ASTM INTERNATIONAL

Designation: E3009 – 23

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

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

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, $P_c = 1/3$ in the tetrad test; while if the products are perceptibly different, $P_c > 1/3$.

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, δ or P_d (that is, a meaningful 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. 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.

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.

---

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

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:

8.2.3.1 $\delta < 0.5$ or $P_d < 20$ % represent small values;

8.2.3.2 $0.5 < \delta < 1.0$ or 20 % $< P_d < 30$ % represent medium sized values; and

8.2.3.3 $\delta > 1.0$ 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 (5). 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) (6). 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.

10.1.1 If the objective is to determine that the products are perceptibly different (that is, testing for a difference), then the null and alternative hypotheses for the test are:

H_O: δ or $P_d = 0$ (that is, $P_c = \frac{1}{3}$) versus H_A: δ or $P_d > 0$ (that is, $P_c > \frac{1}{3}$).

10.1.2 If the objective is to determine that the products are sufficiently similar to use interchangeably (that is, testing for similarity), then the null and alternative hypotheses for the test are (assuming that the value chosen to represent a meaningful difference has been specified as d):

10.1.2.1 H_O: δ or $P_d > $ d versus H_A: δ or $P_d \leq d$.

10.2 Use Table A1.3 to analyze the data obtained from a tetrad test when testing for a difference. If the number of correct responses is greater than or equal to the number given in Table A1.3, conclude that a perceptible difference exists between the samples. Failure to obtain a significant difference using Table A1.3 does not allow the researcher to conclude that the samples are significantly similar. Instead, when testing for similarity, use Table A1.4 to analyze the data from a tetrad test. If the number of correct answers is less than or equal to the number given in Table A1.4, conclude that the samples are sufficiently similar to be used interchangeably. Again, the conclusions are based on the risks accepted when the level of sensitivity (that is, α, β, and δ or P_d) was selected.

11. Report

11.1 Report the test objective, the results, and the conclusions. The following additional information is recommended:

11.1.1 The purpose of the test and the nature of the treatment studied;

11.1.2 *Full Identification of the Samples*—Origin, age, lot number, packaging, where obtained, method of preparation, quantity, shape, storage prior to testing, serving size, temperature. (Sample information should communicate that all storage, handling, and preparation was done in such a way as to yield samples that differ only due to the variable of interest, if at all.);

11.1.3 The number of assessors, the number of correct selections, and the result of the statistical evaluation;

11.1.4 *Assessors*—Age, gender, experience in sensory testing, experience with the product category, experience with the samples in the test;

11.1.5 Any information and any specific instructions given to the assessor in connection with the test;

11.1.6 The test environment: use of booths, simultaneous or sequential presentation, light conditions, whether the identity of the samples was disclosed after the test, and the manner in which it was done; and

11.1.7 The location and date of the test and the name of the test administrator.

12. Precision and Bias

12.1 Because results of sensory difference tests are functions of individual sensitivities, a general statement regarding the precision of results that is applicable to all populations of assessors cannot be made. However, adherence to the recommendations stated in this standard should increase the reproducibility of results and minimize bias.

13. Keywords

13.1 difference testing; discrimination test; sensory analysis; similarity testing; tetrad test

---

ANNEX

(Mandatory Information)

A1. NUMBER OF ASSESSORS AND CORRECT RESPONSES NEEDED FOR A TETRAD TEST

A1.1 See Tables A1.1-A1.4.

TABLE A1.1 Number of Assessors Needed for Tetrad Test When Testing for a Difference

NOTE 1—Entries are the minimum number of assessors required to execute a tetrad test when testing for a difference with a pre-specified level of sensitivity determined by the values chosen for δ or P_d, α, and β. Enter the table in the section corresponding to the chosen value of δ or P_d and the column corresponding to the chosen value of β. Read the minimum number of assessors from the row corresponding to the chosen value of α.

α		β		
	0.30	0.20	0.10	0.05
δ = 0.50; P_d = 7 %				
0.20	226	339	529	728
0.10	380	532	767	1000
0.05	555	727	1004	1265
0.01	949	1169	1519	1831
0.001	1513	1792	2218	2603
δ = 0.75; P_d = 14 %				
0.20	50	75	120	163
0.10	89	119	175	223
0.05	125	158	224	280
0.01	211	257	336	408
0.001	334	396	491	576
δ = 1.00; P_d = 24 %				
0.20	20	28	47	61
0.10	33	46	62	81
0.05	43	61	82	103
0.01	77	97	123	149
0.001	125	143	181	212
δ = 1.25; P_d = 35 %				
0.20	12	17	25	28
0.10	15	22	30	43
0.05	23	30	40	48
0.01	40	47	59	72
0.001	61	69	88	102
δ = 1.5; P_d = 46 %				
0.20	7	9	12	17
0.10	10	15	20	25
0.05	15	16	23	30
0.01	23	26	33	40
0.001	36	41	50	57

