Designation: E2139 – 05 (Reapproved 2011)

Standard Test Method for Same-Different Test[1]

This standard is issued under the fixed designation E2139; 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 describes a procedure for comparing two products.

1.2 This test method does not describe the Thurstonian modeling approach to this test.

1.3 This test method is sometimes referred to as the simple-difference test.

1.4 A same-different test determines whether two products are perceived to be the same or different overall.

1.5 The procedure of the test described in this test method consists of presenting a single pair of samples to each assessor. The presentation of multiple pairs would require different statistical treatment and it is outside of the scope of this test method.

1.6 This test method is not attribute-specific, unlike the directional difference test.

1.7 This test method is not intended to determine the magnitude of the difference; however, statistical methods may be used to estimate the size of the difference.

1.8 This test method may be chosen over the triangle or duo-trio tests where sensory fatigue or carry-over are a concern, or where a simpler task is needed.

1.9 *This standard may involve hazardous materials, operations, and equipment. 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 and health practices and determine the applicability of regulatory limitations prior to use.*

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

2.2 *ASTM Publications:*[2]

Manual 26 Sensory Testing Methods, 2nd Edition

STP 758 Guidelines for the Selection and Training of Sensory Panel Members

STP 913 Guidelines for Physical Requirements for Sensory Evaluation Laboratories

2.3 *ISO Standard:*[3]

ISO 5495 Sensory Analysis—Methodology—Paired Comparison

3. Terminology

3.1 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 (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 (also known as Type II Error).

3.2.3 *chi-square test*—statistical test used to test hypotheses on frequency counts and proportions.

3.2.4 *Δ (delta)*—test sensitivity parameter established prior to testing and used along with the selected values of α, β, and an estimated value of p_1 to determine the number of assessors needed in a study. Delta (Δ) is the minimum difference in proportions that the researcher wants to detect, where the difference is $\Delta = p_2 - p_1$. Δ is not a standard measure of sensory difference. The same value of Δ may correspond to different sensory differences for different values of p_1 (see 9.5 for an example).

3.2.5 *Fisher's Exact Test (FET)*—statistical test of the equality of two independent binomial proportions.

[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 Fundamentals of Sensory.

Current edition approved Aug. 1, 2011. Published August 2011. Originally approved in 2005. Last previous edition approved in 2005 as E2139–05. DOI: 10.1520/E2139-05R11.

[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 American National Standards Institute (ANSI), 25 W. 43rd St., 4th Floor, New York, NY 10036, http://www.ansi.org.

3.2.6 p_1—proportion of assessors in the population who would respond *different* to the matched sample pair. Based on experience with using the same-different test and possibly with the same type of products, the user may have *a priori* knowledge about the value of p_1.

3.2.7 p_2—proportion of assessors in the population who would respond *different* to the unmatched sample pair.

3.2.8 *power 1-β (beta) risk*—probability of concluding that a perceptible difference exists when, in reality, one of size Δ does.

3.2.9 *product*—material to be evaluated.

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

3.2.11 *sensitivity*—term used to summarize the performance characteristics of this test. The sensitivity of the test is defined by the four values selected for α, β, p_1, and Δ.

4. Summary of Test Method

4.1 Clearly define the test objective in writing.

4.2 Choose the number of assessors based on the sensitivity desired for the test. The sensitivity of the test is in part related to two competing risks: the risk of declaring a difference when there is none (that is, α-risk), and the risk of not declaring a difference when there is one (that is, β-risk). Acceptable values of α and β vary depending on the test objective. The values should be agreed upon by all parties affected by the results of the test.

4.3 The two products of interest (A and B) are selected. Assessors are presented with one of four possible pairs of samples: A/A, B/B, A/B, and B/A. The total number of *same* pairs (A/A and B/B) usually equals the total number of *different* pairs (A/B and B/A). The assessor's task is to categorize the given pair of samples as *same* or *different*.

4.4 The data are summarized in a two-by-two table where the columns show the type of pair received (*same* or *different*) and the rows show the assessor's response (*same* or *different*). A Fisher's Exact Test (FET) is used to determine whether the samples are perceptibly different. Other statistical methods that approximate the FET can sometimes be used.

