Designation: E3041 – 17

Standard Guide for Selecting and Using Scales for Sensory Evaluation[1]

This standard is issued under the fixed designation E3041; 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 The objective of this guide is to provide information to be reviewed and considered by the sensory and consumer scientist who wants to select and use scales to measure responses from consumers or trained assessors. For ease of reading, the term sensory scientist is used throughout the guide when statements apply to the sensory and consumer scientists.

1.2 This guide covers a brief definition of scales as well as some fundamental and practical challenges the sensory and consumer scientists should be aware of when using scales. It also provides a list and a description of the most commonly used scales in the field of sensory evaluation and consumer product research along with a classification framework for these scales.

1.3 The scope of this guide is limited to the sensory and consumer science professional's selection and use of rating scales when an assessor assigns one symbol/value to one stimulus, to their perception of a stimulus or an internal attitude/opinion. It does not cover:

1.3.1 Details of analysis of data obtained from rating scales,

1.3.2 Guidelines for questionnaire design including attribute selection,

1.3.3 Fundamentals of measurement such as reliability and validity,

1.3.4 Measurement models used to convert scale responses into measures of unobserved sensory or hedonic quantities,

1.3.5 Tasks in which the assessor assigns a symbol/value to a group of stimuli,

1.3.6 Rankings, and

1.3.7 Multi-item scales.

1.4 *Units*—The values stated in SI units are to be regarded as the standard. No other units of measurement are included in this standard.

1.5 *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.*

1.6 *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

E2299 Guide for Sensory Evaluation of Products by Children and Minors

3. Terminology

3.1 *Definitions*—See Terminology E253 for definitions relating to sensory evaluation of materials and products.

3.2 *Definitions of Terms Specific to This Standard:*

3.2.1 *interval data*—data obtained from a scale for which numerically identical differences on any part of the scale correspond to the same magnitude of difference.

3.2.1.1 *Discussion*—The occurrence of a zero point in interval data does not correspond to the complete absence of the characteristic measured. An example of interval data is a temperature in degrees Fahrenheit where each degree change is the same change in thermal heat regardless of point on the scale and 0°F does not represent the complete absence of thermal energy.

3.2.2 *ordered category scale, n*—rating instrument in which the categories used to encode the responses are ordered by magnitude.

3.2.3 *ordinal data, n*—data obtained when items are ordered with respect to magnitude, but the magnitudes of difference among successively ordered items are not necessarily equal.

3.2.3.1 *Discussion*—Examples include ranking, just-about-right scales, and the Likert scale.

[1] This guide 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 Dec. 1, 2017. Published January 2018. DOI: 10.1520/E3041-17.

[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.

3.2.4 *rating instrument, n*—collection of symbols/values, provided by the sensory scientist to the assessor along with instructions, from which the assessor chooses to communicate an affective, attitudinal, behavioral, or perceptual response to a stimulus.

3.2.4.1 *Discussion*—Examples of rating instruments include category scales, line scales, and list of CATA terms printed on paper or displayed on an electronic device.

3.2.5 *ratio data, n*—data obtained from a scale that has an absolute zero point and for which numerically identical differences on any part of the scale correspond to the same magnitude.

3.2.5.1 *Discussion*—An example of ratio data is a temperature on the Kelvin scale in which each degree change is the same change in thermal heat regardless of the point on the scale and zero represents the complete absence of thermal energy.

3.2.6 *scale, n*—(*1*) rating instrument, sometimes referred to as a rating scale, used to encode human responses to stimuli numerically, an example of which is an ordered category scale and (*2*) continuum on which perceptions are quantified with specified theoretical properties that depend on the type of scale, an example of which is an interval scale.

3.2.6.1 *Discussion*—In the sense of definition (*1*), while all scales have several characteristics in common, for example, all have at least two response options and all are used to encode responses to stimuli, scales differ in the amount of information they provide per data point. In evaluating the sweetness of products, for example, CATA (that is, yes/no) data may provide less information about sweetness than do rank data, in which the relative sweetness of a group of products is ordered from least to most, which in turn provide less information about sweetness than direct intensity ratings obtained from an ordered category or line scale.

