Using a Consensus Degree of Difference Scale as an Alternative to Triangle Testing

Yvonne Koelliker[1], Lydia J.R. Lawless PhD[2], Annlyse N. Retiveau Krogmann PhD[1], Ashley Sobel[1], Gail Vance Civille[1]
[1]SENSORY SPECTRUM, INC., UNITED STATES, [2]THE HERSHEY COMPANY, UNITED STATES

Abstract

Many product development, improvement and quality initiatives require product or process changes without noticeably altering sensory properties of the end product. To assess whether such changes will remain unnoticed in the marketplace, discrimination tests have traditionally been used, as they allow for management of alpha and beta-risks simultaneously and provide clear "pass-fail" decision criteria. However, to ensure optimal risk management, they often require access to a large respondent base, and can be costly and time-consuming. Further, in case of failure, they may not provide sufficient diagnostics to accurately explain product differences and to strategically direct development efforts.

Expert panels trained in Consensus Degree of Difference (DOD) methodology offer an alternative to these traditional methods. Panels are trained to rate the strength of the overall difference across a wide range of consumer products and quantify whether differences will be noticed or unnoticed by consumers; they provide direction for changes, all with a smaller sample size, in a shorter timeframe and at lower cost. The current study outlines how to develop a consensus DOD scale and validate its predictive power. Data from the Spectrum Descriptive Panel trained in DOD is compared to findings from traditional triangle testing.

Results show a high degree of correlation ($r = 0.97$, $p < 0.001$) between the DOD scores and the number of correct responses in discrimination testing for similarity across two beverage systems. Critical DOD scores along the scale that provide adequate protection against both risk for false positives and false negatives are determined. These results indicate DOD scores can be used as an analog for beta-risk, thereby shortening and strengthening the product development process.

Background, Objectives, & Samples

Background

Triangle testing is widely used in the industry, typically to understand whether a perceivable difference exists between two products, often times in the context of a product or process change (ingredient substitution or removal, cost reduction initiatives, alternate supplier) or for QA/QC purposes.

Along with other difference testing methods, triangle testing is typically performed with naïve consumers or panelists oriented to the method. While this testing is fairly quick and simple, it can be costly in terms of internal resources, product costs, incentive costs, and may not be optimal when numerous products need to be screened for similarity or difference.

Degree of difference (DOD) methodology, on the other hand, requires the use of a panel trained to recognize and quantify overall differences among products, so that fewer resources can be used on a day-to-day basis (post-training) to highlight differences or similarities among products. DOD panels may vary in size and degree of training.

In the following study, the use of an Expert Spectrum Panel using a universal DOD scale* is discussed as a possible alternative to or precursor of traditional triangle testing methodology.

Similar to other traditional discrimination methods, one of the challenges of using the DOD methodology relies in risk assessment.

- With triangle testing, for example, one may decide panel size based on predetermined α-risk, β-risk and Proportion Distinguishers (P_d) to manage the risk of falsely concluding for similarity or differences. One may decide to be more conservative in some cases and more lenient in others.
- With a DOD method, empirical evidence and training or the use of a blind control usually aid in determining a critical point on the scale that corresponds to what may be a noticeable difference in the marketplace.

* The scale is considered universal as it can be used across a wide variety of consumer products, food and non-food alike, and scale points document a similar size difference independent of category tested.

Objectives & Samples

The current study proposes:

1. To show alignment between triangle testing (discrimination testing) and DOD methodology, in terms of perceivable differences within products. Samples are selected to show a range of variability from a control.
2. To uncover the relationship between DOD ratings and P_d similar to a triangle test and given pre-determined sample size and β-risk.
3. To highlight how DOD ratings may lead to sound decision-making in light of project objectives and pre-determined risk-levels.

Tropical Punch Kool-Aid

ID	Full Name
Control	Tropical Punch Made to package directions (1.9 L water + 1 packet Kool-Aid mix + 200 g sugar)
Red. Sugar	Tropical Punch Red. Sugar (1.9 L water + 1 packet Kool-Aid mix + 175 g sugar)
3/4 Diluted	Tropical Punch Diluted 3/4 (1.4 L control + 473 ml water)
1/2 Diluted	Tropical Punch Diluted 1/2 (946 ml control + 946 ml water)

Retail Orange Juice

ID	Full Name
Control	Tropicana Pure Premium OJ (no pulp)
Trop PP	Tropicana Original OJ (no pulp) - different date code than control
Trop HK	Tropicana Healthy Kids OJ (no pulp)
FL Nat	Florida's Natural OJ (no pulp)

