Designation: E 1808 – 96

Standard Guide for Designing and Conducting Visual Experiments[1]

This standard is issued under the fixed designation E 1808; 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 is intended to help the user decide on the type of viewing conditions, visual scaling methods, and analysis that should be used to obtain reliable visual data.

1.2 This guide is intended to illustrate the techniques that lead to visual observations that can be correlated with objective instrumental measurements of appearance attributes of objects. The establishment of both parts of such correlations is an objective of Committee E-12.

1.3 Among ASTM standards making use of visual observations are Practices D 1535, D 1729, D 3134, D 4086, and E 1478; Test Methods D 2616, D 3928, and D 4449; and Guide E 1499.

1.4 *This guide 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:*

D 1535 Practice for Specifying Color by the Munsell System[2]

D 1729 Practice for Visual Examination of Color Differences of Opaque Materials[2]

D 2616 Test Method for Evaluation of Visual Color Difference with a Gray Scale[2]

D 3134 Practice for Establishing Color and Gloss Tolerances[2]

D 3928 Test Method for Evaluation of Gloss or Sheen Uniformity[2]

D 4086 Practice for Visual Evaluation of Metamerism[2]

D 4449 Test Method for Visual Evaluation of Gloss Differences Between Surfaces of Similar Appearance[2]

E 284 Terminology of Appearance[2]

E 1478 Practice for Visual Color Evaluation of Transparent Sheet Materials[2]

E 1499 Guide to the Selection, Evaluation, and Training of Observers[2]

3. Terminology

3.1 The terms and definitions in Terminology E 284 are applicable to this guide.

3.2 *Definitions:*

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 stimuli.

3.2.3 *sample*, *n*—a small part or portion of a material or product intended to be representative of the whole.

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

3.2.4.1 *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.

3.2.5 *specimen*, *n*—a piece or portion of a sample used to make a test.

3.2.6 *stimulus*, *n*—any action or condition that has the potential for evoking a response.

3.3 *Definitions of Terms Specific to This Standard:*

3.3.1 *anchor*, *n*—the stimulus from which a just-perceptible difference is measured.

3.3.2 *anchor pair*, *n*—a pair of stimuli differing by a defined amount, to which the difference between two test stimuli is compared.

3.3.3 *interval scale*, *n*—a scale having equal intervals between elements.

3.3.3.1 *Discussion*—Logical operations such as greater-than, less-than, equal-to, and addition and subtraction can be performed with interval-scale data.

3.3.4 *law of comparative judgments*—an equation relating the proportion of times any stimulus is judged greater, according to some attribute, than any other stimulus in terms of just-perceptible differences.

3.3.5 *nominal scale*, *n*—scale in which items are scaled simply by name.

[1] This guide is under the jurisdiction of ASTM Committee E-12 on Appearance and is the direct responsibility of Subcommittee E12.11 on Visual Methods.
Current edition approved May 10, 1996. Published July 1996.

[2] *Annual Book of ASTM Standards*, Vol 06.01.

3.3.5.1 *Discussion*—Only naming can be performed with nominal-scale data.

3.3.6 *ordinal scale*, *n*—a scale in which elements are sorted in order based on more or less of a particular attribute.

3.3.6.1 *Discussion*—Logical operations such as greater-than, less-than, or equal-to can be performed with ordinal-scale data.

3.3.7 *psychometric function*, *n*—the function, typically sigmoidal, relating the probability of detecting a stimulus to the stimulus intensity.

3.3.8 *psychophysics*, *n*—the study of the functions relating the physical measurements of stimuli and the sensations and perceptions the stimuli evoke.

3.3.9 *ratio scale*, *n*—a scale which, in addition to the properties of other scales, has a meaningfully defined zero point.

3.3.9.1 *Discussion*—In addition to the logical operations performable with other types of data, multiplication and division can be performed with ratio-scale data.

3.3.10 *scale*, *n*—a defined arrangement of the elements of a set of stimuli or responses.

4. Summary of Guide

4.1 This guide provides an overview of experimental design and data analysis techniques for visual experiments. Carefully conducted visual experiments allow accurate quantitative evaluation of perceptual phenomena that are often thought of as being completely subjective. Such results can be of immense value in a wide variety of fields, including the formulation of colored materials and the evaluation of the perceived quality of products.

4.2 This guide includes a review of issues regarding the choice and design of viewing environments, an overview of various classes of visual experiments, and a review of experimental techniques for threshold, matching, and scaling experiments. It also reviews data reduction and analysis procedures. Three different threshold and matching techniques are explained, the methods of adjustment, limits, and constant stimuli. Perceptual scaling techniques reviewed include ranking, graphical rating, category scaling, paired comparisons, triadic combinations, partitioning, and magnitude estimation or production. Brief descriptions and examples, along with references to more detailed literature, are given on the appropriate types of data analysis for each experimental technique.

