This document is not an ASTM standard and is intended only to provide the user of an ASTM standard an indication of what changes have been made to the previous version. Because it may not be technically possible to adequately depict all changes accurately, ASTM recommends that users consult prior editions as appropriate. In all cases only the current version of the standard as published by ASTM is to be considered the official document.

~~Designation:E2164–01 (Reapproved 2007)~~ Designation: E 2164 – 08

Standard Test Method for Directional Difference Test[1]

This standard is issued under the fixed designation E 2164; 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 comparing two products using a two-alternative forced-choice task.

1.2 This method is sometimes referred to as a paired comparison test or as a 2-AFC (alternative forced choice) test.

1.3 A directional difference test determines whether a difference exists in the perceived intensity of a specified sensory attribute between two samples.

1.4 Directional difference testing is limited in its application to a specified sensory attribute and does not directly determine the magnitude of the difference for that specific attribute. Assessors must be able to recognize and understand the specified attribute. A lack of difference in the specified attribute does not imply that no overall difference exists.

1.5 This test method does not address preference.

1.6 A directional difference test is a simple task for assessors, and is used when sensory fatigue or carryover is a concern. The directional difference test does not exhibit the same level of fatigue, carryover, or adaptation as multiple sample tests such as triangle or duo-trio tests. For detail on comparisons among the various difference tests, see ~~referenceess.~~Ennis **(1)**~~,~~**, MacRae (2)**~~,~~ ~~**and**~~**, and O'Mahony and Odbert (3)**.[2]

1.7 The procedure of the test described in this document consists of presenting a single pair of samples to the assessors.

1.8 *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:* [3]

E 253 Terminology Relating to Sensory Evaluation of Materials and Products

E 456 Terminology Relating to Quality and Statistics

E 1871 Guide for Serving Protocol for Sensory Evaluation of Foods and Beverages

2.2 ~~*ASTM Publications:*~~

~~Manual 26Sensory Testing Methods, 2nd Edition~~

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

~~STP 913Guidelines. Physical Requirements for Sensory Evaluation Laboratories~~

~~2.3~~ *ISO Standard:*

ISO 5495 Sensory Analysis—Methodology—Paired Comparison

3. Terminology

3.1 For definition of terms relating to sensory analysis, see Terminology E 253, and for terms relating to statistics, see Terminology E 456.

3.2 *Definitions of Terms Specific to This Standard:*

3.2.1 α *(alpha) risk*—the 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*—the probability of concluding that no perceptible difference exists when, in reality, one does (also known as type II error).

3.2.3 *one-sided test*—a test in which the researcher has an a priori expectation concerning the direction of the difference. In this

[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 ~~Sept.~~March 1, ~~2007.~~2008. Published ~~January~~April 2008. Originally approved in 2001. Last previous edition approved in ~~2001~~2007 as E 2164 – 01 (2007).

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

[3] 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.

case, the alternative hypothesis will express that the perceived intensity of the specified sensory attribute is greater (that is, A>B) (or lower (that is, A<B)) for a product relative to the other.

3.2.4 *two-sided test*—a test in which the researcher does not have any a priori expectation concerning the direction of the difference. In this case, the alternative hypothesis will express that the perceived intensity of the specified sensory attribute is different from one product to the other (that is, A≠B).

3.2.5 *common responses*—for a one-sided test, the number of assessors selecting the sample expected to have a higher intensity of the specified sensory attribute. Common responses could also be defined in terms of lower intensity of the attribute if it is more relevant. For a two-sided test, the larger number of assessors selecting sample A or B.

3.2.6 P_{max}—A test sensitivity parameter established prior to testing and used along with the selected values of α and β to determine the number of assessors needed in a study. P_{max} is the proportion of common responses that the researcher wants the test to be able to detect with a probability of 1 β. For example, if a researcher wants to be 90 % confident of detecting a 60:40 split in a directional difference test, then P_{max}= 60% and β = 0.10. P_{max} is relative to a population of judges that has to be defined based on the characteristics of the panel used for the test. For instance, if the panel consists of trained assessors, P_{max} will be representative of a population of trained assessors, but not of consumers.

3.2.7 P_c—the proportion of common responses that is calculated from the test data.

3.2.8 *product*—the material to be evaluated.

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

3.2.10 *sensitivity*—a general term used to summarize the performance characteristics of the test. The sensitivity of the test is rigorously defined, in statistical terms, by the values selected for α, β, and P_{max}.