5. Significance and Use

5.1 This overall difference test method is used when the test objective is to determine whether a sensory difference exists or does not exist between two samples. It is also known as the simple difference test.

5.2 The test is appropriate in situations where samples have extreme intensities, give rapid sensory fatigue, have long lingering flavors, or cannot be consumed in large quantities, or a combination thereof.

5.3 The test is also appropriate for situations where the stimulus sites are limited to two (for example, two hands, each side of the face, two ears).

5.4 The test provides a measure of the bias where judges perceive two same products to be different.

5.5 The test has the advantage of being a simple and intuitive task.

6. Apparatus

6.1 Carry out the test under conditions that prevent contact between assessors until the evaluations have been completed, for example, booths that comply with STP 913.

6.2 For food and beverage tests, sample preparation and serving sizes should comply with Practice E1871, or see Refs (**1**) or (**2**).[4]

7. Definition of Hypotheses

7.1 This test can be characterized by a two-by-two table of probabilities according to the sample pair that the assessors in the population would receive and their responses, as follows:

		Assessor Would Receive	
		Matched Pair (AA or BB)	Unmatched Pair (AB or BA)
Assessor's Response	Same:	$1 - p_1$	$1 - p_2$
	Different:	p_1	$p_2 = (= p_1 + \Delta)$
	Total:	1	1

where p_1 and p_2 are the probabilities of responding *different* for those who would receive the matched pairs and the unmatched pairs, respectively.

7.2 To determine whether the samples are perceptibly different with a given sensitivity, the following one-sided statistical hypothesis is tested:

H_o: $p_1 = p_2$
H_a: $p_1 < p_2$

7.3 The hypothesis test can be expressed in terms of the minimum detectable difference Δ (H_o: Δ = 0 versus H_a: Δ > 0). Delta (Δ) will equal 0 and p_1 will equal p_2 if there is no detectable difference between the samples. This test addresses whether or not Δ is greater than 0. Thus, the hypothesis is one-sided because it is not of interest in this test to consider that responding *different* to the matched pair could be more likely than responding *different* to the unmatched pair.

8. Assessors

8.1 All assessors must be familiar with the mechanics of the same-different test (the format, the task, and the procedure of evaluation). Greater test sensitivity, if needed, may be achieved through selection of assessors who demonstrate above average individual sensitivity (see STP 758).

8.2 In order to perform this test, assessors do not require special sensory training on the samples in question. For example, they do not need to be able to recognize any specific attribute.

8.3 The assessors must be sampled from a homogeneous population that is well-defined. The population must be chosen on the basis of the test objective. Defining characteristics of the population can be, for example, training level, gender, experience with the product, and so forth.

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

9. Number of Assessors

9.1 Choose all the sensitivity parameters that are needed to choose the number of assessors for the test. Choose the α-risk and the β-risk. Based on experience, choose the expected value for p_1. Choose Δ, $p_2 - p_1$, the minimum difference in proportions that the researcher wants to detect. The most commonly used values for α-risk, β-risk, p_1 and Δ are α = 0.05, β = 0.20, p_1 = 0.3, and Δ = 0.3. These values can be adjusted on a case-by-case basis to reflect the sensitivity desired versus the number of assessors.

9.2 Having defined the required sensitivity (α-risk, β-risk, p_1, and Δ), determine the corresponding sample size from Table A1.1 (see Ref (**9**)). This is done by first finding the section of the table with a p_1 value corresponding to the proportion of assessors in the population who would respond *different* to the matched sample pair. Second, locate the total sample size from the intersection of the desired α, p_2 (or Δ), and β values. In the case of the most commonly used values listed in 9.1, Table A1.1 indicates that 84 assessors are needed. The sample size *n* is based on the number of same and different samples being equal The sample sizes listed are the total sample size rounded up to the nearest number evenly divisible by 4 since there are four possible combinations of the samples. To determine the number of same and different pairs to prepare, divide *n* by two.

