Designation: ~~E3000 – 17~~ E3000 – 18

Standard Guide for Measuring and Tracking Performance of Assessors on a Descriptive Sensory Panel[1]

This standard is issued under the fixed designation E3000; 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 guide provides guidelines for measuring and tracking the performance of individual assessors on a descriptive sensory panel.

1.2 This guide provides guidelines to assist sensory professionals in measuring performance for given assessors. Measuring performance will form the basis for (*1*) determining the reliability of the results, and (*2*) establishing remedial actions for an individual assessor.

1.3 This guide examines various aspects of trained assessor performance; such as repeatability, discrimination, and agreement and demonstrates some ways to measure them. The procedures will help the sensory professional determine areas of good performance as well as those that require improvement.

1.4 Individual assessor performance is tracked using established statistical procedures. These procedures depend on whether replicates are collected and if they are collected over multiple sessions or within a single session.

1.5 This guide provides suggested procedures, including statistical procedures that can be done using standard statistical software, for evaluating performance and is not meant to exclude other methods that may be effectively used for a similar purpose.

1.6 Methods for training and screening assessors are not within the scope of this guide. This guide does not address how to communicate performance feedback information to individual assessors. This monitoring of panel reproducibility, a measure of the panel's ability to reproduce the results of other panels, is also not within the scope of this guide.

1.7 *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.*

1.8 *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

2.2 *Other Documents:*[2]

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

ASTM MNL13 Manual on Descriptive Analysis for Sensory Evaluation

2.3 *ISO Standards:*[3]

ISO 11132:2012 Sensory Analysis – Methodology—Guidelines for Monitoring the Performance of a Quantitative Sensory Panel

---

[1] This guide is under the jurisdiction of ASTM Committee E18 on Sensory Evaluation and is the direct responsibility of Subcommittee E18.03 on Sensory Theory and Statistics.

Current edition approved ~~Nov. 1, 2017~~April 1, 2018. Published ~~December 2017~~April 2018. Originally approved in 2017. Last previous edition approved in 2017 as E3000 – 17. DOI: ~~10.1520/E3000-17.~~10.1520/E3000-18.

[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), ISO Central Secretariat, BIBC II, Chemin de Blandonnet 8, CP 401, 1214 Vernier, Geneva, Switzerland, http://www.iso.org.

3. Terminology

3.1 Please refer to Terminologies E253 and E456, *ASTM STP 758,ASTM MNL13* and ISO 11132:2012 for any terms related to assessor performance that are not listed below.

3.2 *Definitions:*

3.2.1 *agreement*—ability of an assessor to give similar scores (rate) or to order the intensity of stimuli similarly to the rest of the panel (rank) on a given attribute.

3.2.2 *performance*—ability of an assessor to make repeatable assessments that are in agreement with other assessors on the panel and discriminate perceptible differences between attributes when they are present.

3.2.3 *scale usage*—the extent to which the assessor(s) uses the scale with respect to the intensities of the attributes being measured.

4. Summary of Practice

4.1 The protocols described in this guide provide a procedure for quantitatively establishing the performance of individual assessors by discussing the minimum level of good performance, determining when a performance problem exists, and detailing specific procedures to address those problems.

5. Significance and Use

5.1 This guide is meant to be used with and applied to individual trained descriptive assessors.

5.2 The procedures recommended in this guide can be used by the panel leader to periodically appraise the performance of individual descriptive assessors.

5.3 Tracking assessor performance will provide information as to the quality of the data being generated. Performance information may be used to decide whether to use the data to interpret product profiles.

5.4 Monitoring assessor performance will enable the panel leader to identify retraining needs or to identify assessors who are not performing well enough to continue participating on a panel.

6. Performance

6.1 *Introduction:*

6.1.1 This section provides sensory approaches for the assessment of assessor performance. It is assumed that good sensory practices are being followed in order to allow for good assessor performance. Panel members must be motivated to carry out the job conscientiously, be in good health both physically and mentally, and must be willing and able to follow instructions. Standard procedures to reduce random variability and systematic bias, including robust experimental design to reduce order and carry over effects must be followed by the sensory professional.

6.1.2 Assessor performance is the measure of the ability of an assessor to make reliable attribute assessments across the products being evaluated. It has been recognized as an important component of descriptive analysis since the method was first developed (**1, 2**). It can be measured at a given time point or tracked over time. Performance is compromised if an assessor cannot repeat their own results (repeatability), discriminate among the products (discrimination) and assess stimuli similarly to other assessors on a given attribute (agreement). This guide will focus on these three key measures. These measures can allow the sensory professional to diagnose sources of poor performance such as the inability to use a scale to correctly indicate intensity (scale usage), and failure to use attributes similarly to other assessors (lexicon usage).

