Designation: E1499 – 16 (Reapproved 2023)

Standard Guide for Selection, Evaluation, and Training of Observers[1]

This standard is issued under the fixed designation E1499; 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 describes criteria and tests for selecting, evaluating, and training human visual-sensory observers for tasks involving the perception and scaling of properties and phenomena relating to appearance.

1.2 Examples of tests requiring the use of trained observers include but are not limited to those described in the following ASTM standards: on color, Practice D1535 and Practice E1360; on color difference, Practice D1729 and Test Method D2616; on gloss, Test Method D4449; on metamerism, Practice D4086; and on setting tolerances, Practice D3134.

1.3 *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.*

1.4 *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]

D1535 Practice for Specifying Color by the Munsell System

D1729 Practice for Visual Appraisal of Colors and Color Differences of Diffusely-Illuminated Opaque Materials

D2616 Test Method for Evaluation of Visual Color Difference With a Gray Scale

D3134 Practice for Establishing Color and Gloss Tolerances

D4086 Practice for Visual Evaluation of Metamerism

D4449 Test Method for Visual Evaluation of Gloss Differences Between Surfaces of Similar Appearance

E284 Terminology of Appearance

E1360 Practice for Specifying Color by Using the Optical Society of America Uniform Color Scales System

3. Terminology

3.1 *Definitions*—Definitions of appearance terms in Terminology E284 are applicable to this guide.

3.2 *Definitions of Terms Specific to This Standard:*

3.2.1 *appearance, n—in psychophysical studies*, perception in which the spectral and geometric aspects of a visual stimulus are integrated with its illuminating and viewing environment.

3.2.2 *observer, n*—one who judges visually, qualitatively or quantitatively, the content of one or more appearance attributes in each member of a set of objects or stimuli.

3.2.3 *scale, v*—to assess the content of one or more appearance attributes in the members of a set of stimuli.

3.2.3.1 *Discussion*—Alternatively, scales may be determined by assessing the difference in content of an attribute with respect to the differences in that attribute among the members of the set.

4. Summary of Guide

4.1 This guide provides descriptions of techniques and tests for the selection of candidates for observers for use in visual testing, for the evaluation of their capabilities in this field, and for their training to enhance these capabilities.

4.2 Discussion is provided of precautions required for the efficient use of observers in visual tests, including avoidance of overtaxing the observers and the control of test variables.

4.3 Other considerations of test design, including the numbers of observers and observations required and the precision of the visual results, are to be covered elsewhere.

5. Significance and Use

5.1 The term *appearance* (see 3.2.1) implies the essential presence of human visual observations. The results of visual observation involve not only the step of observing, accomplished by the eye, but also the inseparable step of interpretation in the brain. Instrumental test methods currently cannot duplicate this second step, and therefore can now only approximate, but not fully measure, appearance. Such instrumental measures of appearance properties are useful only to the

[1] This guide is under the jurisdiction of ASTM Committee E12 on Color and Appearance and is the direct responsibility of Subcommittee E12.11 on Visual Methods.

Current edition approved June 1, 2023. Published July 2023. Originally approved in 1992. Last previous edition approved in 2016 as E1499 – 16. DOI: 10.1520/E1499-23.

[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.

extent that they can be correlated to the results of visual observations by observers of the appearance phenomena being evaluated.

5.2 Almost invariably, too little attention has been paid to ensuring that the essential visual observations have been properly obtained to provide the basis for correlating visual and instrumental test results.

5.3 This guide provides the means for assessing observers, by outlining the requirements and tests for their selection, evaluation, and training. This guide should be useful to all experimenters designing or using visual test methods to provide either direct results in terms of the observation of appearance properties, or the experiments correlating such results with instrumental measures approximating the same appearance properties. The user is cautioned to avoid the substitution of validated vision tests with replicas of any kind, either printed, photographed or digitally displayed.

6. Selection and Evaluation of Observers

6.1 The process used for selecting observers depends a great deal upon the type of experiment being carried out, but should essentially evaluate the potential capability of the observer to execute a series of visual evaluation tasks (**1, 2**).[3] When these tasks involve appearance attributes, color or related spectral phenomena are often among the task subjects, and if instead geometric phenomena such as gloss are involved, many of the same considerations apply. Accordingly, the emphasis in this guide is upon selecting observers for color-related measurements. Thus, observers must be screened to rule out those with any color- or task-oriented deficiencies.