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, a function of two competing risks—the risk of declaring a difference in the attribute when there is none (that is, α-risk) and the risk of not declaring a difference in the attribute 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 In directional difference testing, assessors receive a pair of coded samples and are informed of the attribute to be evaluated. The assessors report which they believe to be higher or lower in intensity of the specified attribute, even if the selection is based only on a guess.

4.4 Results are tallied and significance determined by direct calculation or reference to a statistical table.

5. Significance and Use

5.1 The directional difference test determines with a given confidence level whether or not there is a perceivable difference in the intensity of a specified attribute between two samples, for example, when a change is made in an ingredient, a process, packaging, handling, or storage.

5.2 The directional difference test is inappropriate when evaluating products with sensory characteristics that are not easily specified, not commonly understood, or not known in advance. Other difference test methods such as the same-different test should be used.

5.3 A result of no significant difference in a specific attribute does not ensure that there are no differences between the two samples in other attributes or characteristics, nor does it indicate that the attribute is the same for both samples. It may merely indicate that the degree of difference is too low to be detected with the sensitivity (α, β, and P_{max}) chosen for the test.

5.3.1 The method itself does not change whether the purpose of the test is to determine that two samples are perceivably different versus that the samples are not perceivably different. Only the selected values of P_{max}, α, and β change. If the objective of the test is to determine if the two samples are perceivably different, then the value selected for α is typically smaller than the value selected for β. If the objective is to determine if no perceivable difference exists, then the value selected for β is typically smaller than the value selected for α and the value of P_{max} needs to be stated explicitly.

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 **(4)**.

6.2 Sample preparation and serving sizes should comply with Guide E 1871~~, or see Refs.~~ **, or see Herz and Cupchik (~~4~~5) ~~or~~ or Todrank et al (~~5)~~ 6).**

7. Assessors

7.1 All assessors must be familiar with the mechanics of the directional difference test (format, task, and procedure of evaluation). For directional difference testing, assessors must be able to recognize and quantify the specified attribute.

7.2 The characteristics of the assessors used define the scope of the conclusions. Experience and familiarity with the product or the attribute 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 for selecting assessors with increased sensitivity.

Consumers can be used, as long as they are familiar with the format of the directional difference test. If a sufficient number of employees are available for this test, they too can serve as assessors. If trained descriptive assessors are used, there should be sufficient numbers of them to meet the agreed-upon risks appropriate to the project. Mixing the types of assessors is not recommended, given the potential differences in sensitivity of each type of assessor.

7.3 The degree of training for directional difference testing should be addressed prior to test execution. Attribute-specific training may include a preliminary presentation of differing levels of the attribute, either shown external to the product or shown within the product, for example, as a solution or within a product formulation. 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 assessors' acuity for the taint (see STP 758 **(7)** for details). Allow adequate time between the exposure to the training samples and the actual test to avoid carryover or fatigue.

7.4 During the test sessions, avoid giving information about product identity, expected treatment effects, or individual performance until all testing is comple.

8. Number of Assessors

8.1 Choose the number of assessors to yield the sensitivity called for by the test objectives. The sensitivity of the test is a function of four factors: α-risk, β-risk, maximum allowable proportion of common responses (P_{max}), and whether the test is one-sided or two-sided.

8.2 Prior to conducting the test, decide if the test is one-sided or two-sided and select values for α, β, and P_{max}. The following can be considered as general guidelines:

8.2.1 One-sided versus two-sided: The test is one-sided if only one direction of difference is critical to the findings. For example, the test is one-sided if the objective is to confirm that the sample with more sugar is sweeter than the sample with less sugar. The test is two-sided if both possible directions of difference are important. For example, the test is two-sided if the objective of the test is to determine which of two samples is sweeter.

8.2.2 When testing for a difference, for example, when the researcher wants to take only a small chance of concluding that a difference exists when it does not, the most commonly used values for α-risk and β-risk are $\alpha = 0.05$ and $\beta = 0.20$. These values can be adjusted on a case-by-case basis to reflect the sensitivity desired versus the number of assessors available. When testing for a difference with a limited number of assessors, hold the α-risk at a relatively small value and allow the β-risk to increase to control the risk of falsely concluding that a difference is present.

