Designation: ~~E3009 – 23a~~ E3009 – 24

Standard Test Method for Sensory Analysis—Tetrad Test[1]

This standard is issued under the fixed designation E3009; the number immediately following the designation indicates the year of original adoption or, in the case of revision, the year of last revision. A number in parentheses indicates the year of last reapproval. A superscript epsilon (ε) indicates an editorial change since the last revision or reapproval.

1. Scope

1.1 This test method covers a procedure for determining whether a perceptible sensory difference exists between samples of two products or to estimate the magnitude of the perceptible difference.

1.2 This test method applies whether a difference may exist in a single sensory attribute or in several.

1.3 This test method is applicable when the nature of the difference between the samples is unknown. The attribute(s) responsible for the difference are not identified.

1.4 The tetrad test is more efficient statistically than the triangle test (Test Method E1885) or the duo-trio test (Test Method E2610).

1.5 *This standard does not purport to address all of the safety concerns, if any, associated with its use. It is the responsibility of the user of this standard to establish appropriate safety, health, and environmental practices and determine the applicability of regulatory limitations prior to use.*

1.6 *This international standard was developed in accordance with internationally recognized principles on standardization established in the Decision on Principles for the Development of International Standards, Guides and Recommendations issued by the World Trade Organization Technical Barriers to Trade (TBT) Committee.*

2. Referenced Documents

2.1 *ASTM Standards:*[2]

E253 Terminology Relating to Sensory Evaluation of Materials and Products
E456 Terminology Relating to Quality and Statistics
E1871 Guide for Serving Protocol for Sensory Evaluation of Foods and Beverages
E1885 Test Method for Sensory Analysis—Triangle Test
E2262 Practice for Estimating Thurstonian Discriminal Distances
E2610 Test Method for Sensory Analysis—Duo-Trio Test

2.2 *ISO Standards:*[3]

ISO 4120 Sensory Analysis – Methodology – Triangle Test
ISO 10399 Sensory Analysis – Methodology – Duo-Trio Test

[1] This test method is under the jurisdiction of ASTM Committee E18 on Sensory Evaluation and is the direct responsibility of Subcommittee E18.04 on Test Methods.
Current edition approved ~~Dec. 15, 2023~~Feb. 15, 2024. Published ~~December 2023~~March 2024. Originally approved in 2015. Last previous edition approved in 2023 as ~~E3009 – 23.~~E3009 – 23a. DOI: ~~10.1520/E3009-23A.~~10.1520/E3009-24.

[2] For referenced ASTM standards, visit the ASTM website, www.astm.org, or contact ASTM Customer Service at service@astm.org. For *Annual Book of ASTM Standards* volume information, refer to the standard's Document Summary page on the ASTM website.

[3] Available from International Organization for Standardization (ISO), 1, ch. de la Voie-Creuse, CP 56, CH-1211 Geneva 20, Switzerland, http://www.iso.org.

3. Terminology

3.1 *Definitions*—For definition of terms relating to sensory analysis, see Terminology E253, and for terms relating to statistics, see Terminology E456.

3.2 *Definitions of Terms Specific to This Standard:*

3.2.1 *α (alpha) risk*—probability of concluding that a perceptible difference exists when, in reality, one does not.

3.2.1.1 *Discussion*—

Also known as Type I Error or significance level.

3.2.2 *β (beta) risk*—probability of concluding that no perceptible difference exists when, in reality, one does.

3.2.2.1 *Discussion*—

Also known as Type II Error.

3.2.3 *δ*—Thurstonian measure of sensory difference (effect size) relative to perceptual noise (standard deviation) (see Practice E2262).

3.2.4 P_c—the probability of obtaining a correct answer from an assessor in the test.

3.2.4.1 *Discussion*—

If the products are indistinguishable sensorially, $P_c = ⅓$ in the tetrad test; while if the products are perceptibly different, $P_c > ⅓$.

3.2.5 P_d—proportion of assessors who can discriminate the two products in the test.

3.2.5.1 *Discussion*—

P_d is the measure of sensory difference used in the guessing model.

3.2.6 *product*—material to be evaluated.

3.2.7 *sample*—unit of product prepared, presented, and evaluated in the test.

3.2.8 *sensitivity*—general term used to summarize the performance characteristics of the test.

3.2.8.1 *Discussion*—

The sensitivity of the test is rigorously defined, in statistical terms, by the values selected for α, β, δ, or P_d.

4. Summary of Test Method

4.1 Clearly define the test objective in writing.

4.2 Choose the number of assessors based on the objective of the test (that is, testing for a difference or testing for similarity) and the level of sensitivity desired for the test. The sensitivity of the test is, in part, a function of two competing risks, α and β and the maximum acceptable difference between the samples, as measured by δ or P_d. ~~(that is,~~ The value chosen by the researcher for δ or P_d ~~a meaningful difference).~~ prior to the test represents the threshold of a meaningful difference, where a value larger than δ or P_d represents a meaningfully large perceptible difference. When testing for a difference, α is the risk of declaring the samples different when they are not, and β is the risk of not declaring the samples different when they ~~are. When testing for similarity, the meanings of α and β are reversed.~~ are (and the difference is, in fact, equal to δ or P_d). When testing for similarity, α is the risk of declaring the samples similar when they are not, and β is the risk of not declaring the samples similar when they are. Acceptable values of α and β vary depending on the test objective and should be determined before the test (see Appendix X1 and Appendix X2).

