Reprinted from IFPress (2018) 21(3) 3,4

Derived Preference from Applicability Scoring

Daniel M. Ennis and Benoît Rousseau

Introduction: Traditional preference testing is expensive with more than two products because of the large number of comparisons that require evaluation. Applicability scoring[1,2] was originally used in product testing partly to develop an alternative method for deriving an analog preference measure in sequential monadic testing[3]. Since the method is used sequentially, it can be used for more than two products. When the attribute is *liking* (the item scored is *I like this product*), the method allows the separation of *like both* from *like neither* which is not provided using a preference question with a *no preference option*. Therefore, this capability provides more information about the acceptability of both products than can be obtained from a preference test.

Applicability scoring requires the consumer to indicate whether each term or statement 'applies' or 'does not apply' to the sample evaluated. When it was first used in the sensory field[3] and compared to traditional preference testing, it was found to be comparable in sensitivity.

In this technical report, we extend the learnings from our previous report[2] to illustrate how applicability scoring can be a viable alternative to traditional preference or other paired comparison tests. We will show that theoretically it exhibits a similar statistical power to paired testing, supporting the original comparative testing, and can be executed far more cost-effectively using a sequential monadic design.

Scenario: You work for a major yogurt manufacturer. You often conduct consumer testing to investigate the potential of new prototypes as well as the strength of your products compared to current and new competitors. Your consumer research involves liking ratings on a 7-point hedonic scale, followed by an ANOVA and individual mean comparisons. When few products are involved (2 or 3), you also conduct paired preference testing with all product pairs.

Since your research typically involves 5-10 products, collecting all pairwise preferences, along with product attribute intensities and JAR information, is not an option. You incorporate applicability scoring into your next project. This testing includes six vanilla flavored plant-based yogurts evaluated by 300 regular consumers of plant-based yogurt products. From attribute applicability scoring on individual samples (e.g., 'I like this product'), you can derive pairwise information for all product pairs without conducting any paired tests. In addition to your standard hedonic and intensity rating scales, the ballot you use includes 15 applicability statements, a subset of which is found in Table 1.

	Applies	Does not apply
I like this yogurt		
This yogurt has a strong off-flavor		
This yogurt tastes natural		
This yogurt tastes like a regular dairy vanilla yogurt		
This yogurt tastes like a low calorie product		

Table 1. Subset of applicability scoring portion of the plant-based yogurt ballot.

Applicability Scoring and Derived Preference: Applicability scoring is an easier task for consumers to perform than the process of rating products on sets of 7- or 9-point category scales. It is also very efficient and involves the monadic evaluation of a single product. While the applicability scores of the attributes first provide useful information on the profile of each product, the same scores can also be combined by pairs of products to derive pairwise information. When comparing two products, respondents are split into four categories: Those who thought the statement applied to both products, to neither product, and to one product but not the other (2 groups). A typical result is summarized in Table 2.

		Product 2	
		Applies	Does not apply
Product 5	Applies	130	80
	Does not apply	51	39

Table 2. Data example for the 'I like this yogurt' statement.

Using this data, we can conduct a chi-square test (McNemar's test)[2] between the two counts on the off-diagonal (80 vs. 51) to see if Product 5 and Product 2 differ significantly. We find that they do, $p = 0.01$. We also note that the derived preference result is 61% in favor of Product 5 (80/131). The similarity of the preference proportion and the derived proportion was discussed in the original tests of this method[3].

While preference information can be derived from applicability data, there is a need for a theoretical analysis to compare the power of the two approaches. Figure 1 summarizes the models that will be used to account for the applicability data and also the preference data. The perceptual distribution of each product is assumed to be normal with a variance of 1. The distributions of the two samples are assumed to be separated by a distance of δ, with Y liked more/preferred over X. Both of these Thurstonian models have been discussed in detail in previous papers[1,4].

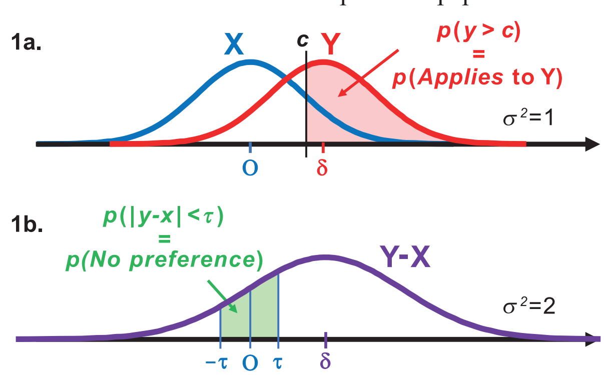

Figure 1. Graphical representation of the applicability scoring and paired preference decisions rules.

Figure 1a illustrates the applicability decision rule: The respondent uses a criterion c on the sensory/hedonic continuum. If the momentary perception is greater than c, the response will be 'Applies', otherwise 'Does not apply'.

The shaded area represents the proportion of times Y will be higher than c, i.e., the proportion of times the statement will apply to Y. We can see that the statement will apply more to Y than X, based on their respective areas above c.

Figure 1b illustrates the preference test decision rule: For the 2-Alternative Choice method, the decision rule is better represented using the Y-X difference distribution which has a variance of 2. To decide whether one product is preferred over the other, the respondent uses a τ criterion. If the perceived difference is greater than τ, the respondent will choose Y over X if at that moment $y > x$, or X over Y if at that moment $x > y$. If the perceived difference is smaller than τ, the respondent will answer "No preference".

