Preference without a Difference

Benoît Rousseau and Daniel M. Ennis

Background: There are certain universally accepted tenets that support the continued adoption of sensory evaluation methods. One of these strongly held beliefs is that a person cannot have a preference between two perceptually identical products. This idea justifies the existence of difference testing sensory programs in many consumer product companies and is the basis for the argument that if an internal panel cannot perceive a difference, then a consumer will not. Consequently, a consumer will not have a preference. To use the principle requires a method to decide if two products are different. Typically the basis for that decision depends on the use of a difference testing method, such as the triangle test, followed by a statistical test of the results based on a null hypothesis of no difference. In many cases, failure to reject the null hypothesis from the results of an expert or experienced internal panel are used as an indication that consumers are unlikely to have a preference. An advantage to this type of procedure is that relatively rapid and inexpensive testing can be conducted that may obviate the need to conduct expensive consumer preference tests when a difference is not detected.

As with any declaration of a fundamental principle, it is usually in the application of the principle that the difficulties arise. In this technical report it will be shown how a frequently used difference testing method may rarely identify a difference but a consumer preference test may show that one product is preferred to another. It will be shown that although the tenet stated earlier may be abstractly justified, difference and preference test results depend critically on the methods chosen to obtain the data and the form of analysis used to interpret them. Difference testing interpretation using significance tests leads to binary decision-making ('go/no go'). Differences in methodologies and the magnitude and precision of the detected sensory difference are not considered and may lead to inexplicable results. While internal testing may lead to a conclusion that there is no evidence of a difference, a consumer preference test may disagree.

Scenario: Your company markets confectionery brands that use sucrose and light versions prepared with an artificial sweetener, Sw_1. A change in regulation calls for a reduction in the amount of Sw_1 in your light products. Based on preliminary research, your product development group recommends using a mixture of Sw_1 and another sweetener, Sw_2 which is sweeter on an equal weight basis. This change will address the new regulation requirement as well as result in lower product costs, since Sw_2 is a less expensive option. Your task is to find the best Sw_1/Sw_2 ratio that will provide lower costs without creating a perceivable sensory difference from your current brand.

	Sweetener	P_1	P_2	P_3	P_4	P_5
Percentage	Sw_1	100%	80%	60%	40%	20%
	Sw_2	0%	20%	40%	60%	80%

Table 1. Composition of the five samples used in the sweetener modification research.

You first select five samples for testing. The sample compositions are summarized in Table 1. All comparisons will be made to P_1, your current brand artificially sweetened with Sw_1. You decide to use a discrimination test and find the sample with the highest Sw_2 percentage that does not result in a significant sensory difference at the 5% level.

When conducting discrimination testing, your sensory group uses the triangle test and uses a sample size of about 40 subjects who are experienced company employees participating regularly in difference tests. The results are shown in Table 2.

Product Comparison		# Correct	Sample Size	*p*-value
P_1 vs.	P_2 (20% Sw_2)	14	41	*> 0.05*
	P_3 (40% Sw_2)	15	42	*> 0.05*
	P_4 (60% Sw_2)	15	39	*> 0.05*
	P_5 (80% Sw_2)	20	43	*< 0.05*

Table 2. Triangle test results for four comparisons.

Based on these results, you conclude that a modified candy with 60% of Sw_2 was not significantly different from P_1 using the triangle test with your internal panel. At 80% the change is too large and results in a significant difference at the 5% level.

With insight from the difference testing, you recommend that a confirmatory preference test with consumers be conducted for 60% and 80% Sw_2 inclusion levels. You expect that the preference tests will show a difference at the 80% level but not at the 60% level. For this research you recruit 120 users of the candy category and they perform the two paired preference tests within a session. The results are shown in Table 3.

Product Comparison		# Consumers Choosing P_1	Sample Size	*p*-value *(two-tailed)*	*Preference Proportion*
P_1 vs.	P_4 (60% Sw_2)	87	120	*< 0.05*	*73%*
	P_5 (80% Sw_2)	94	120	*< 0.05*	*78%*

Table 3. Paired preference results for 120 category users.

These results are not what you expected since it is assumed that if an internal experienced panel cannot detect a difference, then consumers should not have a basis for preference. *How can a preference exist if there is not a sensory difference large enough to be detected by experienced panelists?*

Since you expect the inclusion of Sw_2 to increase sweetness, you consider using the 2-alternative forced choice (2-AFC) method to further explore sensory differences between the control and levels of Sw_2 inclusion. These tests are based on the knowledge that Sw_2 is sweeter than Sw_1 and greater inclusion should lead to higher sweetness. You also conduct additional preference tests to evaluate P_1 compared to P_2 and P_3 because your initial preference tests failed to find an acceptable inclusion level. The results of these experiments are given in Tables 4 and 5.

From these results you have learned there is evidence that consumers prefer P_1 to all inclusion levels of Sw_2 from 40% to 80%. The results do not support a preference for P_1 over the 20% inclusion level. You also learned from the 2-AFC

Product Comparison		# Consumers Choosing P_1	Sample Size	*p*-value (two-tailed)	*Preference Proportion*
P_1 vs.	P_2 (20% Sw_2)	65	120	*> 0.05*	*54%*
	P_3 (40% Sw_2)	75	120	*< 0.05*	*63%*

Table 4. Second paired preference test results for 120 category users.