6.1.3 It is important to track panel performance as a whole, since panel data are used for decision making; however, this guide describes how to measure and interpret performance criteria for the individual assessors, since assessor performance influences panel results. An assessor who does not discriminate among the samples may impact the panel data, causing the mean values for a specific attribute to be close together and preventing overall discrimination between samples. In some cases, a poorly performing assessor can cause panel data to be inconsistent and non-repeatable. All of the assessors must be using the vocabulary in the same way and utilizing the scales in a consistent manner in order for the panel to succeed.

6.2 *Individual Assessor Performance:*

6.2.1 In the early stages of training, performance evaluation should be analyzed for each individual assessor prior to participation in a panel. During this phase, the panel leader typically monitors panel agreement on ranking or rating stimuli for intensity and on scale usage. Specific examples for each attribute need to be introduced, experienced and defined to ensure that all assessors understand the sensory qualities and range of intensity of the attribute. During training, attribute definitions and references should be reviewed and possibly revised, to ensure that attributes are understood and used consistently by individual assessors across all samples. Assessors should be selected for continued panel participation based upon performance.

6.2.2 Once assessors are trained, they should be monitored for the three key measures of performance (repeatability, discrimination, agreement). It is important to evaluate assessor performance periodically in order to detect any change in individual performance over time and to identify an assessor or assessors who are not performing well. Assessor performance on an individual

study should be monitored if you are making high value or high risk business decisions with your panel data. The panel member should be engaged for a sufficient period to have established a history of performance which has been monitored.

6.2.3 In cases of poor performance, initially check with the individual assessor for any reasons they may not have been performing as usual. This would indicate the need to eliminate their responses on relevant data sets.

6.2.4 Rule out the possibility that assessor variation may be due to potential variability within the samples. Some types of products such as meat, seafood, or crop-based products can be quite variable and this variability must also be understood before concluding that there is an issue with assessor or panel performance.

6.2.5 Verification of test procedures, such as correct samples evaluated, correct instructions given, no data transcription errors, should also be done.

6.2.6 Fundamental issues such as insufficient training (issues with scale or lexicon usage), not understanding the procedure, boredom, over-use, or being unable to perceive certain attributes of the stimuli (physiological differences) can also contribute to poor assessor performance. It is important to identify early signs of performance inconsistency and correct the problem before the assessor has an impact on the overall panel's results. Additional training should be given as a part of panel maintenance to address these issues. By correcting the problem of the inconsistent assessor, one can achieve the aim of having a consistent panel.

6.3 *Key Measures of Performance*—There are three main elements of poor performance—lack of repeatability, inability to discriminate, and lack of agreement—that should be examined regularly.

6.3.1 *Repeatability*—Lack of repeatability occurs when assessor(s) cannot replicate their ratings from one evaluation to another of the same sample. It should be noted that assessment of repeatability is only possible if assessors evaluate the same sample on at least two occasions, either during the same session or on different sessions.

6.3.1.1 Inadequate training, inconsistent scale usage, lexicon usage, and various psychological and physiological factors can impair repeatability. Assessor fatigue, improper spacing of samples, poor instructions, inconsistent reference samples, and sample variation can also contribute to the problem. Study redesign or retraining, or both, may be necessary to reduce the variation in repeatability.

6.3.2 *Discrimination*—An assessor's inability to find significant differences among samples that are found to be different by the panel as a whole may occur for several reasons, such as general or specific ageusia and anosmia or differences in lexicon usage.

6.3.2.1 Using the same ratings across all samples for an attribute may indicate low sensory acuity resulting in the assessor's inability to use the scale as they were trained. Poorly discriminating assessors may use similar ratings across all samples in a "safe scale range" to cover their inability to discern the attribute.

6.3.2.2 The non-discriminating attributes should be identified and training provided to the assessor on those attributes for which the samples are expected to differ. It may be necessary to change the reference standards to better represent the attribute if previously used references are not helpful for the panel.

6.3.3 *Agreement*—Agreement is obtained when assessors rate samples similarly in relation to each other. Similar ratings indicate that the assessors are scoring the samples consistently for each attribute.