TABLE A1.2 Number of Assessors Needed for Tetrad Test When Testing for Similarity

Note 1—Entries are the minimum number of assessors required to execute a tetrad test when testing for similarity with a pre-specified level of sensitivity determined by the values chosen for δ or P_d, α, and β. Enter the table in the section corresponding to the chosen value of δ or P_d and the column corresponding to the chosen value of β. Read the minimum number of assessors from the row corresponding to the chosen value of α.

	β			
α	0.30	0.20	0.10	0.05
δ = 0.5; P_d = 7 %				
0.20	227	339	532	727
0.10	383	529	767	1004
0.05	554	728	1000	1265
0.01	962	1181	1526	1842
0.001	1538	1819	2240	2618
δ = 0.5; P_d = 14 %				
0.20	53	75	119	158
0.10	90	120	175	224
0.05	125	163	223	280
0.01	218	261	337	407
0.001	347	406	495	581
δ = 0.5; P_d = 24 %				
0.20	19	28	46	61
0.10	33	47	62	82
0.05	47	61	81	103
0.01	79	100	125	147
0.001	125	146	181	213
δ = 0.5; P_d = 35 %				
0.20	11	17	22	30
0.10	16	25	30	40
0.05	22	28	43	48
0.01	41	47	59	74
0.001	61	72	89	103
δ = 0.5; P_d = 46 %				
0.20	8	9	15	16
0.10	11	12	20	23
0.05	15	17	25	30
0.01	22	28	33	42
0.001	36	39	51	58

TABLE A1.3 Number of Correct Responses Need for Significance in a Tetrad Test When Testing for a Difference

Note 1—Entries are the minimum number of correct responses required for significance at the stated α-level (that is, column) for the corresponding number of respondents, n (that is, row). Reject the null hypothesis assumption, H_0: There is no perceptible difference between the samples, if the number of correct responses is greater than or equal to the tabled value.

Note 2—For values of n not in the table, compute the missing entry as follows: Minimum number of correct responses (x) = nearest whole number greater than $= \frac{n}{3}+z\sqrt{\frac{2n}{9}}$, where z varies with the significance level as follows: 0.84 for α = 0.20; 1.28 for α = 0.10; 1.64 for α = 0.05; 2.33 for α = 0.01; 3.09 for α = 0.001.

n	α 0.2	0.1	0.05	0.01	0.001	n	α 0.2	0.1	0.05	0.01	0.001
						31	14	15	16	18	20
						32	14	15	16	18	20
3	3	3	3	—	—	33	14	15	17	18	21
4	3	4	4	—	—	34	15	16	17	19	21
5	4	4	4	5	—	35	15	16	17	19	22
6	4	5	5	6	—	36	15	17	18	20	22
7	4	5	5	6	7	42	18	19	20	22	25
8	5	5	6	7	8	48	20	21	22	25	27
9	5	6	6	7	8	54	22	23	25	27	30
10	6	6	7	8	9	60	24	26	27	30	33
11	6	7	7	8	10	66	26	28	29	32	35
12	6	7	8	9	10	72	28	30	32	34	38
13	7	8	8	9	11	78	30	32	34	37	40
14	7	8	9	10	11	84	33	35	36	39	43
15	8	8	9	10	12	90	35	37	38	42	45
16	8	9	9	11	12	96	37	39	41	44	48
17	8	9	10	11	13	102	39	41	43	46	50
18	9	10	10	12	13	108	41	43	45	49	53
19	9	10	11	12	14	114	43	45	47	51	55
20	9	10	11	13	14	120	45	48	50	53	57
21	10	11	12	13	15	126	47	50	52	56	60
22	10	11	12	14	15	132	50	52	54	58	62
23	11	12	12	14	16	138	52	54	56	60	64
24	11	12	13	15	16	144	54	56	58	62	67
25	11	12	13	15	17	150	56	58	61	65	69
26	12	13	14	15	17	156	58	61	63	67	72
27	12	13	14	16	18	162	60	63	65	69	74
28	12	14	15	16	18	168	62	65	67	71	76
29	13	14	15	17	19	174	64	67	69	74	79
30	13	14	15	17	19	180	66	69	71	76	81