9.3 If the user has no prior experience with the same-different test and has no specific expectation for the value of p_1, then two options are available. Either use p_1 = 0.3 and proceed as indicated in 9.2, or use the last section of Table A1.1. This section gives samples sizes that are the largest required, given α, β, and Δ, regardless of p_1.

9.4 Often in practice, the number of assessors is determined by practical conditions (for example, duration of the experiment, number of available assessors, quantity of product, and so forth) However, increasing the number of assessors increases the likelihood of detecting small differences. Thus, one should expect to use larger numbers of assessors when trying to demonstrate that products are similar compared to when one is trying to demonstrate that they are different.

9.4.1 When the number of assessors is fixed, the power of the test (1-β) may be calculated by establishing a value for p_1, defining the required sensitivity for α-risk and the Δ, locating the number of assessors nearest the fixed amount, and then following up the column to the listed β-risk.

9.5 If a researcher wants to be 90 % certain of detecting response proportions of p_2 = 60 % versus the expected p_1 = 40 % with an α-risk of 5 %, then Δ = 0.60 − 0.40 = 0.20 and β = 0.10 or 90 % power. The number of assessors needed in this case is 232 (Table A1.1). If a researcher wants to be 90 % certain of detecting response proportions of p_2 = 70 % versus the expected p_1 = 50 % with an α-risk of 5 %, then Δ = 0.70 − 0.50 = 0.20 and β = 0.10 or 90 % power. The number of assessors needed in this case is 224 (Table A1.1).

10. Procedure

10.1 Determine the number of assessors needed for the test as well as the population that they should represent (for example, assessors selected for a specific sensory sensitivity).

10.2 It is critical to the validity of the test that assessors cannot identify the samples from the way in which they are presented. One should avoid any subtle differences in temperature or appearance, especially color, caused by factors such as the time sequence of preparation. It may be possible to mask color differences using light filters, subdued illumination or colored vessels. Prepare samples out of sight and in an identical manner: same apparatus, same vessels, same quantities of product (see Practice E1871). The samples may be prepared in advance; however, this may not be possible for all types of products. It is essential that the samples cannot be recognized from the way they are presented.

10.3 Prepare serving order worksheet and ballot in advance of the test to ensure a balanced order of sample presentation of the two products, A and B. One of four possible pairs (A/A, B/B, A/B, and B/A) is assigned to each assessor. Make sure this assignment is done randomly. Design the test so that the number of *same* pairs equals the number of *different* pairs. The presentation order of the *different* pairs should be balanced as much as possible. Serving order worksheets should also include the identification of the samples for each set.

10.4 Prepare the response ballots in a way consistent with the product you are evaluating. For example, in a taste test, give the following instructions: *(1)* you will receive two samples. They may be the same or different; *(2)* evaluate the samples from left to right; and *(3)* determine whether they are the same or different.

10.4.1 The researcher can choose to add an instruction to the ballot indicating whether the assessor may re-evaluate the samples or not.

10.4.2 The ballot should also identify the assessor and date of test, as well as a ballot number that must be related to the sample set identification on the worksheet.

10.4.3 A section soliciting comments may be included following the initial forced-choice question.

10.4.4 The example of a ballot is provided in Fig. X2.2.

10.5 When possible, present both samples at the same time, along with the response ballot. In some instances, the samples may be presented sequentially if required by the type of product or the way they need to be presented, or both. This may be the case, for example, for the evaluation of a fragrance in a room where the assessor must change rooms to evaluate the second sample.

10.6 Collect all ballots and tabulate results for analysis.

11. Analysis and Interpretation of Results

11.1 The data from the test is summarized in a two-by-two table, as illustrated in the table below.

		Assessor Received Matched Pair (AA or BB)	Unmatched Pair (AB or BA)	Total
Assessor's Response	Same:	17	9	26
	Different:	13	21	34
	Total:	30	30	60

11.1.1 Before computing any test statistic, determine if the number of *different* responses from those who received the unmatched pair is less than or equal to the number of *different* responses from those who received the matched pair. If this is the case, conclude that the hypothesis of no difference cannot be rejected. If this is not the case, the computation of a test statistic is needed to determine whether the samples are perceptibly different or not.