4.3 For reviews of topics in other than visual sensory testing within ASTM, see Refs **(1, 2)**.[3]

5. Viewing Conditions

5.1 *Light Source*—The illumination of the specimens in scaling experiments must be reproducible over the course of the experiments. To achieve this, it is essential to control both the spectral character and the amount of illumination closely in both space and time. Failure to accomplish this can seriously undermine the integrity of the experiments. The spectral power distribution of the illumination should be known or, if this is not possible, the light source should be identified as to type and manufacturer. Information such as daylight-corrected fluorescent light, warm-white fluorescent light, daylight-filtered incandescent light, incandescent light, etc., together with parameters such as correlated color temperature and color rendering index, if available, should be noted in the report of the experiment.

5.2 *Viewing Geometry*—Almost all specimens exhibit some degree of gonioapparent or goniochromatic variation; therefore the illuminating and viewing angles must be controlled and specified. This is particularly important in the study of specimens exhibiting gloss variations, textiles showing directionality, or gonioapparent (containing metallic or pearlescent pigments) or retroreflective specimens, among others. This control and specification can range from correct positioning of the source and observer and the elimination of any secondary light sources visible in the specimens, for the judgment of gloss specimens at and near the specular angle, to more elaborate procedures specifying a range of angles and aperture angles of illumination and viewing for gonioapparent and retroreflective specimens. When fluorescent specimens are studied, the spectral power distribution of the source must closely match that of a designated standard source.

5.3 *Surround and Ambient Field*—For critical visual scaling work, the surround, the portion of the visual field immediately surrounding the specimens, should have a color similar to that of the specimens. The ambient field, the field of view when the observer glances away from the specimens, should have a neutral color (Munsell Chroma less than 0.2) and a Munsell Value of N6 to N7 (luminous reflectance 29 to 42); see Practice D 1729).

5.4 *Observers*—Guide E 1499 describes the selection, evaluation, and training of observers for visual scaling work. Of particular importance is the testing of the observers' color vision and their color discrimination for normality. Color vision tests for this purpose are described in Guide E 1499.

6. Categories of Visual Experiments

6.1 Visual experiments tend to fall into two broad classes: (*1*) threshold and matching experiments designed to measure visual sensitivity to small changes in stimuli (or perceptual equality), and (*2*) scaling experiments intended to generate a psychophysical relationship between the perceptual and physical magnitudes of a stimulus. It is critical to determine first which class of experiment is appropriate for a given application.

6.1.1 *Threshold and Matching Experiments*—Threshold experiments are designed to determine the just-perceptible difference in a stimulus, or JPD. Threshold techniques are used to measure the observers' sensitivity to a given stimulus. Absolute thresholds are defined as the JPD for a change from no stimulus, while difference thresholds represent the JPD from a particular stimulus level greater than zero. The stimulus from which a difference threshold is measured is known as an anchor stimulus. Often, thresholds are measured with respect to the difference between two stimuli. In such cases, the difference of a pair of stimuli is compared to the difference in an anchor pair. Absolute thresholds are reported in terms of the physical units used to measure the stimulus, for example, a brightness

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

threshold might be measured in luminance units of candelas per square metre. Sensitivity is measured as the inverse of the threshold, since a low threshold implies high sensitivity. Threshold techniques are useful for defining visual tolerances, such as color-difference tolerances. Matching techniques are similar, except that the goal is to determine when two stimuli are not perceptibly different. Measures of the variability in matching can be used to estimate thresholds. Matching experiments provided the basis for CIE colorimetry through the metameric matches used to derive the color-matching functions of the CIE standard observers.

6.1.2 *Scaling Experiments*—Scaling experiments are intended to derive relationships between perceptual magnitudes and physical magnitudes of stimuli. Several decisions must be made, depending on the type and dimensionality of the scale required. It is important to identify the type of scale required and decide on the scaling method to be used before any scaling data are collected. This seems to be an obvious point, but in the rush to acquire data it is often overlooked, and later it may be found that the data obtained do not yield the answer required or cannot be used to perform desired mathematical operations. See Refs **(3, 4)** for further details. Scales are classified into the following four classes:

6.1.2.1 *Nominal Scales*—Nominal scales are relatively trivial in that they scale items simply by name. For color, a nominal scale might consist of reds, yellows, greens, blues, and neutrals. Scaling in this case would simply require deciding which color belonged in which category. Only naming can be performed with nominal data.

