The Institute for Perception

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

Reprinted from IFPress (2010) 13(4) 2,3

Action Standards in a Successful Sensory Discrimination Program

Benoît Rousseau

Background: Sensory quality assurance is a critical activity in the management of consumer product companies. Cost-savings initiatives, ingredient changes, and modifications to comply with new regulations or dietary changes lead to a need for product sensory measurement. Discrimination testing involves the use of a specific class of sensory evaluation methods. Such procedures are often used by food, beverage and personal care product companies to measure differences between a gold standard product and various alternatives. In the vast majority of cases, interest centers on whether the modified alternative is a substitute for the gold standard. The objective of this report is to discuss the background to setting up action standards in a sensory quality control program using discrimination methodologies to qualify product modifications.

Scenario: Working within a company that markets tomato juice products, you were recently assigned the responsibility to establish a sensory evaluation program to make decisions on the suitability of product modifications. Prior to this assignment, several costly product changes were approved through internal product testing but resulted in negative consumer reactions. Consequently, you want to design a program that will allow you to detect differences that might be perceived by consumers, while providing assurance that differences that are negligible will be categorized as such.Your background in quality control equips you to understand the necessity of setting specifications which will allow you to determine whether a given product meets present standards. You want these standards to be consumer relevant, since consumer reactions are the benchmark by which the success of your program will be measured. You decide to test various sample pairs both with your internal discrimination panel (N=25) and with consumers (N=325) based on the experience that a preference detected by a consumer sample of that size is meaningful (power of 80% to detect a 57/43 preference split at α=0.05). The idea is to estimate the sensory difference between the samples and to relate it to preference responses from consumers. Figure 1 illustrates the process planned to set up specifications to be used as action standards.

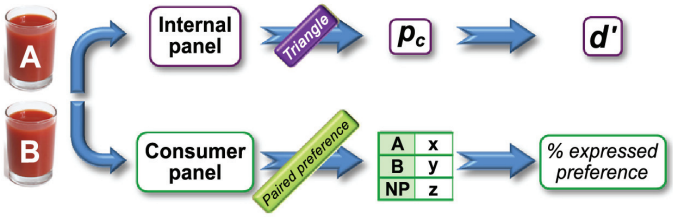

Figure 1. Initial steps for setting-up action standards.

Power and Acceptable Quality Levels: The importance of establishing bounds for acceptable product is not universally appreciated. Often the approach used involves a panel of internal subjects performing a discrimination test, such as a triangle or duo-trio, and if no significant difference is found, the project will go ahead to the next stage towards a product modification. However, how can one be sure that whatever difference existed between the products was truly negligible? Is it possible the test simply lacked power relative to some defined bounds? Would a conclusion of no difference be reached if the sample size was doubled or tripled? There are numerous cases of products that successfully passed internal testing only to fail in subsequent consumer investigations. Also, it is possible that if a test is very powerful, a significant difference might be unimportant to the consumer. This may result in a missed opportunity to successfully implement a modification.

In order to use the concept of power (or an alternative based on equivalence theory[1,2], which will not be discussed in this report), one needs a quantified measure of sensory difference to define the point at which a difference matters from a practical standpoint. One possibility is to use proportion of discrimintors[3], but this measure is method-specific. One could set an action standard based on experience or subjectively, as sometimes occurs when setting specifications for acceptance sampling. However, if the standard could be validated using consumer-relevant behavior it would be more robust.

Table 1 illustrates four outcomes possible from a difference test. The importance or negligibility of the size of the difference relates to the consumer's perception. The objective is to maximize the number of correct decisions, while minimizing the number of false alarms and misses. The two latter outcomes are much more likely when action standards are not used.

		Result significant	
		Yes	No
Size of the difference	Negligible	False alarm	Correct decision
	Important	Correct decision	Miss

Table 1. Four possible results from a discrimination test.

Data Collection and Results: You select the sample pairs based on various process and formula modifications of your gold standard. You have a total of twelve pairs of products. Over three successive days, consumers evaluate four pairs per day. Panelists perform three replications for each pair of samples for a total sample size of seventy five trials. You establish that your panelists are homogenous[4] so replicated measures from one panelist are equivalent to separate measures from different panelists. Consumers taste each pair once, indicating which of the two samples they prefer. Figure 2 summarizes the results from the testing of the 12 pairs with your internal panelists