11.2 Analyze the data using a Fisher's Exact Test **(3, 4, 5)**. The FET is widely available in industry standard software. See computation examples in X1.5.2 and X2.5.2.

11.3 Other statistical tests can also be used as an approximation to the FET, provided the data table is not sparse. A sparse table is defined as one that has at least one expected frequency less than 5. The expected frequency in row i and column j is computed as:

$$E_{ij} = \frac{(\text{Row } i \text{ Total})\ (\text{Column } j \text{ Total})}{(\text{Grand Total})} \quad (1)$$

11.3.1 For example, the expected frequency for Row 1: Column 1 (that is, same response on a matched pair) is:

$$E_{11} = \frac{(26)\ (30)}{60} = 13 \quad (2)$$

11.4 Available tests that approximate the FET include the one-tailed continuity corrected Chi-square (χ^2) **(6)**, the one-tailed non-continuity corrected Chi-square (χ^2) **(7)** and the z-test **(8)**.

11.4.1 In the case of either Chi-square test, compare the calculated statistic to the critical value of a χ^2 distribution with one degree of freedom and an α level of twice the desired level. The critical values for a number of α levels are given in Table 1. For example, the critical value for a 5 % a level is 2.71.

11.4.2 Computation examples of the one-tailed continuity, corrected Chi-square are given in X1.5.3 and X2.5.3.

11.4.3 In the case of a z-test, compare the calculated statistic to the one-tailed critical value of the z distribution for the chosen α level.

TABLE 1 Critical Values for a One Sided, 1 Degree of Freedom χ^2 Test

α Level	Critical Value (one sided[A] 1df χ^2)
0.01	5.41
0.05	2.71
0.1	1.64
0.2	0.708
0.3	0.275
0.4	0.0642

[A] A one sided value is obtained by using the χ^2 value corresponding to twice the desired a level.

12. Report

12.1 Report the test objective, the results, the conclusions, and the population to which they can be generalized. The following additional information is recommended:

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

12.1.2 Full identification of the samples: origin, method of preparation, quantity, shape, storage prior to testing, serving size, and temperature. (Sample information should communicate that all storage, handling, and preparation was done in such a way as to yield samples that differed only in the variable of interest);

12.1.3 The number of assessors, the number of selections of each sample, and the result of the statistical analysis;

12.1.4 Relevant assessor information such as age, gender, experience in sensory testing, and experience with the product and test samples. Provide all details necessary to clearly define the population represented by the assessors;

12.1.5 Any information or instructions given to the assessor in connection with the test;

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

12.1.7 The location and date of the test and name of the panel leader.

13. Precision and Bias

13.1 Because results of this test are a function 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 in this test method should increase the reproducibility of results and minimize bias.

14. Keywords

14.1 difference test; minimize carry-over; minimize sensory fatigue; sensory test for difference; two-sample sensory test

ANNEX

(Mandatory Information)

A1. NUMBER OF ASSESSORS REQUIRED FOR THE SAME-DIFFERENT TEST

A1.1 See Table A1.1.

TABLE A1.1 Number of Assessors Required for Same-Different Test Based on Fishers Exact Test (One-Tailed) (see Ref 9)

NOTE 1—Please note that this table is divided into sections based upon the value of p_1. The sample size specified for Δ in the table will apply only to that p_1; if p_1 changes, a different sample size may be needed even if the value of Δ remains the same.

NOTE 2—First, select the appropriate value for p_1 and then find the section of the table that corresponds to it. If you do not know your actual p_1 it is proposed that a value of $p_1 = 0.3$ is a reasonable generic starting point. Alternatively, you can use the last section of this table which gives sample sizes that are the largest required given α, β, and Δ.

NOTE 3—The values recorded in this table have been rounded to the nearest whole number evenly divisible by four to allow for equal presentation of all possible paired combinations of the same and different samples.

NOTE 4—The values in this table were determined by calculating the appropriate *N* divisible by 4 that is at least equal to the power (1-β) listed.