4. Significance and Use

4.1 Rating instruments or rating scales are commonly used in many areas such as sensory evaluation, marketing research, experimental psychology, survey research, and economics in which there is interest in quantifying perceptions such as liking, preference, level of purchase interest, intensity of an attribute, degree of difference, or level of agreement with statements. This guide is concerned with the scales that are used to record human responses to physical stimuli rather than measuring physical entities. Many types of rating scales already exist and have been used in the above fields. Specific rating scales each have their own properties, advantages, and disadvantages. Some rating scales are intended for specific applications, while others have broader applications. Some rating scales have been extensively studied and modeled and have well-established properties.

4.2 Given the overwhelming number of scales available to practitioners when designing research, it is necessary for the researcher to have some knowledge about the scales that are available along with the many considerations that surround their use and applications. This guide will be useful to the sensory researcher who wants to use a scale as a measuring tool for their work. Selecting the right scale is a critical step towards meeting the research objective and making valid conclusions.

5. Data Properties

5.1 This section concerns the properties of the data obtained from numerically encoding responses obtained from rating scales. Data generated using rating scales are classified by the type of information supplied, often referred to as "levels of measurement." Note that the levels of measurements outlined in the following pertain to the data generated rather than to the rating instrument itself. The data properties should be considered when determining which statistical analyses are appropriate.

5.2 The four levels of measurement are:

5.2.1 *Nominal Data*—Differentiates samples or assessors based on arbitrary categories or qualitative classifications. The categories or classifications do not have numerical significance.

5.2.1.1 Examples are gender, ethnicity, and religion.

5.2.2 *Ordinal Data*—Ordinal data arise from ranking items or a set of ordered categories. In either case, the data do not include information about the relative spacing between these scores. In other words, numerically identical differences on any part of the scale are not necessarily identical in magnitude with respect to the variable measured.

5.2.2.1 Many of the scales presented throughout this guide are ordered category scales. Strictly speaking, they do not generate anything more complex than an order of the items or sensations being evaluated.

5.2.2.2 *Example*—A five-category "meets expectations" scale with anchoring points ranging from "much worse than expected" to "much better than expected" allows each assessor to categorize items based on how well the items met his or her expectations. However, the difference between scale categories is not likely to be interpreted the same among assessors.

5.2.3 *Interval Data*—Interval data are obtained from a rating instrument that does not have a true zero point even though one of the scale point labels may be called "zero" and has numerically identical differences on any part of the scale. In other words, points on the scale are equally spaced such that the numbers assigned represent actual degrees of difference between samples. Since an interval scale does not have a meaningful zero point, ratio comparisons are not appropriate. However, the numeric differences between values assigned to the categories are meaningful. Differences on an interval scale do have ratio properties.

5.2.3.1 *Example*—The Fahrenheit temperature scale is an interval scale. A 5° change represents the same degree of difference at all points on the scale (that is, the difference between 5 and 10° is the same as the difference between 25 and 30°). However, since the 0 value is arbitrary, it is not appropriate to apply ratio comparisons such as "80° is twice as hot as 40°." However, one could say that (212 - 68)/(68 - 32) = (100 - 20)/(20 - 0), which connects a Fahrenheit-based ratio to a Celsius-based ratio and shows that ratios of differences on interval scales are meaningful.

5.2.4 *Ratio Data*—Ratio data are obtained from a rating instrument that has a true zero point and numerically identical differences on any part of the scale. In other words, points on the scale are equally spaced such that the numbers assigned represent actual degrees of difference between samples. As the name implies, ratios of these assigned values are meaningful.

5.2.4.1 Examples are length, mass, age, and the Kelvin temperature scale. It is appropriate to say that 12 m is twice as long as 6 m.

5.2.5 The visual look of the rating instrument does not guarantee any property of the data collected with it. Many rating instruments in sensory science appear to generate data with interval or ratio properties when they do not. The researcher needs an understanding of the data properties to choose the appropriate statistical analysis approach (**1**).[3]

5.2.6 *Reliability and Validity*—Rating scales do not have inherent reliability and validity. How the panel uses the rating scale, the experimental procedure, and many other factors impact the reliability and validity of a sensory method. For an explanation of reliability and validity in the context of sensory studies, refer to the major sensory science publications (**2-5**).

6. Classification of Scales

6.1 *Classification Based on Objective:*

6.1.1 Hedonic scales are used when the research objective is to assess how much assessors like products or samples.

6.1.2 Relative scales are used when the research objective is to assess samples relative to another sample or to an ideal.

