deliza, R., Gimenez, A., and Gámbaro, A. (2010). Application of a check-all-that-apply question to the development of chocolate milk desserts. *Journal of Sensory Studies,* **25**(s1), 67-86.
2. Dooley, L., Lee, Y., and Meullenet, J.F. (2010). The application of check-all-that-apply (CATA) consumer profiling to preference mapping of vanilla ice cream and its comparison to classical external preference mapping. *Food Quality and Preference,* **21**(4), 394-401.
3. Plaehn, D. (2011). CATA penalty/reward. *Food Quality and Preference,* **24**(1), 141-152.
4. Loh, C.F. and Ennis, D.M. (1982). Glossary of attributes. *Philip Morris Internal Report,* accession number 82-288. Retrieved from http://www.pmdocs.com/pdf/2001298236_8263_xhmfyd55uzdszmvz5bnqoqvt.pdf
5. Ennis, D.M. and Ennis, J.M. (2011). Accounting for no difference/preference responses or ties in choice experiments. *Food Quality and Preference*, **23**(1), 13-17.
6. $\chi^2_1 = (104\text{-}50)^2/154 = 18.94$, $p < 0.0001$
7. Thurstone, L.L. (1927). A law of comparative judgement. *Psychological Review,* **34**(4), 273-286.
8. Hacker, M. J. and Ratcliff, R. (1979). A revised table of d' for m-alternative forced choice. *Perception & Psychophysics,* **26**, 168-170.
9. Ennis, D.M. (1993). The power of sensory discrimination methods. *Journal of Sensory Studies,* **8**(4), 353-370.
10. Ennis, J.M., Ennis, D.M., Yip, D., and O'Mahony, M. (1998). Thurstonian models for variants of the method of tetrads. *British Journal of Mathematical and Statistical Psychology,* **51**, 205-215.
11. Ennis, J.M. and Jesionka, V. (2011). The power of sensory discrimination methods revisited. *Journal of Sensory Studies,* **26**, 371-382.