6.2 *Color Vision Tests* (**3**):

6.2.1 *Pseudoisochromatic Plates*—As a preliminary color vision test, a pseudoisochromatic-plate test should be administered to the candidate observers. The instructions and scoring techniques supplied by the manufacturer should be followed. In particular, the illumination level should be kept well within the photopic range (1000 1x is recommended as a minimum value) and the spectral quality of the illuminating source should be near that of north-sky daylight. Failure to identify correctly the required number of the plates in the test used should be considered grounds for dismissing the candidate observer.

6.2.2 *Farnsworth-Munsell 100 Hue Test*—The Farnsworth-Munsell 100 Hue Test (**4**) should next be administered to the candidate. While the pseudoisochromatic-plate tests isolate certain factors of color deficiency, the Farnsworth-Munsell 100 Hue Test measures color discrimination directly and in detail. This test was not designed strictly for pass-fail categorization of observers but is recommended as an adjunct test for the analysis of color defectives. (It is also useful as an observer evaluation test; see 6.3.1.) In the Farnsworth-Munsell 100 Hue Test, abnormal color vision is indicated by the observer's failure to place the test chips in correct order. The chips consist of 85 colored papers varying in hue at approximately constant value and chroma, and the observer's failure is usually by wide margins in one or more limited regions of the hue circle. The presence of such abnormal results of the test should be grounds for dismissing the candidate observer.

6.3 *Visual Acuity and Discrimination Tests*—Having determined that the candidate observers have normal color vision, it is next necessary to test their level of discrimination of small differences in color or another appearance attribute of interest.

6.3.1 *Farnsworth-Munsell 100 Hue Test*—Use of the Farnsworth-Munsell 100 Hue Test as a color-discrimination test does not require readministration of the test, but merely reexamination of the test results. For the purposes of assessing color (more precisely, hue) discrimination, the test results are examined for the presence of an approximately constant but significant error level in the arrangement of the test chips throughout the hue circle. This may be interpreted as an inability to discriminate the small color differences between neighboring chips. While a weakness of this type might, for example, interfere with an observer's ability to participate in threshold scaling experiments, the observer might still be competent to perform magnitude scaling of larger differences among specimens.

6.3.2 *Triangle Test*—This test is part of a series known as the Japanese Color Aptitude Test. The candidate observers are shown, one at a time, a series of 20 sets of three colored chips each. In each set, two of the chips are identical and the third is slightly different in color. The observer is asked to identify which one is different, the differences being so small that there is considerable uncertainty in the judgment. A lower than average score in this test indicates that the observer does not differentiate small differences well.

6.4 *Magnitude Scaling Tests:*

6.4.1 *Length Estimation*—A simple magnitude-scaling test may be devised to familiarize the observer with scaling procedures and the experimental task of matching a given anchor scale with a perceived difference in stimuli. In an example (**1**), the observer was asked to judge the apparent length of a line in comparison to the length of a standard line. The lines were drawn with a heavy black marker on 100 mm by 150 mm index cards, one line to each of 21 cards. The standard or anchor line, 125 mm long, was assigned a value of 10 units of length. The other 20 cards had lines of various lengths, both longer and shorter than the anchor line. The anchor and one test card were displayed side by side at a distance of 0.6 m. Of course, no rulers or other aids were allowed. The observer's task was to assign a value to the length of each line relative to that of 10 units assigned to the anchor. Means of assessing the data obtained from a test such as this are discussed by Lodge (**5**).

6.4.2 *Color Estimation*—Another set of 20 cards from the Japanese Color Aptitude Test may be used to assess the candidate observer's "feel" for the type of judgment required in magnitude scaling of an appearance attribute. These cards each contain three color chips in a horizontal row. The left-hand chip is identified with the value 1, and the right-hand chip with the value 10. The chip in the middle lies between these two ends on some color-attribute scale. The task is to assign a scale value

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

between 1 and 10 to the center chip. The color-attribute scales of hue, value, and chroma are used randomly in the set.

