Sensory Equivalence testing - the reversed null hypothesis and the size of a difference that matters

Paul Arents, C.A.A. Duineveld, Bonnie M. van der Pers - King

Quest International Nederland BV,
P.O. Box 2, 1400 CA Bussum
The Netherlands

E-mail: paul.arents@questintl.com

Summary

Two procedures are given for determining a value for *delta*, the smallest degree of difference detectability that matters, when one is using a rating scale for intensity measurement. This value is then implemented in a Two One-Sided Tests procedure (TOST) in order to demonstrate equality in sensory equivalence testing. The procedure is of importance to those involved in product matching after process change. An example is given for which limited sample availability permitted only one triangle test and one profiling test. The lack of significant difference between the two products as shown by results from the triangle test was demonstrated to be due in fact to sensory equality of the products as measured by profiling.

Two One-Sided Tests (TOST)

A t-test for determining differences between two products T and R has null and alternative hypotheses

$$H_0: \mu_T - \mu_R = 0$$
$$H_1: \mu_T - \mu_R \neq 0$$

Under normality assumption H_0 will be rejected at level α if

$$t = \frac{|\hat{d}|}{se(\hat{d})} > t_{\frac{1}{2}\alpha,r}$$

The Two One-sided Tests procedure (TOST) is a combined technique to determine equivalence of two products [Schuirmann (1987). *A comparison of the two one-sided tests procedure and the power approach for assesing the equivalence of average bioavailability*. J. of Pharm. and Biopharm.]. Two sets of interval null hypothesis of inequivalence and alternative hypotheses are defined as:

$$H_{01}: \mu_T - \mu_R \leq \delta_L$$
$$H_{11}: \mu_T - \mu_R > \delta_L$$

and

$$H_{02}: \mu_T - \mu_R \geq \delta_U$$
$$H_{12}: \mu_T - \mu_R < \delta_U$$

Graphical representation of these hypotheses are shown in Figure 1. The limits δ_L and δ_U represent the *delta* value which need to be defined by the sensory scientist beforehand.

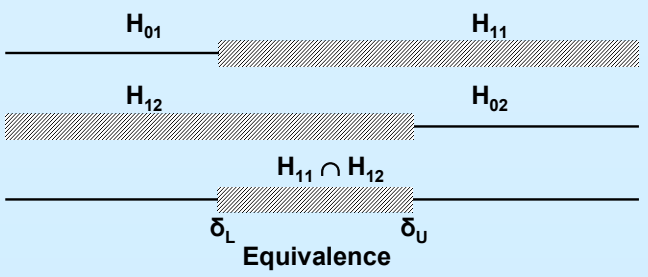

Fig. 1 Hypotheses used in TOST procedure

Reject $H_0 = H_{01} \cup H_{02}$ at level α if

$$T_L = \frac{\hat{d}-\delta_L}{se(\hat{d})} > t_{\alpha,r} \text{ and } T_U = \frac{\hat{d}-\delta_U}{se(\hat{d})} < -t_{\alpha,r}$$

For simplicity reasons we take $\delta_L = -\delta_U$.

Delta, the minimum detectable difference

Procedure 1 [Cohen, J. (1977), *Statistical Power Analysis for the Behavioral Sciences*, Academic Press]

Cohen defined three classes for effect size (ES): small (0.2), medium (0.5) and large (0.8). ES is a standardized difference:

$$ES = \frac{\mu_1 - \mu_2}{\sigma}$$

The minimum detectable difference, *delta*, is defined as: $\delta = ES \cdot \sigma$, where σ is the panel standard deviation or panel precision. Per descriptor a one-way ANOVA with product as factor gives a MSE. The estimate of the panel precision, σ, for a descriptor class is the square root of the geometric mean of the MSEs. Values for σ were obtained from 18 projects carried out over the last year (Table 1).

Procedure 2 'Multiples of the Weber ratio (k)'

For 2 stimuli S_0 and S_1 the Weber ratio (k) is

$$\frac{S_1 - S_0}{S_0} = \frac{\Delta S}{S} = k$$

Literature values for Weber ratios, range from 1/10 to 1/2, for odor and taste. Sucrose is a common and well-suited substance that can be used to provide an estimate for a general value of k. Based on reported values and our measurements, an average value of k=0.15 was chosen. The corresponding *delta* can be obtained via our panel's dose-response curve for sucrose.

$$I = -64 + 364 \cdot \log(conc)$$

If $S_1 = (1+k) \cdot S_0$ then the difference in intensity equals

$$\begin{aligned} I(S_1) - I(S_0) &= (-64 + 346 \cdot \log(S_1)) - (-64 + 346 \cdot \log(S_0)) \\ &= 346 \cdot \log(S_1 / S_0) \\ &= 346 \cdot \log((1+k)S_0 / S_0) \\ &= 346 \cdot \log(1+k) \end{aligned}$$

This difference $346 \cdot \log(1+k)$ can be seen as a small *delta*. If $S_2 = (1+k)^2 \cdot S_0$ and $S_3 = (1+k)^3 \cdot S_0$ then S_2 and S_3 are associated with a medium *delta* of size $2 \cdot 346 \cdot \log(1+k)$ and a large *delta* $3 \cdot 346 \cdot \log(1+k)$ (see Figure 2).