8.2.3 When testing for similarity, for example, when the researcher wants to take only a small chance of missing a difference that is there, the most commonly used values for α-risk and β-risk are $\alpha = 0.20$ and $\beta = 0.05$. These values can be adjusted on a case-by-case basis to reflect the sensitivity desired vs. the number of assessors available. When testing for similarity with a limited number of assessors, hold the β-risk at a relatively small value and allow the α-risk to increase in order to control the risk of missing a difference that is present.

8.2.4 For P_{max}, the proportion of common responses falls into three ranges:

$P_{max} < 55\ \%$ represents "small" values;
$55\ \% \leq P_{max} \leq 65\ \%$ represents "medium-sized" values; and
$P_{max} > 65\ \%$ represents "large" values.

8.3 Having defined the required sensitivity for the test using 8.2, use Table 1 or Table 2 to determine the number of assessors necessary. Enter the table in the section corresponding to the selected value of P_{max} 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, Table 1 or Table 2 can be used to develop a set of values for P_{max}, α, and β that provide acceptable sensitivity while maintaining the number of assessors within practical limits.

8.4 Often in practice, the number of assessors is determined by material conditions (e.g., duration of the experiment, number of available assessors, quantity of sample). 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 samples are similar compared to when one is trying to show they are different.

9. Procedure

9.1 Prepare serving order worksheet and ballot in advance of the test to ensure a balanced order of sample presentation of the two samples, A and B. Balance the serving sequences AB and BA across all assessors. Serving order worksheets should also include complete sample identification information. See ~~Example~~ Appendix X1.

9.2 It is critical to the validity of the test that assessors cannot identify the samples from the way in which they are presented. For example, in a test evaluating flavor differences, one should avoid any subtle differences in temperature or appearance 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. Code the vessels containing the samples in a uniform manner using 3-digit numbers chosen at random for each test. Prepare samples out of sight and in an identical manner: same apparatus, same vessels, same quantities of sample (see Guide E 1871-91).

9.3 Present each pair of samples simultaneously whenever possible, following the same spatial arrangement for each assessor (on a line to be sampled always from left to right, or from front to back, etc.). Within the pair, assessors are typically allowed to

TABLE 1 Number of Assessors Needed for a Directional Difference Test One-Sided Alternative

NOTE 1—The values recorded in this table have been rounded to the nearest whole number evenly divisible by two to allow for equal presentation of both pair combinations (AB and BA).

NOTE 2—Adapted from Meilgaard et al **(8)**.

		β							
α		0.50	0.40	0.30	0.20	0.10	0.05	0.01	0.001
0.50	P_{max}=75 %	2	4	4	4	8	12	20	34
0.40		2	4	4	6	10	14	28	42
0.30		2	6	8	10	14	20	30	48
0.20		6	6	10	12	20	26	40	58
0.10		10	10	14	20	26	34	48	70
0.05		114	16	18	24	34	42	58	82
0.01		22	28	34	40	50	60	80	108
0.001		38	44	52	62	72	84	108	140
0.50	P_{max}=70 %	4	4	4	8	12	18	32	60
0.40		4	4	6	8	14	26	42	70
0.30		6	8	10	14	22	28	50	78
0.20		6	10	12	20	30	40	60	94
0.10		14	20	22	28	40	54	80	114
0.05		18	24	30	38	54	68	94	132
0.01		36	42	52	64	80	96	130	174
0.001		62	72	82	96	118	136	176	228
0.50	P_{max}=65 %	4	4	4	8	18	32	62	102
0.40		4	6	8	14	30	42	76	120
0.30		8	10	14	24	40	54	88	144
0.20		10	18	22	32	50	68	110	166
0.10		22	28	38	54	72	96	146	208
0.05		30	42	54	70	94	120	174	244
0.01		64	78	90	112	144	174	236	320
0.001		108	126	144	172	210	246	318	412
0.50	P_{max}=60 %	4	4	8	18	42	68	134	238
0.40		6	10	24	36	60	94	172	282
0.30		12	22	30	50	84	120	206	328
0.20		22	32	50	78	112	158	254	384
0.10		46	66	86	116	168	214	322	472
0.05		72	94	120	158	214	268	392	554
0.01		142	168	208	252	326	392	536	726
0.001		242	282	328	386	480	556	732	944
0.50	P_{max}=55 %	4	8	28	74	164	272	542	952
0.40		10	36	62	124	238	362	672	1124
0.30		30	72	118	200	334	480	810	1302
0.20		82	130	194	294	452	618	1006	1556
0.10		170	240	338	462	658	862	1310	1906
0.05		282	370	476	620	866	1092	1584	2238
0.01		550	666	820	1008	1302	1582	2170	2928
0.001		962	1126	1310	1552	1908	2248	2938	3812