4.3 Each assessor receives four coded samples where two samples are of one product and the other two samples are of the other product being tested. The assessors are instructed to group the four samples into two groups of two based on similarity.

4.4 Results are tallied and significance determined by reference to a statistical table or software that calculates binomial probabilities.

5. Significance and Use

5.1 The test method is effective for the following test objectives:

5.1.1 To determine whether a perceptible difference results or a perceptible difference does not result, for example, when a change is made in ingredients, processing, packaging, handling, or storage; or

5.1.2 To select, train, and monitor assessors.

5.2 The test method itself does not change whether the purpose of the test is to determine that the products are perceptibly different versus that the products are sufficiently similar to be used interchangeably. Only the selected values of α, β, and δ or P_d change. If the objective of the test is to determine if there is a perceptible difference between two products, then initially the products are assumed to be indistinguishable (for example, H_O: δ or $P_d = 0$) and the data are examined to determine if the assumption can be rejected (that is, conclude that the products are perceptively different). If the objective is to determine if the two products are sufficiently similar to be used interchangeably, then initially the products are assumed to be meaningfully different (for example, H_O: δ or $P_d >$ the value chosen to represent a meaningful difference) and the data are examined to determine if the assumption can be rejected (that is, conclude that the samples are sufficiently similar to be used interchangeably).

5.3 The tetrad method involves the evaluation of four samples. When the products being tested cause excessive sensory fatigue, carryover, or adaptation, methods that involve the evaluation of fewer samples (same-different, triangle test, etc.) may be preferred.

6. Apparatus

6.1 Carry out the test under conditions that prevent contact between assessors until the evaluations have been completed, for example, using booths that comply with STP 913 (**1**).[4]

6.2 Sample preparation and serving sizes should comply with Practice E1871. See Refs (**2**) or (**3**).

7. Assessors

7.1 All assessors must be familiar with the mechanics of the tetrad test (the format, the task, and the procedure of evaluation). Experience and familiarity with the product and test method may increase the sensitivity of an assessor and may therefore increase the likelihood of finding a significant difference. Monitoring the performance of assessors over time may be useful.

7.2 Choose assessors in accordance with test objectives. For example, if the project results are to represent the general consumer population, assessors with unknown sensitivity might be selected. To increase protection of product quality, assessors with demonstrated acuity should be selected.

7.3 The decision whether or not to train assessors on the samples before testing should be addressed prior to testing. Training may include a preliminary presentation on the nature of the samples and the problem concerned. For example, if the test concerns the detection of a particular taint, consider the inclusion of samples during training that demonstrate its presence and absence. Such demonstration will increase the panel's acuity for the taint but may detract from other differences. See STP 758 for details (**4**). Allow adequate time between the exposure to the training samples and the actual tetrad test to avoid carryover.

7.4 During the test sessions, do not give any information about product identity, expected treatment effects, or individual performance until all testing is complete.

7.5 Avoid replicate evaluations by the same assessor whenever possible. However, if replications are needed to produce a sufficient number of total evaluations, every effort should be made to have each assessor perform the same number of replicate evaluations.

8. Number of Assessors

8.1 Choose the number of assessors to yield the level of sensitivity called for by the test objectives. The sensitivity of the test is a function of three values: α, β, and the maximum allowable sensory difference, expressed as either δ or P_d.

[4] The boldface numbers in parentheses refer to a list of references at the end of this standard.

8.2 Prior to conducting the test, select values for α, β, and δ or P_d. The following can be considered as general guidelines.

8.2.1 For α-risk, when testing for a difference: a statistically significant result at:

8.2.1.1 10 % to 5 % (0.10 to 0.05) indicates “slight” evidence that a difference was apparent;

8.2.1.2 5 % to 1 % (0.05 to 0.01) indicates “moderate” evidence that a difference was apparent;

8.2.1.3 1 % to 0.1 % (0.01 to 0.001) indicates “strong” evidence that a difference was apparent; and

8.2.1.4 Below 0.1 % (<0.001) indicates “very strong” evidence that a difference was apparent.

8.2.2 For α-risk, when testing for similarity: a statistically significant result at:

8.2.2.1 10 % to 5 % (0.10 to 0.05) indicates “slight” evidence that no meaningful difference was apparent;

8.2.2.2 5 % to 1 % (0.05 to 0.01) indicates “moderate” evidence that no meaningful difference was apparent;

8.2.2.3 1 % to 0.1 % (0.01 to 0.001) indicates “strong” evidence that no meaningful difference was apparent; and

8.2.2.4 Below 0.1 % (<0.001) indicates “very strong” evidence that no meaningful difference was apparent.

8.2.3 For δ and P_d, the value that defines a meaningful sensory difference is affected by several factors, such as the importance of the product in the company’s portfolio, the stage in the development process at which testing is being done, etc. ~~As a general guide, meaningful differences fall into three ranges:~~

Presently, there is no consensus on the values of δ or P_d that represent small, medium, and large sensory differences. However, based on input from researchers who use δ in discrimination testing, the following ranges are presented as general guidance:

8.2.3.1 A more risk-averse business unit might consider
$\delta \leq 0.5$ to be small values,
$0.5 < \delta \leq 1.0$ to be medium values, and
$\delta > 1.0$ to be large values.