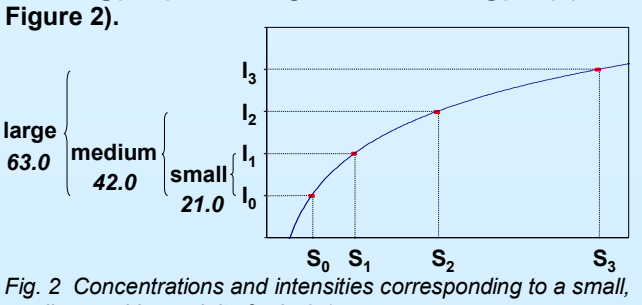

Fig. 2 Concentrations and intensities corresponding to a small, medium and large delta for k=0.15.

Example

Two snack products from different production sites were profiled once by 19 panelists who used 18 descriptors. A triangle test showed no significant difference between the two snack products: 6 / 19 proportion of correct identifications (p=0.6481).

Table 1 shows that estimates of the panel precision per descriptor class are relatively similar. Therefore an overall value of 69.4 was used to calculate *deltas* for each of the 3 effect sizes.

Descriptor Class	σ (95% Confidence Interval)	Effect Size		
		0.2	0.5	0.8
Basic Taste	60.8 (41.8, 88.4)	12.2	30.4	48.7
Retronasal	71.9 (49.3, 104.8)	14.4	35.9	57.5
Texture	69.7 (51.7, 93.8)	13.9	34.8	55.7
Tri-Geminal	69.7 (50.5, 96.1)	13.9	34.8	55.7
Lingering Sensation	66.5 (51.1, 86.5)	13.3	33.2	53.2
Overall	69.4 (47.7, 100.9)	13.9	34.7	55.5

Table 1 Panel precision σ per descriptor class and δ for 3 Effect Sizes. Values (Hz) reflect intensity scaling by the audio method.

Table 2 shows the resemblance between the *delta* values obtained for the 2 procedures.

	Cohen's ES	*delta*	Weber ratios	*delta*
small	0.2	13.9	1 x	21.0
medium	0.5	34.7	2 x	42.0
large	0.8	55.5	3 x	63.0

Table 2 Three levels of minimum detectable differences delta for the Cohen and the method based on Weber ratios

The profiling test (Figure 3) showed that the largest differences between the two products were to be found on descriptors 8, 12, 9 and 13, although a classical t-test indicated no significant difference between the products for any descriptor. The TOST for *delta* = 21.0 gave no significant equalities despite rejection of at least one of the one-sided t-tests for 9 descriptors at the 10% level, 3 of which were also rejected at the 5% level. As suspected, the small difference value for *delta* is not realistic for measurements by profiling.

Using the middle difference value of *delta* = 42.0, however, revealed significant equality by TOST at the 10% level for descriptors 3, 5, 7, 10 and 15. Under this criterion, each descriptor had at least one rejection of the one-sided tests for the 10% level, and for 12 descriptors this rejection was at the 5% level.

Choosing the large value of *delta* = 63.0 allowed one to claim TOST equality at the 10% level for all descriptors, 13 of which were also significantly equal at the 5% level (double green circles). The sensory scientist examining these data can therefore express confidence that the samples produced by both production sites are identical unless the descriptors 1, 8, 9, 12 and 13 are so important to product quality that they merit special attention.

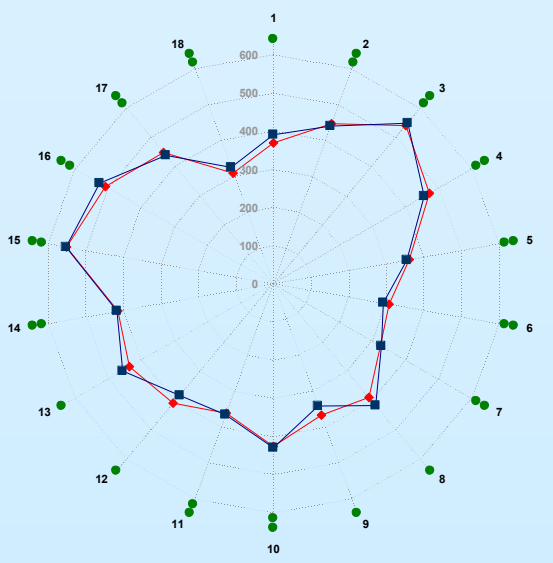

Fig. 3 Profiles for Reference ( ) and Test ( ) products and TOST significances at 5%( ) and 10% ( ) for delta=63.0