Simulations were conducted while varying c, τ, δ, and the experiment's sample size. For each of these scenarios, the power of the method was determined. In the case of the simulated preference testing, the results were tested among those who expressed a preference. Figure 2 shows the power curves for parameters typically found in practice.

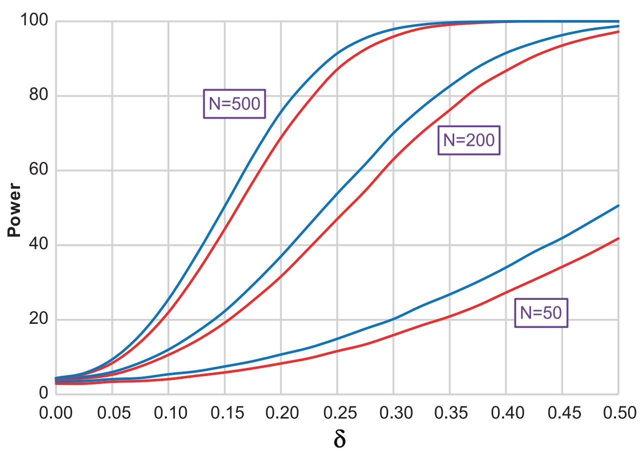

Figure 2. Power comparisons between applicability and preference testing based on δ for $c = 0.25$ and $\tau = 0.3$ with samples sizes of 50, 200 and 500.

Based on these simulations it was found that, for a given sample size, the paired preference results, based on those who expressed a preference (the most powerful method of analysis for preference) was only slightly higher than applicability scoring. This is a valuable finding for a derived preference measure given the advantages in cost and execution of applicability scoring. The first advantage is that applicability scoring does not require pairwise testing and can be obtained in a sequential monadic design involving more than two products. This allows the derivation of a preference analog, or other attribute, in the form of pairwise ratios over all product pairs, as illustrated in our example (80/131 = 61% in favor of Product 5). The second advantage is that the 'both' and 'neither' categories are also useful to get a measure of the number of consumers who might not have a strong preference between the samples. In the traditional paired preference test with a 'no preference' option, these consumers should answer 'no preference'. However, as has been reported repeatedly, even putatively identical products will result in 'no preference' proportions of only about 20%[4], far from the 100% that would be expected intuitively. Applicability scoring has the potential to reduce this paired preference bias related to demand characteristic (consumers feeling that they should prefer one product over the other). In our example, we would conclude that 56% of the respondents had no meaningful preference.

While this example focused on liking and preference, any other attribute information collected using applicability scoring can be treated in a similar way, providing relative strength insights that may be more sensitive than average rating scores on a category scale.

Deriving Preference for the Yogurt Samples: Using the applicability scores you collected for the 6 products over the 15 statements, you create 15 tables summarizing the paired ratios calculated using the 2x2 table off-diagonal as described previously. Figure 3 summarizes the preference proportions from the 'I like this product' statements.

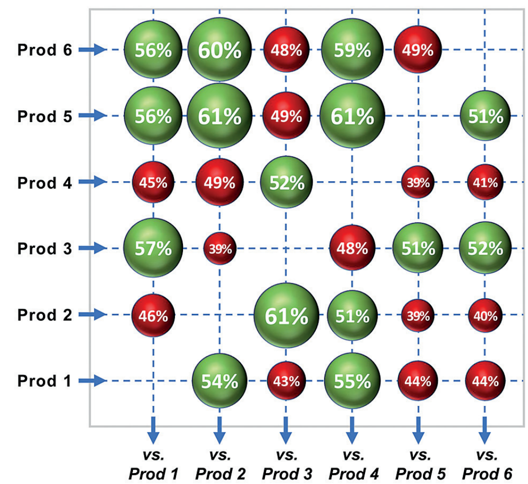

Figure 3. The circle is green if the product on the Y axis is preferred over the product on the X axis, otherwise red.

Conclusion: Applicability scoring has several advantages over traditional preferential choice data collection techniques. These advantages include ease of implementation and execution, generation of an applicability-based attribute profile, derivation of pairwise information from monadic evaluations, information on 'like both' and 'like neither', and potentially more meaningful preference ratios. Coupled with the theoretical work supporting previous experiments referred to in this report, the method discussed for deriving analog preference and attribute measures deserves further exploration and research to uncover greater insights from sensory and consumer testing.

References

1. Ennis, D.M. and Ennis, J.M. (2013). Analysis and Thurstonian scaling of applicability scores. *Journal of Sensory Studies,* **28**(3), 188-193.
2. Ennis, D. M. and Ennis, J. M. (2018). Interpreting Applicability Scores. In D. M. Ennis, J. M. Ennis, and B. Rousseau (Eds.), *Tools and Applications of Sensory and Consumer Science* (pp. 18-19). Richmond, VA : The Institute for Perception. (Original work published 2011)
3. 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_xhmfyd55uzds zmvz5bnqoqvt.pdf.
4. Ennis, D. M. and Ennis, J. M. (2012). Accounting for no difference/preference responses or ties in choice experiments. *Food Quality and Preference,* **23**(1), 13-17.