NOTE—The values recorded in this table have been rounded to the nearest whole number evenly divisible by two to allow for equal presentation of both pair combinations (AB and BA).

make repeated evaluations of each sample as desired. If the conditions of the test require the prevention of repeat evaluations, for example, if samples are bulky, leave an aftertaste, or show slight differences in appearance that cannot be masked, present the samples monadically (or sequential monadic) and do not allow repeated evaluations.

9.4 Ask only one question about the samples. The selection the assessor has made on the initial question may bias the reply to subsequent questions about the samples. Responses to additional questions may be obtained through separate tests for preference, acceptance, degree of difference, ~~etc..~~etc. See ~~Manual 26~~Chambers and Baker Wolf **(9)**. A section soliciting comments may be included following the initial forced-choice question.

9.5 The directional difference 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 should be instructed to make a guess and select one of the samples, and can indicate in the comments section that the selection was only a guess.

10. Analysis and Interpretation of Results

10.1 The procedure used to analyze the results of a directional difference test depends on the number of assessors.

10.1.1 If the number of assessors is equal to or greater than the value given in Table 1 (for a one-sided alternative) or Table 2 (for a two-sided alternative) for the chosen values of α, β, and P_{max}, then use Table 3 to analyze the data obtained from a one-sided test and Table 4 to analyze the data from a two-sided test. If the number of common responses is equal to or greater than the number given in the table, conclude that a perceptible attribute difference exists between the samples. If the number of common responses is less than the number given in the table, conclude that the samples are similar in attribute intensity and that no more than P_{max} of the population would perceive the difference at a confidence level equal to 1-β. Again, the conclusions are based on the risks accepted when the sensitivity (that is, P_{max}, α, and β) was selected in determining the number of assessors.

TABLE 2 Number of Assessors Needed for a Directional Difference Test Two-Sided Alternative

NOTE 1—The values recorded in this table have been rounded to the nearest whole number evenly divisible by two to allow for equal presentation of both pair combinations (AB and BA).

NOTE 2—Adapted from Meilgaard et al **(8)**.

					β				
α		0.50	0.40	0.30	0.20	0.10	0.05	0.01	0.001
0.50	P_{max}=75 %	2	6	8	12	16	24	34	52
0.40		6	6	10	12	20	26	40	58
0.30		6	8	12	16	22	30	42	64
0.20		10	10	14	20	26	34	48	70
0.10		14	16	18	24	34	42	58	82
0.05		18	20	26	30	42	50	68	92
0.01		26	34	40	44	58	66	88	118
0.001		42	50	58	66	78	90	118	150
0.50	P_{max}=70 %	6	8	12	16	26	34	54	86
0.40		6	10	12	20	30	40	60	94
0.30		8	14	18	22	34	44	68	102
0.20		14	20	22	28	40	54	80	114
0.10		18	24	30	38	54	68	94	132
0.05		26	36	40	50	66	80	110	150
0.01		44	50	60	74	92	108	144	192
0.001		68	78	90	102	126	148	188	240
0.50	P_{max}=65 %	8	14	18	30	44	64	98	156
0.40		10	18	22	32	50	68	110	166
0.30		14	20	30	42	60	82	126	188
0.20		22	28	38	54	72	96	146	208
0.10		30	42	54	70	94	120	174	244
0.05		44	56	68	90	114	146	200	276
0.01		74	92	108	132	164	196	262	346
0.001		122	140	162	188	230	268	342	440
0.50	P_{max}=60 %	16	28	36	64	98	136	230	352
0.40		22	32	50	78	112	158	254	384
0.30		32	44	66	90	134	180	284	426
0.20		46	66	86	116	168	214	322	472
0.10		72	120	158	214	268	392	554	
0.05		102	126	158	200	264	328	456	636
0.01		172	204	242	292	374	446	596	796
0.001		276	318	364	426	520	604	782	1010
0.50	P_{max}=55 %	50	96	156	240	394	544	910	1424
0.40		82	130	194	294	452	618	1006	1556
0.30		110	174	254	360	550	722	1130	1702
0.20		170	240	338	462	658	862	1310	1906
0.10		282	370	476	620	866	1092	1584	2238
0.05		390	498	620	786	1056	1302	1834	2544
0.01		670	802	964	1168	1494	1782	2408	3204
0.001		1090	1260	1462	1708	2094	2440	3152	4064

NOTE—The values recorded in this table have been rounded to the nearest whole number evenly divisible by two to allow for equal presentation of both pair combinations (AB and BA).