Methodology

Triangle Test vs. DOD

	Triangle Test	DOD
Respondents	N = 60, untrained evaluators, all employees of Sensory Spectrum. Respondents are oriented to the task	N = 8, members of the Sensory Spectrum Expert Food Panel, trained in the Spectrum Method and Degree of Difference methodology.
Task	Each respondent is presented with 3 samples. Two are the same, one is different. Respondents are asked to select the odd sample. Each respondent completes six triangle test within 1 hour. For each test, the number of correct responses are recorded.	Expert panelists use the degree of difference scale. The DOD method uses a 0.0 to 10.0 scale indicating how different a product is from a reference product or control, with 0.0 meaning no difference and 10.0 being extremely different. A consensus score is provided.
Criteria and interpretation	Prior to testing α-risk is set at **0.10, β-risk=0.05, P_d=30%**. A minimum of 54 responses are needed. A total of **60 respondents** complete the evaluations. Out of 60 responses: • **26** or more correct responses is evidence of a difference. • **25** or fewer responses is evidence of similarity (95% sure that no more than 30% of respondents can detect a difference when samples are evaluated side-by-side.	The degree of difference (DOD) rating quantifies the magnitude of the difference but is not directional. • A DOD between **2.0 and 3.0** is low enough that a number of consumers would not be able to tell the two samples apart. • A **DOD of 5.0** or greater is high enough that a number of consumers would be able to tell the samples apart. • A **DOD of 4.0** is a questionable area ("grey area") where consumers may or may not be able to tell the difference.

Note - In either case, perceiving a difference does not necessarily indicate a difference in liking

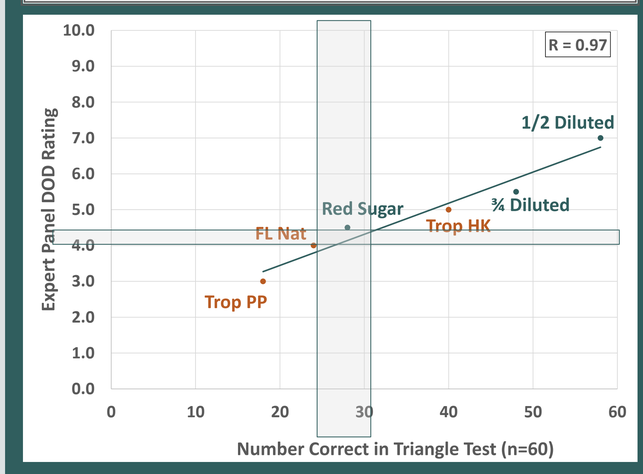
Results

And Risk Assessment

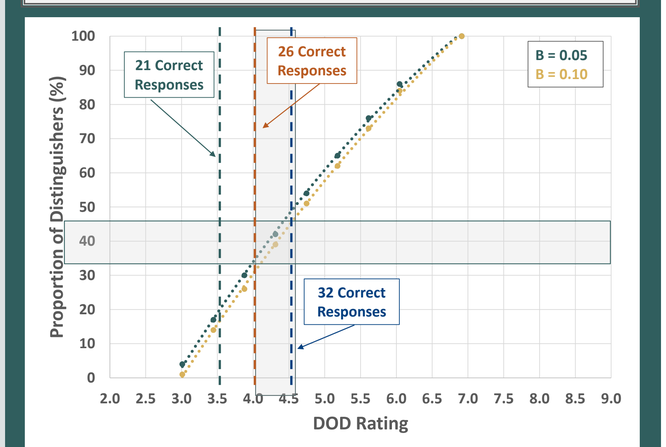
Triangle testing and DOD testing lead to similar conclusions.

	Triangle Test		DOD	
	Results (# Correct out of 60 observations)	Conclusion (α=0.10, β=0.05, Pd=30%)*	Results	Conclusion
KOOL- AID				
Red. Sugar	28	Different	4.5	Grey Area
3/4 Diluted	48	Different	5.5	Different
1/2 Diluted	58	Different	7.0	Different
ORANGE JUICE				
Trop PP	18	Similar	2.0	Similar
FL Nat	24	Similar	4.0	Grey Area
Trop HK	40	Different	5.0	Different

Results from DOD scaling are highly correlated with results from triangle testing.



Risk Assessment in the Grey Area



The Grey Area

The horizontal grey area corresponds to 4.0 - 4.5 on the DOD scale.

The vertical grey area corresponds to n = 26 - 32 correct responses for 60 triangle test panelists. This area may move, depending on level of risk one is willing to take.

Concluding for similarity when DOD is 4.0 - 4.5 would be akin to concluding for similarity with 90% certainty that no more than 30-45% of the population would be able to tell samples apart when samples are seen side-by-side.

One would reach the same conclusion with 26 - 32 correct responses for 60 triangle test panelists.

Applications & Advantages

- Degree of Difference methodology may be implemented at all stages of the development process; benchtop screening, selection of samples for consumer testing, quality control once products are in production, etc.
- While this methodology is shown to be an analog for beta-risk, it is not meant as a replacement to traditional discrimination testing. Rather it can reduce the use of this type of testing (thereby reducing costs) when a panel trained in DOD methodology is available.
- DOD methodology is less time-intensive from a training standpoint than descriptive analysis panel training and panels can be trained and validated within a few months.
- Once the panel is trained and validated in one category, the use of a universal DOD scale allows for quick and easy application to other, non-related categories, resulting in rapid turnaround in terms of panel validity and data collection.

Next Steps

To further validate the relationship between the universal Spectrum DOD scale and risk-based discrimination methods (beyond triangle testing) across non-food categories and food categories other than beverages.

Presented at the Society of Sensory Professionals Meeting – September 2014

www.sensoryspectrum.com

sensoryspectrum