6.3.3.1 The data set should be carefully examined to determine which individual assessors contribute to the dissimilarity of the attribute ratings. A lack of agreement may be due to a difference in the assessors' discrimination, differences in scale or lexicon usage, or both.

6.3.3.2 Sometimes the main cause of a lack of agreement may not be due to a poor assessor, but rather, an assessor who may be more discriminating or more sensitive to an attribute. The identification of the origin of a disagreement is therefore essential for identification of the appropriate corrective action.

6.3.3.3 Assessors who vary on the perceived intensity in relation to other assessors, but still show the same sample ranking pattern as the other assessors (magnitude type interactions), usually differ in scale usage. A disagreement in assessor ratings may also indicate that assessors do not associate the same sensory perception with the attributes or vary on the perceived intensity due to individual differences in sensory acuity, thus causing cross-over interactions. A cross-over interaction occurs when an assessor's mean score for a specific sample is reversed in response pattern from those of other panel members. All cross-over interactions should be carefully examined since they reduce chances of the panel finding significant sample differences.

6.3.3.4 Lexicon usage may also contribute to agreement issues when one assessor understands the attribute to mean something different from the other panel members.

6.4 *Performance Diagnostics*—Scale usage and lexicon usage are two diagnostics that can be examined to understand what is causing issues with agreement, discrimination, and repeatability.

6.4.1 *Scale Usage*—Inconsistent scale usage occurs when different assessors use different ranges of the scale and also different areas or locations of the scale while rating the same sample (note: this is an assessor effect in ANOVA). Inconsistent scale usage for an overall panel can be considered acceptable to a certain degree as long as assessors are consistent with their own behavior (across all samples) and are in agreement with the rest of the panel (for example, rank the samples in the same order). Poor assessor calibration, inadequate training, insensitivity or super-sensitivity to the problem attribute, or lack of reference standards is usually the source of the inconsistent scale usage.

6.4.2 *Lexicon Usage*—Correct lexicon usage is the ability of an assessor(s) to understand and use attributes in a similar manner. It is important that each attribute being assessed has a definition that is precise and clearly understood by the assessor. A discussion during panel training can uncover inconsistent lexicon usage. References should also be developed that supports the attribute

definition and provides clarity to the assessor. An assessor who is having issues with lexicon usage should be given the opportunity to review the definitions and references during a training session.

6.5 *Procedure to Evaluate Assessor Performance:*

6.5.1 Follow the statistical procedures outlined in Section 7 of this guide to analyze the performance of the assessor for a single session over time. Evaluation of an assessor's performance should involve, at a minimum, the examination of performance data for potential issues with repeatability, discrimination, and agreement.

6.5.2 Historical data enable the panel leader to review assessors' performance over time. By tracking performance over time the panel leader can identify patterns of agreement or disagreement across assessors, and recognize improvement or deterioration of discrimination over time for individual assessors and for the panel as a whole.

6.5.3 Decide what corrective action (for example, further training, ad hoc deletion of data or assessor, or both) is required for the assessor based on their performance results. Refer to Section 9 Corrective Action for more information.

7. Procedure and Statistics for Evaluating Assessor Performance

7.1 This section outlines a procedure for evaluating assessor performance. It covers different statistical methods commonly used to calculate or visually inspect each performance measure including repeatability, discrimination, and agreement. Table 1 summarizes the statistical process for evaluating assessor performance. This section does not give exact details on how to calculate each measure but rather describes the statistics and how to use them.

7.2 All the listed techniques are available through statistical and graphical computer software packages. Other methodologies can be used; refer to the Bibliography for suggestions. This guide assumes a sufficient level of statistical knowledge to run the suggested statistical procedures. If you are not familiar with how to run these statistical procedures, please consult a statistician or a relevant textbook. More advanced assessor performance statistics can be done with specialized assessor performance or statistical software packages (see for example, Naes et al. (~~1~~3) and www.panelcheck.com).