6.1.3 Attitude scales are used when the research objective is to assess consumers' attitudes or opinions.

6.1.4 Intensity scales are used when the research objective is to assess the perceived intensity of samples' sensory attributes or the perceived intensity of the difference between samples.

6.2 *Classification Based on the Objective of the Response*—All scales can be classified according to whether the assessor communicates an internal reaction, attitude, or intention or whether the assessor communicates the property of an external product or stimulus. If the response is a function of the person making the rating, it is an "assessor-focused" scale. Responses on an assessor-focused scale can change when the researcher changes the characteristics of the respondent sample. Liking, attitudes, emotion, and agreement are all assessor-focused scales. When the response is a function of the product, the scale is considered a "product-focused scale." With product-focused scales, the responses are a function of the product and thus are not expected to change unless there is variation in the product, even when the characteristics of the respondent sample are changed. Product-focused scales are intensity scales, quality scales, grading scales, and relative to reference rating scales.

6.3 *Structural Classification:*

6.3.1 *Scale Polarity:*

6.3.1.1 Scales are either unipolar or bipolar.

6.3.1.2 A unipolar scale is used to record responses that are increasing from low (or zero) at one end to high at the other end. An example of a unipolar scale for sweetness intensity is a scale anchored at "not at all sweet" at one end and "extremely sweet" at the other end. In addition to rating intensity, unipolar scales are appropriate for rating amount (for example, amount of sauce) and frequency (for example, of consuming a certain product) to name a few applications.

6.3.1.3 In a bipolar scale, the endpoints are semantic opposites with an implied or stated midpoint. Examples of a bipolar scale are the hedonic scale (see Fig. 1), the just-about-right scale (see Fig. 2), the purchase intent scale (see Fig. 3), and Likert scales (see Fig. 4). With bipolar scales, careful consideration shall be given to whether the endpoints are truly opposites, as the bipolarity implies. For example, a scale ranging from "much too sour" to "much too sweet" or a scale with a mid-point of "neither sweet, nor sour" are incorrect because "sour" and "sweet" are both present in many products and are not semantic opposites, and therefore, they should be separated into two different scales. One of these scales should range from "not sweet enough" to "much too sweet;" while the other scale should range from "not sour enough" to "much too sour." Bipolar scales can either be balanced (with an equal number of categories or an equal line length on either side of the midpoint) or unbalanced. The BASES scale (Fig. 5) is a bipolar unbalanced scale. Unbalanced scales are less common than balanced scales. When translating scale(s) to another language, care should be taken to assure semantic opposites are maintained.

6.3.2 *Scale Continuity:*

6.3.2.1 Scales are category scales or line scales; they can also be a hybrid of the two.

6.3.2.2 Category scales offer the assessors a limited number of responses (typically ten or fewer) from which to choose. More specifically ordered category scales indicate an increasing or decreasing degree of a variable labeled with numbers, words, or symbols. The intensity scales in Figs. 6 and 7 are examples of ordered category scales.

Please check the box that best describes your liking of this product.

- ☐ Like extremely
- ☐ Like very much
- ☐ Like moderately
- ☐ Like slightly
- ☐ Neither like nor dislike
- ☐ Dislike slightly
- ☐ Dislike moderately
- ☐ Dislike very much
- ☐ Dislike extremely

FIG. 1 Nine-Point Hedonic Scale

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

Please indicate your opinion about the sweetness level in the sample.

1	2	3	4	5
Much too weak	Slightly too weak	Just about right	Slightly too strong	Much too strong

FIG. 2 Just-About-Right (JAR) Scale

Please select the statement that best describes your intent to purchase this product.

- ☐ Definitely would buy
- ☐ Probably would buy
- ☐ Might or might not buy
- ☐ Probably would not buy
- ☐ Definitely would not buy

FIG. 3 Purchase Intent Scale

6.3.2.3 Line scales, also known as visual analog scales (VAS) or unstructured line scales, indicate an increasing or decreasing degree of a variable using an anchored continuum for responding and are not restricted to a fixed set of categories. The assessors are free to place a mark anywhere along the line implying, at least theoretically, an infinite number of response options. The intensity line scale in Fig. 8 is an example of such a scale.

6.3.2.4 It is also common to find scales that are a hybrid of category and line scales. The visual thickness scale in Fig. 9 is an example. These scales are used by the panel in the same way as a line scale (that is, as a continuum) but the categories provide anchoring points. They are more often used with a trained panel that was calibrated to the scale and its anchoring points.