10.1.2 If the number of assessors is less than the value given in Table 1 or Table 2 for the chosen values of α, β, and P_{max} and the researcher is primarily interested in testing for a difference, then use Table 3 to analyze the data obtained from a one-sided test or Table 4 to analyze the data obtained from a two-sided test. If the number of common responses is equal to or greater than the number given in the table, conclude that a perceptible attribute difference exists between the samples at the α-level of significance.

10.1.3 If the number of assessors is less than the value given in Table 1 or Table 2 for the chosen values of α, β, and P_{max} and the researcher is primarily interested in testing for similarity, then a one-sided confidence interval is used to analyze the data obtained from the test. The calculations are as follows:

$$P_c = c/n$$

$$S_c \text{ (standard error of } P_c\text{)} = \sqrt{P_c(1 - P_c)/n}$$

$$\text{confidence limit} = P_c + z_\beta S_c$$

where:

z_β = the one-sided critical value of the standard normal distribution, and
c = the number of common responses.

Values of z_β for some commonly used values of β-risk are:

TABLE 3 Number of Selected Responses Needed For Significance in a Directional Difference Test, One-Sided Alternative

NOTE—Entries are the minimum number of common responses required for significance at the stated significance level (column) for the corresponding number of assessors *n* (row). Reject the assumption of "no difference" 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 responses (x) = nearest whole number greater than $x = (n/2) + z\sqrt{n/4}$, 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; 3.10 for α = 0.001. This calculation is an approximation. The value obtained may differ from the exact value as presented in the table, but the difference never exceeds one response. Exact values can be obtained from binomial distribution functions widely available in statistical computer packages.

NOTE 3—Adapted from Meilgaard et al (**8**).

	Significance level (%)					
n	.50	.20	.10	.05	.01	.001
4	3	4	4	. . .	. . .	. . .
5	4	4	5	5	. . .	. . .
6	4	5	6	6	. . .	. . .
7	4	6	6	7	7	. . .
8	5	6	7	7	8	. . .
9	6	7	7	8	9	. . .
10	6	7	8	9	10	10
11	6	8	9	9	10	11
12	7	8	9	10	11	12
13	7	9	10	10	12	13
14	8	10	10	11	12	13
15	9	10	11	12	13	14
16	9	11	12	12	14	15
17	9	11	12	13	14	16
18	10	12	13	13	15	16
19	10	12	13	14	15	17
20	11	13	14	15	16	18
21	12	13	14	15	17	18
22	12	14	15	16	17	19
23	12	15	16	16	18	20
24	13	15	16	17	19	20
25	13	16	17	18	19	21
26	14	16	17	18	20	22
27	14	17	18	19	20	22
28	15	17	18	19	21	23
29	16	18	19	20	22	24
30	16	18	20	20	22	24
31	16	19	20	21	23	25
32	17	19	21	22	24	26
33	17	20	21	22	24	26
34	18	20	22	23	25	27
35	19	21	22	23	25	27
36	19	22	23	24	26	28
40	21	24	25	26	28	31
44	23	26	27	28	31	33
48	25	28	29	31	33	36
52	27	30	32	33	35	38
56	29	32	34	35	38	40
60	31	34	36	37	40	43
64	33	36	38	40	42	45
68	35	38	40	42	45	48
72	37	41	42	44	47	50
76	39	43	45	46	49	52
80	41	45	47	48	51	55
84	43	47	49	51	54	57
88	45	49	51	53	56	59
92	47	51	53	55	58	62
96	49	53	55	57	60	64
100	51	55	57	59	63	66

NOTE 1—For values of n not in the table, compute the missing entry as follows: Minimum number of responses (x) = nearest whole number greater than $x = (n/2) + z\sqrt{n/4}$, 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; 3.10 for α = 0.001. This calculation is an approximation. The value obtained may differ from the exact value as presented in the table, but the difference never exceeds one response. Exact values can be obtained from binomial distribution functions widely available in statistical computer packages.