	$p_1 = 0.1$					β			
α	p_2	Δ	0.5	0.4	0.3	0.2	0.1	0.05	0.01
0.4	0.2	0.1	32	44	60	88	168	224	364
0.4	0.3	0.2	16	20	28	36	52	68	124
0.4	0.4	0.3	12	16	16	24	32	40	60
0.4	0.5	0.4	8	12	12	16	20	28	40
0.4	0.6	0.5	8	8	12	12	16	20	28
0.4	0.7	0.6	8	8	8	8	12	16	20
0.4	0.8	0.7	4	8	8	8	12	12	16
0.4	0.9	0.8	4	4	8	8	8	8	12
0.3	0.2	0.1	52	68	88	136	200	276	436
0.3	0.3	0.2	16	24	40	48	68	88	144
0.3	0.4	0.3	12	16	20	28	40	48	72
0.3	0.5	0.4	8	12	12	16	28	32	48
0.3	0.6	0.5	8	8	12	12	20	24	36
0.3	0.7	0.6	8	8	8	8	12	20	24
0.3	0.8	0.7	4	8	8	8	12	12	20
0.3	0.9	0.8	4	4	8	8	8	8	12
0.2	0.2	0.1	72	96	132	180	260	348	536
0.2	0.3	0.2	28	40	48	60	88	112	172
0.2	0.4	0.3	20	20	28	32	48	60	92
0.2	0.5	0.4	16	16	20	24	32	40	56
0.2	0.6	0.5	12	12	16	16	20	28	40
0.2	0.7	0.6	12	12	12	12	16	20	28
0.2	0.8	0.7	4	8	12	12	12	16	20
0.2	0.9	0.8	4	4	8	8	12	12	16
0.1	0.2	0.1	116	156	200	264	368	464	684
0.1	0.3	0.2	44	52	68	88	116	152	216
0.1	0.4	0.3	24	32	40	48	64	76	112
0.1	0.5	0.4	16	20	24	32	40	48	68
0.1	0.6	0.5	12	16	20	24	32	36	48
0.1	0.7	0.6	12	12	16	16	20	28	36
0.1	0.8	0.7	8	8	12	16	16	20	28
0.1	0.9	0.8	8	8	8	8	12	16	20
0.05	0.2	0.1	176	216	272	348	464	576	820
0.05	0.3	0.2	64	76	92	112	148	184	256
0.05	0.4	0.3	36	40	48	60	80	96	132
0.05	0.5	0.4	24	28	32	40	52	60	84
0.05	0.6	0.5	20	20	24	28	36	40	56
0.05	0.7	0.6	12	16	20	20	24	32	40
0.05	0.8	0.7	12	12	12	16	20	24	32
0.05	0.9	0.8	12	12	12	12	16	20	24
0.01	0.2	0.1	312	372	444	540	688	824	1116
0.01	0.3	0.2	104	120	144	172	216	260	344
0.01	0.4	0.3	56	68	76	92	112	136	176
0.01	0.5	0.4	36	40	48	60	72	84	112
0.01	0.6	0.5	28	28	36	40	48	60	76
0.01	0.7	0.6	20	24	28	32	36	44	56
0.01	0.8	0.7	16	20	20	24	28	32	40
0.01	0.9	0.8	12	12	16	20	20	24	32
	$p_1 = 0.2$					β			
α	p_2	Δ	0.5	0.4	0.3	0.2	0.1	0.05	0.01
0.4	0.3	0.1	32	48	68	136	212	292	532
0.4	0.4	0.2	16	20	28	40	60	100	156
0.4	0.5	0.3	12	16	20	24	32	44	80
0.4	0.6	0.4	8	12	12	16	24	28	44
0.4	0.7	0.5	8	8	12	12	16	20	32
0.4	0.8	0.6	8	8	8	12	12	16	20
0.4	0.9	0.7	4	8	8	8	12	12	16
0.3	0.3	0.1	52	88	120	180	272	392	636
0.3	0.4	0.2	24	32	40	56	92	116	192
0.3	0.5	0.3	12	20	24	32	44	56	92
0.3	0.6	0.4	8	12	12	24	28	36	52
0.3	0.7	0.5	8	8	12	12	20	28	36