8.2.3.2 A more risk-tolerant business unit might consider
$\delta \leq 1.0$ to be small values,
$1.0 < \delta \leq 1.5$ to be medium values, and
$\delta > 1.5$ to be large values.

8.2.3.3 Smaller values of δ are usually chosen in late-stage testing or for very important products in the company’s portfolio, whereas larger values are chosen, for example, in early-stage testing or for less important products.

8.2.3.4 ~~δ < 0.5 or~~ The value of P_d ~~< 20 % represent small values;~~ that corresponds to a value chosen for δ can be obtained, for example, by using a software like the **rescale** function in the R package sensR (**5**).

For example,

```
>library(sensR)
>rescale(d.prime=c(0.5,0.75,1.00,1.25,1.50), method="tetrad")
```

will yield the following output:

Estimates for the tetrad protocol:

pc	pd	d.prime
1 0.3777187	0.06657806	0.50
2 0.4290391	0.14355862	0.75
3 0.4938084	0.24071264	1.00
4 0.5663366	0.34950487	1.25
5 0.6409366	0.46140484	1.50

~~8.2.3.2 0.5 < δ < 1.0 or 20 % < P_d < 30 % represent medium sized values; and~~

~~8.2.3.3 δ > 1.0 or P_d > 30 % represent large values.~~

8.3 Having defined the required level of sensitivity for the test using 8.2, use Table A1.1 (when testing for a difference) or Table A1.2 (when testing for similarity) to determine the number of assessors necessary. Enter the appropriate table in the section corresponding to the selected value of δ or P_d and the column corresponding to the selected value of β. The minimum required number of assessors is found in the row corresponding to the selected value of α. Alternatively, Tables A1.1 and A1.2 can be used to develop a set of values for δ or P_d, α, and β that provide acceptable sensitivity while maintaining the number of assessors within practical limits. The approach is presented in detail in Ref (**56**). Software that performs the same calculations may also be used, for example by using the discrimSS or d.primeSS function in the R package sensR.

8.4 Often in practice, the number of assessors is determined by material conditions (for example, duration of the experiment, number of available assessors, quantity of product). Increasing the number of assessors increases the likelihood of detecting small values of δ or P_d.

9. Procedure

9.1 Prepare worksheets and scoresheets, either manually or using software designed for this purpose (see Appendix X3), in advance of the test so as to utilize an equal number of the six possible sequences of two products, A and B:

AABB	BBAA
ABAB	BABA
ABBA	BAAB

Distribute these at random among the assessors so that serving order is balanced.

9.2 Present each set of four samples simultaneously if possible, following the same spatial arrangement for each assessor. Within the set of four samples, assessors are typically allowed to make repeated evaluations, for example, retasting, of each sample as desired. If the conditions of the test require the prevention of repeat evaluations, for example, if samples are too large to serve simultaneously or leave an aftertaste, present the samples sequentially and do not allow repeated evaluations. In addition, if the samples change over time, for example, cereal with milk, samples should be tested sequentially (or consider using an alternative testing method).

9.3 Instruct the assessors to evaluate the four test samples in the order presented. The assessor should then group the four samples into two groups of two based on similarity. It is critical that the instructions to the assessors say, "Group the four samples into two groups of two based on similarity," and not, "Identify the two samples that are most similar to each other." The latter wording does not correctly represent the tetrad task the assessor is to perform. It should be confirmed that the assessors understand the instructions and the tetrad task in general, for example, when they are being familiarized with the mechanics of the test.

9.4 Each scoresheet should provide for a single group of samples. If a different set of products is to be evaluated by an assessor in a single session, the completed scoresheet and any remaining product from the evaluation just completed should be returned to the test administrator prior to receiving the subsequent set of test samples. The assessor cannot go back to any of the previous samples or change the verdict on any previous test.

9.5 Do not ask questions about preference, acceptance, or degree of difference after the initial grouping of samples into pairs. The selection the assessor has just made may bias the reply to any additional questions. Responses to such questions may be obtained through separate tests for preference, acceptance, degree of difference, etc. (see Manual 26) (**67**). A comment section asking why the choice was made may be included for the assessor's remarks.

9.6 The tetrad test is a forced-choice procedure; assessors are not allowed the option of reporting "no difference." An assessor who detects no difference between the samples and requests to report "no difference," should be instructed to group the test samples into two pairs randomly. In such situations the assessor can indicate that the selection was only a guess in the comments section of the scoresheet.

10. Analysis and Interpretation of Results

10.1 Prior to conducting the test decide whether the objective of the test is to determine that the products are perceptibly different or that the products are sufficiently similar to be used interchangeably.