6.1.2.2 *Ordinal Scales*—Ordinal scales are scales in which elements are sorted in ascending or descending order based on more or less of a particular attribute. A box of multicolored crayons could be sorted by hue, and then in each hue family, say red, the crayons could be sorted from the lightest to the darkest. In a box of crayons the colors are not evenly spaced, so one might have, for example, three dark, one medium, and two light reds. If these colors were numbered from one to six in increasing lightness, an ordinal scale would be created. Note that there is no information on such a scale as to the magnitude of difference from one of the reds to another, and it is clear that they are not evenly spaced. For an ordinal scale, it is sufficient that the specimens be arranged in increasing or decreasing amounts of an attribute. The spacing between specimens can be large or small and can change up and down the scale. Logical operations such as greater-than, less-than, or equal-to can be performed with ordinal-scale data.

6.1.2.3 *Interval Scales*—Interval scales have equal intervals. On an interval scale, if a pair of specimens were separated by two units, and a second pair at some other point on the scale were also separated by two units, the differences between the pair members would appear equal. However, there is no meaningful zero point on an interval scale. A common example of an interval scale is the Celsius temperature scale. In addition to the mathematical operations listed for nominal and ordinal scales, addition and subtraction can be performed with interval-scale data.

6.1.2.4 *Ratio Scales*—Ratio scales have all the properties of the above scales plus a meaningfully defined zero point. Thus it is possible to equate ratios of numbers meaningfully with a ratio scale. Ratio scales are often impossible to obtain in visual work. An example of a ratio scale is the absolute, or Kelvin, temperature scale. All of the mathematical operations that can be performed on interval-scale data can also be performed on ratio-scale data, and in addition, multiplication and division can be performed.

7. Threshold and Matching Methods

7.1 Several basic types of threshold experiments are presented in this section in order of increasing complexity of design and utility of the data generated. Many modifications of these techniques have been developed for specific applications. Experimenters should strive to design an experiment that removes as much control of the results from the observers as possible, thus minimizing the influence of variable observer judgment criteria. Generally, this comes at the cost of implementing a more complicated experimental procedure.

7.1.1 *Method of Adjustment*—The method of adjustment is the simplest and most straightforward technique for deriving threshold data. In it, the observer controls the stimulus magnitude and adjusts it to a point that is just perceptible (absolute threshold) or just perceptibly different (difference threshold). The threshold is taken to be the mean setting across a number of trials by one or more observers. The method of adjustment has the advantage that it is quick and easy to implement. However, it has a major disadvantage in that the observer is in control of the stimulus. This can bias the results due to variability of observers' criteria and adaptation effects. If an observer approaches the threshold from above, adaptation might result in a higher threshold than if it were approached from below. Often the method of adjustment is used to obtain a first estimate of the threshold, to be used in the design of more sophisticated experiments. The method of adjustment is also commonly used in matching experiments.

7.1.2 *Method of Limits*—The method of limits is only slightly more complex than the method of adjustment. In the method of limits, the experimenter presents the stimuli at predefined discrete magnitude levels in either ascending or descending series. For an ascending series, the experimenter presents a stimulus, beginning with one that is certain to be imperceptible, and asks the observer if it is visible. If the observer responds no, the experimenter increases the stimulus magnitude and presents another trial. This continues until the observer responds yes. A descending series begins with a stimulus magnitude that is clearly perceptible and continues until the observer responds no, the stimulus cannot be perceived. The threshold is taken to be the average stimulus magnitude at which the transition between yes and no responses occurs for a number of ascending and descending series. Averaging over both types of series minimizes adaptation effects. However, the observers are still in control of their criteria since they can respond yes or no at their own discretion.

7.1.3 *Method of Constant Stimuli*—In the method of constant stimuli, the experimenter chooses several stimulus magnitude levels (usually five or seven) around the level of the threshold. These stimuli are each presented to the observer several times, in random order. The frequency, over the trials,

with which each stimulus is perceived is determined. From such data, a "frequency-of-seeing" curve, or psychometric function, can be derived that allows determination of the threshold and its uncertainty. The threshold is generally taken to be the stimulus magnitude at which it is perceived in 50 % of the trials. Psychometric functions can be derived for either a single observer (through multiple trials) or a population of observers (one or more trials per observer). Two types of response can be obtained: yes-no (or pass-fail) and forced choice.