TABLE 1 Statistical Procedure for Evaluating Assessor Performance

Key Steps	Statistics
Step 1: Initial data check Initial data check and validation to confirm that data for the correct samples were entered, that the data set is complete, and to identify and correct any obvious data entry and transcription errors.	1. CHECK: Raw data
Step 2: Assessor Agreement (initial check) and Repeatability Check assessor **repeatability:** how consistent are they?	2. CALCULATE: Mean CHECK: for Assessor agreement 3. CALCULATE: Standard Deviation CHECK: for Assessor repeatability 4. GRAPH: Individual assessors' data across the samples, one attribute per chart.
Step 3: ANOVA	5. Run appropriate 'assessor monitoring' ANOVA model in statistical software. The model used depends on whether data are replicated. See 7.6 for more details.
Step 4: Assessor Agreement Check **agreement** among assessors: Does the assessor agree with other assessors on the panel for each attribute?	6. CHECK: ANOVA Assessor main effect (α = 5 %) 7. CHECK: ANOVA Assessor*Sample interaction effect (α = 1 %) to determine agreement between assessors for each sample. 8. GRAPH: Generate an Assessor*Sample interaction graph for each attribute.
Step 5: Discrimination Check assessor **discrimination:** Can the assessors discriminate between samples?	9. For each assessor: CHECK ANOVA Sample main effect for each attribute (α = 5 %)

7.3 All statistical output used in Section 7 is based on an apple data set. The trained descriptive apple panel consisted of twelve assessors. This research project compared ten apple varieties (that is, samples) for ten flavor-related attributes (attributes are labelled with an 'F' prefix) and twelve texture-related attributes (attributes are labelled with a 'T' prefix). Assessors scored each of the ten samples on a 100 mm line scale. Each of the twelve assessors evaluated all ten samples in three separate sessions (that is, one replicate per session across three sessions). Refer to ASTM Research Report RR:E18-1001[4] for full data set details, including raw data and the full statistical output from the procedure described in this section.

7.4 *Initial Data Check and Validation*—The data should be checked to confirm that data for the correct samples were entered, that the data set is complete, that all obvious data entry and transcription errors were identified and corrected, and that the data will give a true representation of the samples. Knowledge of the samples is useful for checking that the sample means make sense and that the correct samples is useful for checking that the sample means make sense and that the correct samples were presented to assessors (for example, since Johnson's Red is a red apple does it have a higher red apple flavor intensity; Granny Smith is typically a sour apple, does it have a high sour intensity?). Step 1 can be done with both non-replicated and replicated data. It is assumed that replicated data are available for subsequent analysis steps.

7.4.1 *Raw Data*—Scan the raw data for each assessor to check for any obvious inconsistencies in the data. In Table 2, for Assessor 1, were Braeburn and Top Red samples swapped in Replicate 2? F_red apple and F_green apple scores are reversed for these two samples. This step is not specifically related to assessor performance. Instead, it is done to ensure that all subsequent analyses are performed on a data set for which all identifiable errors have been removed.

7.5 *Step 2. Assessor Agreement (Initial Check) and Repeatability:*

7.5.1 *Mean*—Calculate the mean for each assessor for each attribute (refer to Table 3 for examples of Sweet, Sour, and Bitter). Means can be calculated across samples or for each individual sample. It is recommended to use means in conjunction with their standard deviation to understand the basics of agreement. Calculating means across samples provides a gross measure of agreement among assessors. Assessors should have similar mean values. If not, scale usage and lexicon usage should be examined to uncover the source of the differences. The similarity of the assessor means across samples should be assessed by taking into account the significance of the Assessor Main Effect (see 7.7.2). When means are calculated for each assessor and sample separately, good agreement among the assessors is evidence by similar rank orders of the samples among the assessors. Alternatively, a graph of the assessors' means across the samples could be plotted. Good agreement among assessors is evidenced by similar patterns of

TABLE 2 Raw Apple Data, Assessor 1—Green Apple, Red Apple, and Sweet Attributes

Assessor	Apple	Replicate	F_Green apple	F_Red apple
1	Braeburn	1	53	0
1	Braeburn	2	0	37
1	Braeburn	3	57	7
1	Fuji	1	19	45
1	Fuji	2	41	14
1	Fuji	3	10	59
1	Gibson's Green	1	62	0
1	Gibson's Green	2	50	0
1	Gibson's Green	3	37	0
1	Golden Delicious	1	37	0
1	Golden Delicious	2	42	0
1	Golden Delicious	3	36	0
1	Granny Smith	1	71	0
1	Granny Smith	2	48	0
1	Granny Smith	3	48	0
1	Johnson's Red	1	0	70
1	Johnson's Red	2	0	80
1	Johnson's Red	3	0	52
1	Pink Lady	1	47	23
1	Pink Lady	2	29	34
1	Pink Lady	3	19	58
1	Royal Gala	1	0	45
1	Royal Gala	2	0	41
1	Royal Gala	3	0	51
1	Sun Gold	1	55	0
1	Sun Gold	2	47	28
1	Sun Gold	3	65	11
1	Top Red	1	0	42
1	Top Red	2	59	0
1	Top Red	3	0	48

[4] Supporting data have been filed at ASTM International Headquarters and may be obtained by requesting Research Report RR:E18-1001. Contact ASTM Customer Service at service@astm.org.