6.3.2.5 Many attributes (for example, sensory intensity) can be measured using category scales or line scales. The choice is up to the researcher. See Section 7 for an overview of considerations the researcher should take into account.

7. Important Considerations in Scale Selection

7.1 *Objectives of the Research:*

7.1.1 The main applications presented in this guide are those that the sensory scientist will typically encounter. They are: (*1*) descriptive analysis when the scientist intends to measure the specific sensory properties of samples; (*2*) quality control when the scientist intends to assess the sensory quality of samples; and (*3*) consumer testing when the scientist intends to assess the consumers' perception, liking or preference of samples, or attitudes.

7.1.2 The sensory scientist may occasionally need to work with marketing research scales. These are outside the scope of this guide as they often are multi-item scales. The *Marketing Scales Handbook* series **(6)** is a compendium of such published marketing research scales.

7.1.3 The research objective will provide some direction as to which scale to choose. This guide presents commonly used scales for each research objective in Sections 8 – 10.

7.1.4 The researcher needs to consider the features of different scales to determine which is the most appropriate for the needed information. There are risks and benefits that shall be considered for each situation.

7.1.5 Scales typically used in descriptive analysis and quality control are intensity and relative scales.

7.1.6 Scales typically used in consumer testing are hedonic scales, behavioral choice, relative, and attitude scales.

7.1.7 Scales are used in many different situations. Some scales may be specific to a single application, while others may be appropriate for multiple applications. For example the nine-point hedonic scale (see 10.1.1.1) is only used in consumer testing applications. For example Check-All-That-Apply scales (see 10.3.3) which were originally used in consumer testing have been more recently used in descriptive analysis or quality control situations.

7.2 *Ease of Use:*

7.2.1 When choosing a rating scale, the sensory scientist needs to consider who the assessors are along with their cultural background, their cognitive abilities, and if relevant, their level of sensory training. All these factors impact their understanding of the rating scale and their ability to perform the task. It is also important to consider how easy the data entry will be and how the resulting data will be analyzed. In general, the rating scale should be easy to use for the assessors as well as for the sensory scientist.

7.2.2 Familiar scales are more comfortable for assessors to use. When developing a full questionnaire/ballot for naïve assessors, it is advisable to not have too many different types of scales in a single questionnaire to avoid confusion.

7.2.3 Line scales are better suited when one uses a computerized data collection; this is because data entry from a paper ballot requires measuring responses on the line scale, which can be a cumbersome process before actual data entry.

7.2.4 Special considerations are necessary when using rating scales with children and assessors across cultures and countries, as research has shown that different geographies think about and use scales differently. The *International Consumer Product Testing across Cultures and Countries*, MNL55-EB **(7)** provides guidelines on the use of rating scales in different countries and Guide E2299 provides guidelines when testing with children. All the considerations outlined in 7.3 – 7.9 need to be reviewed ahead of time. A preliminary study to validate the rating scale with the intended population of assessors is recommended. When interpreting the data from multiple geographies or cultures or both, it is important to understand if the product acceptance ratings are actually different or reflect different use of the scale.

7.3 *Number of Points on the Rating Scale:*

7.3.1 The number of points on the rating scale shall be sufficient to allow assessors to express their range of responses. Too few scale points may force assessors to report similar scale values for stimuli that elicited different responses. It is up to the sensory scientist to determine what that number should be to allow separation of responses.

7.3.2 The number of discrete scale points shall not be so large that it becomes cumbersome or hinders ease of use on the part of the assessors or the sensory scientist. When too many

Please indicate your level of agreement with each of the following items (circle the appropriate number).

Item	*Strongly Disagree*	*Disagree*	*Neutral*	*Agree*	*Strongly Agree*
This product is appropriate for kids	1	2	3	4	5
This beverage tastes refreshing	1	2	3	4	5
This is the best cereal bar I ever tried	1	2	3	4	5

FIG. 4 Example of a Likert (Agree-Disagree) Scale

Please indicate your degree of liking of this product.