Product Comparison		# Panelists Choosing P_1	Sample Size	*p*-value (one-tailed)	*Choice Proportion*
P_1 vs.	P_2 (20% Sw_2)	22	40	*> 0.05*	*55%*
	P_3 (40% Sw_2)	26	40	*< 0.05*	*65%*
	P_4 (60% Sw_2)	29	41	*< 0.05*	*71%*
	P_5 (80% Sw_2)	32	43	*< 0.05*	*75%*

Table 5. 2-AFC test results for four pairwise comparisons.

results reported in Table 5 that consumer preference seems to track quite well with sweetness level, so preference may depend almost exclusively on sweetness. However, your triangle test results reported in Table 2, in contrast to the 2-AFC results in Table 5, show that you cannot rely on *p*-values for particular methods to make decisions and that methods differ in sensitivity. The latter conclusion means that the application of the principle – *a preference cannot be expressed for identical products* or *a preference depends on demonstrating a difference* – may hinge on the methodology used to determine *difference*.

Methodology and the Scaling of Sensory Magnitudes: Hypothesis tests and their associated *p*-values alone do not provide a complete basis for decision making, particularly in sensory research[1,2]. Increasing the sample size can transform a non-significant difference into a significant one without changing the size of the difference. Determining the size of the sensory difference and its precision is often required. Thurstonian models[3,4,5] for difference tests provides a method to determine the required scale values for different methods and predict why difference testing methods differ enormously in power. Scaled difference testing results can be used to set a consumer relevant action standard that, under similar training conditions, is independent of methodology.

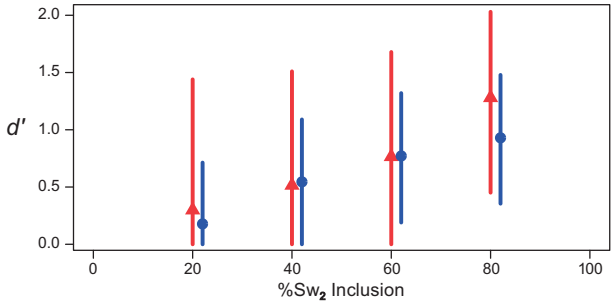

Figure 1. Relationship between the Sw_2 percentage and corresponding *d′* value when compared to P_1 (100% Sw_1) for the triangle test (red triangles) and the 2-AFC test (blue filled circles). The vertical lines are 95% confidence intervals which are truncated when they fall below zero. Although both methods are tested at the 20%, 40%, 60%, and 80% inclusion levels, the 2-AFC test results are offset slightly so as not to obscure the triangle test results.

Resolving the Methodological Conundrum: You calculate the *d′* values (scaled sensory difference) from your internal triangle test and the 2-AFC data. Figure 1 summarizes the relationship between the Sw_2 percentage in each sample and the corresponding sensory difference in terms of *d′* values from both methodologies. You conclude that the results from both panels were based on similar underlying differences within each product pair, but the confidence intervals for the 2-AFC are much shorter than those of the triangle test. The triangle test, due to its insufficient power, missed differences that may be important to your consumers, as demonstrated by the preference results. A preference or difference of 55:45 or 45:55 has been used to establish bounds on equivalence[6]. These splits correspond to bounds slightly less than 0.2 δ values (*d′ is an estimate of* δ). Since the 20% inclusion level is at about 55% for the preference test and less than 0.2 *d′* with the 2-AFC, you recommend this level of Sw_2 inclusion. The estimate for the triangle test was 0.30 but its variance was also much higher than that of the 2-AFC. Due to its apparent low power, you resolve to research other more sensitive methods, such as the tetrad method[7,8,9], to replace the triangle test. You have learned that the tenet you began with cannot be employed without considering the method used to apply it.

Conclusion: The idea that a person cannot have a preference between two perceptually identical products seems self-evident. However, the method used to evaluate differences makes implementation of this principle far from trivial[9,10]. Difference testing methods differ in power and in some cases are so insensitive that experienced panelists may not detect differences that drive consumer preference. Setting standards for differences that are method-independent using the most powerful difference testing methods is far superior to relying exclusively on *p*-values to make decisions.

References

1. Ennis, D. M. (1990). The relative power of difference testing methods in sensory evaluation. *Food Technology,* **44,** 114, 116 & 117.
2. Chew, V. (1977). Statistical Hypothesis Testing: An Academic Exercise in Futility. Proceedings of Florida State Horticultural Society, Vol. 90, p. 214, Lake Alfred, FL.
3. Ennis, D. M. (2016).*Thurstonian Models: Categorical Decision Making in the Presence of Noise,* Richmond, VA: The Institute for Perception.
4. Ennis, D. M. (1993). The power of sensory discrimination methods. *Journal of Sensory Studies,* **8,** 353–370.
5. Rousseau, B. (2015). Sensory discrimination testing and consumer relevance. *Food Quality and Preference,* **43,** 122-125.
6. ASTM. (2013). *Standard guide for sensory claim substantiation - E1958 - 12.* ASTM international.
7. 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.
8. Ennis, J.M. and Christensen, R. H. B. (2014). Precision of measurement in Tetrad testing. *Food Quality and Preference,* **32,** 98-106.
9. Ishii, R., O'Mahony, M., and Rousseau, B. (2014). Triangle and tetrad protocols: Small sensory differences, resampling and consumer relevance. *Food Quality and Preference,* **31,** 49-55.
10. Geelhoed, E. N., MacRae, A.W., and Ennis, D. M. (1994). Preference gives more consistent judgments than oddity only if the task can be modeled as forced choice. *Perception & Psychophysics,* **55**(4), 473-477.