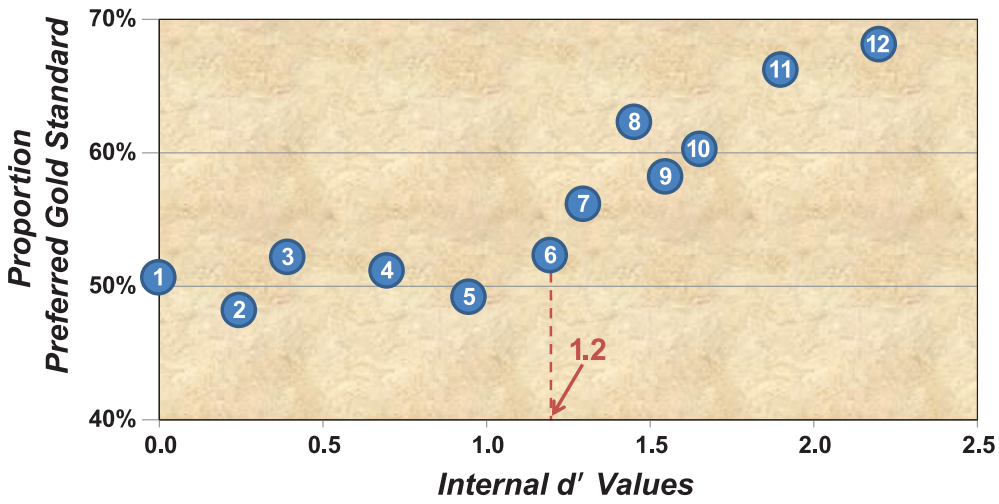

Figure 2.
Counts of each response category for the 12 product pairs with significance levels.

(d' values[5]) and the heavy users of your products (proportion of preference for the gold standard). Statistical analysis of the preference data shows that pairs 7-12 exhibit a significant result at $p=0.05$, but pairs 1-6 do not.

The results obtained are straightforward to interpret. When the d' is 0 (pair 1) and no sensory difference is perceivable, the proportion of expressed preferences splits about equally between the tested items. As the sensory difference becomes larger, up to a d' of 1.20 (pair 6), the gold standard is not preferred significantly over the alternative. However, for a d' of 1.30 (pair 7), the proportions significantly split in favor of the gold standard product. Consequently, you conclude that your action standard could be initially set with a d' of 1.20.

Using the Action Standard: Now that you have determined the size of the difference you will initially target, you need to establish the remaining aspects of your action plan, namely the errors linked to the hypothesis testing: Type I (α, probability of wrongly concluding that there is difference) and Type II (β, probability of wrongly concluding that the difference is negligible). These values will then lead to the recommended sample size for your discrimination investigations. IFPrograms™ software[6] or published tables[7] that relate α, β, d' and sample size can be used for these calculations. For standard values of α = 5% and β = 20% (power = 80%), the predicted sample size needed for the triangular method is 106. You typically conduct tests with 25 panelists who perform three replications for a total of 75 trials (assuming no overdispersion). You decide to increase your base testing by two panelists and one extra replication, for a total of 108 judgments. Using this sample, any future test yielding a significant outcome will result in the rejection of the change, while a non-significant result is unlikely to occur if the real difference exceeds a d' of 1.20 (β=20%). A higher power level (lower β value) would necessitate a greater sample size or an increased Type I error.

With the research completed, you also consider the possibility of switching to an alternative protocol, such as the 2-Alternative Forced Choice (2-AFC) or the same-different method, which would provide you with greater power for the same sample size. Fortunately, you won't have to re-conduct all of the above research to make this change as the use of a standardized unit such as d' will allow you to use the research conducted with the triangular method to establish action standards using other methodologies.

Conclusion: Many companies in the food and personal care industries rely solely on statistical significance with internal panels to decide whether a process or formulation change is judicious. Depending on the sample size and the methodology used, this can result in 'false alarms' (falsely concluding a change would be detected/rejected by consumers) or 'misses' (falsely concluding a change would not be detected/rejected by consumers). In standard industrial practice the latter is much more common, due to the use of discrimination techniques that lack power to detect sensory differences. The use of action standards offers a compelling solution to this issue by taking into account the size of the sensory difference that is relevant to consumers. A sensory discrimination program validated with consumer relevant measures contributes to improved confidence in the management of consumer sensory quality.

References and Notes

1. Ennis, D.M. and Ennis, J.M. (2008). New developments in equivalence testing. *IFPress*, **11**(4), 2,3.
2. Ennis, D.M. and Ennis, J.M. (2009). Hypothesis testing for equivalence based on symmetric open intervals. *Communications in Statistics*, **38**(11), 1792-1803.
3. Rousseau, B. and Ennis, D.M. (2007). Why proportion of discriminators is method-specific. *IFPress*, **10**(3), 2,3.
4. Ennis, D. M. and Bi, J. (1998). The beta-binomial model: Accounting for inter-trial variation in replicated difference and preference tests. *Journal of Sensory Studies*, **13**, 389-412.
5. Ennis, D.M. (1998). Thurstonian scaling for difference tests. *IFPress*, **1**(3), 2,3.
6. Information regarding IFPrograms™ software can be found at http://www.ifpress.com.
7. Ennis, D. M. (1993). The power of sensory discrimination methods. *Journal of Sensory Studies*, **8**, 353-370