TABLE 3 Example of Calculated Means and Standard Deviation—Sweet, Acidic/Sour, and Bitter ~~(Individual Assessor and Sample Scores)~~(Atypically Large Standard Deviations are Highlighted in Red)

Assessor	Sample	F_Sweet Mean	F_Sweet Standard deviation	F_Acidic/sour Mean	F_Acidic/sour Standard deviation	F_Bitter Mean	F_Bitter Standard deviation
3	Braeburn	72.3	7.0	44.7	10.3	0.0	0.0
3	Fuji	49.3	26.5	48.0	26.0	8.7	15.0
3	Gibson's Green	53.0	36.4	62.7	20.8	17.0	29.4
3	Golden Delicious	56.0	28.8	36.3	22.5	25.0	38.2
3	Granny Smith	42.7	17.9	78.0	5.3	27.7	16.2
3	Johnson's Red	57.7	25.7	31.0	24.0	1.3	2.3
3	Pink Lady	42.0	12.5	66.3	6.4	23.7	4.5
3	Royal Gala	70.7	15.0	20.0	19.1	16.7	28.9
3	Sun Gold	29.0	7.2	68.0	9.5	14.0	13.2
3	Top Red	33.7	23.3	41.7	34.5	36.7	10.5
6	Braeburn	29.7	5.0	19.3	9.1	4.3	3.8
6	Fuji	42.7	23.2	2.3	4.0	0.7	1.2
6	Gibson's Green	34.3	8.1	29.3	10.1	4.3	2.1
6	Golden Delicious	38.0	3.6	6.7	6.1	12.0	11.5
6	Granny Smith	30.0	9.5	38.0	3.6	10.3	9.3
6	Johnson's Red	37.0	7.2	0.0	0.0	17.0	7.9
6	Pink Lady	44.3	8.1	46.7	8.5	3.0	5.2
6	Royal Gala	32.0	12.3	11.0	3.6	10.3	8.7
6	Sun Gold	41.3	9.5	35.3	3.2	6.0	3.0
6	Top Red	25.7	9.3	0.0	0.0	20.0	29.6
7	Braeburn	10.0	9.5	19.7	13.6	4.0	6.1
7	Fuji	11.7	2.1	16.7	14.0	4.3	4.0
7	Gibson's Green	7.7	7.2	13.3	7.4	5.0	4.6
7	Golden Delicious	11.3	8.1	13.7	2.5	1.0	1.7
7	Granny Smith	8.3	1.2	30.3	6.5	4.3	4.5
7	Johnson's Red	23.0	6.1	6.0	7.0	2.3	2.1
7	Pink Lady	44.0	7.8	25.3	3.1	0.0	0.0
7	Royal Gala	18.0	6.9	12.7	9.0	1.0	1.7
7	Sun Gold	19.7	14.2	26.0	8.0	1.0	1.7
7	Top Red	9.7	1.5	12.3	10.0	15.3	7.6

sample-to-sample differences among all assessors. The similarity of the sample-to-sample differences among the assessors should be assessed by taking into account the significance of the Assessor*Sample Interaction Effect (see 7.7.3).

7.5.2 *Standard Deviation*—To assess the repeatability of the assessors, calculate the square root of the mean square for error from a two-way ANOVA (with sample and session as the effects) on each assessor's data for each attribute individually (refer to Table 3 for examples of Sweet, Sour, and Bitter). These pooled standard deviations provide measures of the repeatability of the assessors. All assessors should have approximately equal standard deviations. The data from assessors with extremely large or extremely small standard deviations (compared to the rest of the assessors) should be examined to determine the cause of the excessively low or excessively high level of repeatability. Sensory panel data do not provide sufficient sample sizes for sensitive tests for differences among the assessors' standard deviations, so determination of what represents an extremely large or extremely small standard deviations needs to be acquired through experience with the analysis of many sets of sensory panel data.

7.5.2.1 In the apple example presented in this guide, a standard deviation greater than 10 % of the range of the scale (in this case, a standard deviation of 10 on the 100 point intensity scale) was chosen as the action standard to define a high lack of repeatability. Standard deviations for Assessors 3, 6, and 7 that are greater than 10 % are highlighted in red in the Table 3.