6.5 The results of the above magnitude-scaling tests should be compared to the observer's performance on the color-discrimination tests, particularly the Farnsworth-Munsell 100 Hue Test, to determine how observers with superior to normal color vision perform. Any significant disparities should be examined to see if they resulted from poor directions or improper viewing conditions. If this was not the case, the observer should be asked to repeat the judgments at another time. If the results are poor a second time, the conclusion may be drawn that the candidate does not have a substantial skill for this type of judgment, and this observer should be dismissed.

7. Training of Observers

7.1 The importance of following the steps of selection and evaluation of observers by training sessions using the exact conditions of the scaling experiment in accordance with Section 6 cannot be overemphasized. The details of these training sessions will depend on the experimental design being used and cannot be stated in general terms here. Examples are found in Refs (**1**) and (**2**). Only by training the observer under the actual experimental conditions to be used can that observer learn exactly what is expected in the task.

7.1.1 It is recommended that each observer make a dummy set of observations before each observing session. This has several advantages. The first few observations always exhibit more "noise," as the observer refamiliarizes himself with the task. Use of the same sample set for each of these preliminary observation sets allows consistency to be tested and puts the observer at ease. At least one such set of observations should be one from which the results are discarded. The experimenter should watch the observer during the first session to be sure that the instructions for the experiment are understood.

7.1.2 There appears to be virtually no information reported in Refs (**1 and 2**), (**6-9**) about several important aspects of observer training, such as how long the training sessions need to be, when they should occur, whether they should be repeated and at what intervals, and how the experimenter knows when the observer is adequately trained. A few comments on the number of observers required in a typical case are given in Ref (**6**), but this should be considered as part of the experimental design.

7.1.3 Specific comments on magnitude scaling are found in 6.4.

8. Precautions in Using Observers

8.1 The report (**7**) of a conference on color-discrimination psychophysics points out several areas in which care must be taken to provide optimum working conditions for the observers if their visual scaling results are to be reliable. Reference should also be made to the requirements for observers for other sensory testing procedures (**8**).

8.1.1 It is all too easy to overtax the capabilities of observers to make critical judgments over long periods of time. Sessions should be limited in length of time or number of observations, or both. It is not yet possible to place firm limits on these numbers or durations. Boredom as well as physical fatigue must be considered. Some experimenters suggest rewarding observers for consistent results in order to combat boredom and ensure attention and responsiveness. In any case, the performance of the observers should be reviewed to ensure that their level of accuracy does not degrade with increase in length of the observing session.

8.2 Care must be taken in the design of the experiments to ensure that all variables that might influence the results or make the visual judgments more difficult are controlled. Some parameters requiring control that might not be immediately obvious are the sizes of the test specimens (see, for example, Practice D1729), their proximity, and the nature of any border or dividing line between them; the nature and level of the illumination; the nature and luminance of the surround; and the absence of any distraction in the observer's field of vision, such as light reflected from glossy surfaces or brightly colored clothing.

8.2.1 The optimum levels of illumination for the judgment of surface colors given in Practice D1729 are far above those that can be obtained with video display units and similar devices. This unavoidable difference must be properly taken into account in the design of the experiments in which judgments of colors and color differences generated on such displays are required.

8.3 It is well known (**10**) that the visual judgment of color differences is affected if the state of adaptation of the observer's eyes is changed, since the sensitivity of the eye to color differences decreases for the colors corresponding to the adapting color. On examination of a color-difference pair under normal conditions (unless the pair specimens are unusually small) their mean color is the adapting color, and prolonged viewing can lead to adaptation to that color and a decrease in sensitivity to the color difference. It is therefore usual to view small color differences in quick glances, as they tend to appear less prominent on prolonged viewing.

8.4 In the report noted (**7**), it was also pointed out that the differences among stimuli to be scaled can be too large as well as too small. There appears to be an upper limit to the size of differences that the human visual system can scale, as well as a threshold, as is well known.

9. Precision and Bias

9.1 Matters leading to estimations of the precision and bias of visual measurements will depend strongly on the experimental design to be used, but should be explored and considered as an integral part of establishing that design. References (**1-10**) provide few examples in which this has been done.

10. Keywords

10.1 appearance; observers; training; visual examination-color