7.1.3.1 *Yes-No Procedures*—In a yes-no or pass-fail method of constant stimuli procedure, the observers are asked to respond yes if they detect the stimulus (or stimulus change) and no if they do not. The psychometric function is the percent of yes responses as a function of stimulus magnitude. Fifty percent yes responses would be taken as the threshold level. Alternatively, this procedure can be used to measure visual tolerances above threshold by providing a reference stimulus magnitude (for example, a color-difference anchor pair) and asking the observers to pass stimuli that fall below the magnitude of the reference (have a smaller color difference than the anchor pair), and fail those that fall above it (have a larger color difference). The psychometric function is the percent of fail responses as a function of stimulus magnitude and the 50 % fail level is taken as the point of visual equality.

7.1.3.2 *Forced-Choice Procedures*—A forced-choice procedure eliminates the influence of varying observer criteria on the results, by presenting the stimulus in one of two intervals with a defined boundary between them. The observers are asked to indicate in which of the two intervals the stimulus was presented. They are not allowed to respond that the stimulus was not present in either interval, and are forced to guess which interval it was in if they are unsure, hence the name "forced choice". The psychometric function is the percent of correct responses as a function of stimulus magnitude. The psychometric function ranges from 50 % correct when the observers are simply guessing to 100 % correct for stimulus magnitudes at which the stimulus can always be detected. Thus the threshold is defined as the stimulus magnitude at which the observers are correct 75 % of the time and therefore detecting the stimulus 50 % of the time. As long as the observers respond honestly, their criteria, whether liberal or conservative, cannot influence the results.

7.1.3.3 *Staircase Procedures*—Staircase procedures are modifications of the forced-choice procedure designed to measure only the threshold point on the psychometric function. Staircase procedures are particularly applicable to situations in which the stimulus presentations can be fully automated. A stimulus is presented and the observer is asked to respond. If the response is correct, the same stimulus magnitude is presented again. If the response is incorrect, the stimulus magnitude is increased for the next trial. Generally, if the observer responds correctly on three consecutive trials, the stimulus magnitude is decreased. The stimulus magnitude steps are decreased until some desired precision in the threshold is reached. The sequence of 3-correct or 1-incorrect response prior to changing the stimulus magnitude results in convergence to a stimulus magnitude that is correctly identified in 79 % of the trials, very close to the nominal threshold of 75 %. Often several independent staircase procedures are run simultaneously to randomize the experiment further. A staircase procedure can also be run with yes-no or pass-fail responses.

8. Scaling Methods

8.1 *Dimensionality*—Scaling methods can be divided into two groups: unidimensional (one-dimensional) and multidimensional scaling.

8.1.1 *Unidimensional Scaling*—This method assumes that both the attribute to be scaled and the physical variation of the stimulus are unidimensional. The observers are asked to make their judgments on a single perceptual attribute. In color work, common examples include judging the color difference in a pair of specimens or judging the lightness of one specimen relative to that of another in a series of colors in which hue and chroma are constant.

8.1.1.1 *Cross-Modality Scaling*—It is also possible in color work to judge one attribute of a pair of specimens but express the results in terms of another attribute, displayed on a scale made up of anchor pairs. An example is the use of a gray scale, in which differences in total color difference, or chroma, or hue are judged by comparison to anchor pairs presented in the form of gray-scale pairs, in which the variable attribute is lightness (see Test Method D 2616).

8.1.2 *Multidimensional Scaling*—This method of scaling is similar to unidimensional scaling but it does not make the assumption that a single attribute is to be scaled. The dimensionality of the experiment is found as part of the analysis. In multidimensional scaling the data are interval or ordinal scales of the similarities or dissimilarities between all possible pairs of stimuli and the resulting output is a multidimensional geometric configuration of the perceptual relationships among the stimuli. For example, the flying distances among a well-distributed sampling of USA cities can be used to reconstruct a map of the country (see 9.1.3.1 and 9.1.3.2).

8.2 *Scaling Methods*—A variety of scaling techniques has been devised. It is important to determine first the level of scale required, that is, nominal, ordinal, interval, or ratio, and then choose the technique that provides the simplest task for the observer while still generating data that can be used to derive the required scale.

8.2.1 *Rank Order*—Given a set of specimens, the observer is asked to arrange them according to increasing or decreasing magnitudes of a particular perceptual attribute. With a large number of observers, the data may be averaged and re-ranked to obtain an ordinal scale. To obtain an interval scale, certain assumptions about the data must be made and additional analyses performed. In general it is not recommended that one attempt to derive interval scales from rank-order data.

8.2.2 *Graphical Rating*—Graphical rating allows direct determination of an interval scale. Observers are presented stimuli and asked to indicate the magnitude of their perceptions on a unidimensional scale with fixed anchor points. For example, in a lightness scaling experiment a line might be drawn with one end labeled white and the other black. When the observers are presented with a medium gray specimen that is perceptually half way between white and black, they would make a mark on the line at the midpoint. If the specimen was