- ☐ Like extremely
- ☐ Like very well
- ☐ Like quite well
- ☐ Like somewhat
- ☐ Like slightly
- ☐ Do not like at all

FIG. 5 The BASES Scale

Please indicate the intensity of sweetness in this product

- ☐ Extreme
- ☐ Moderate to Extreme
- ☐ Moderate
- ☐ Slight to Moderate
- ☐ Slight
- ☐ Very Slight
- ☐ None

FIG. 6 Intensity Ordered Category Scale

Please indicate the intensity of the sweetness in this product.

1 2 3 4 5 6 7 8 9

Weak **Strong**

FIG. 7 Intensity Ordered Category Scale

scale points are used relative to the discrimination ability of the assessors, they will likely simplify the scale and use a smaller number of the points.

7.3.3 Research to date has not defined a single optimal number of response categories on a rating scale. Most attitude scales, including Likert scales, have five or seven response categories because it is easy to differentiate and understand five to seven different labels. For intensity scales, a wide number of options are used (for example, 3- to 100-point scales). Often intensity scales with larger numbers of scale points do not have every scale point labeled. Preston and Colman (**8**) found that rating scales with seven, nine, or ten response categories were better liked by consumers than scales with fewer response categories.

7.3.4 Note that assessors have different levels of comfort using the ends of the scale. The number of scale points or scale length or both should be sufficient to still allow discrimination for the samples even if assessors tend to avoid the end points of the scale.

7.3.5 A specific number of scale points cannot be recommended. Research of scale usage has found high correlations between scales with a different number of points (**2**); as long as the scale fits the research objective appropriately, many scales will perform well if the task is reasonable and clear. When the researcher is interested in comparing samples, similar trends are typically observed between samples regardless of the specific scale used.

7.3.6 Bipolar scales are best used with an odd number of response categories. These scales should be balanced; have the same number of response categories on either side of the mid-point; and allow for a full range of probable responses, including a neutral response. Unbalanced scales run the risk of biasing assessors either positively or negatively by suggesting they focus on one part of the scale.

7.3.7 Unstructured line scales are advantageous because no consideration needs to be given to the number of points on the scale. Many unstructured line scales are 15 cm long. However, consideration needs to be given to the length of the scale. Analogous to the number of points required for a category scale, a line scale should be sufficiently long as to allow discrimination. Consumers may find it more difficult to use line scales consistently across sessions and they may benefit from the anchors that category scales can provide. Additional consideration should be given to the device on which the line scale will be displayed and whether it will allow the same degree of discrimination across multiple devices. A line scale may not allow the same degree of discrimination when projected on a smaller screen (such as a phone) as when projected on a larger screen (such as a computer monitor). Scale length should be consistent for a specific panel throughout training and data collection.

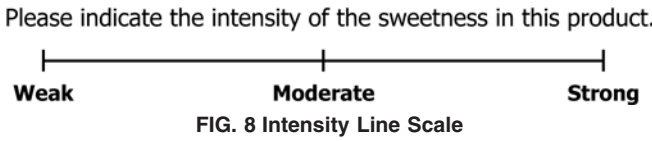

FIG. 8 Intensity Line Scale

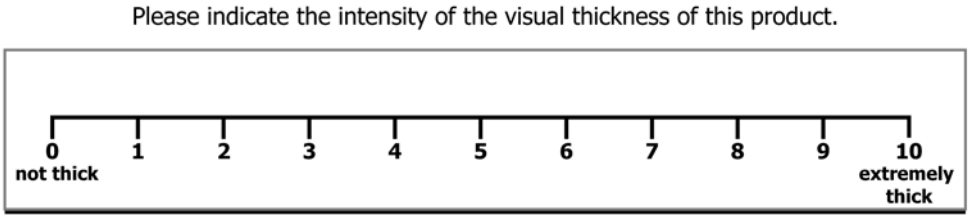

FIG. 9 Visual Thickness Scale

7.4 *Scale Anchoring Points:*

7.4.1 Anchoring points are an important part of any scale. They define one or several points on the scale and indicate its directionality and they help assessors use the scale in a common way. A minimum of two anchoring points are needed to anchor the ends of the scale. Anchoring points may also be used at intermediate points along the scale. Anchoring points indicate whether the scale is unipolar or bipolar. They are especially important when the same question and scale are used over multiple time points or multiple sessions—they help the assessor use the scale more consistently and indicate that the response is similar or different to other questions in the same study. Maintaining consistent anchoring points from study to study helps the researcher to compare data across studies.

7.4.2 Anchoring points may be single words but are often word phrases (for example, “none” or “like extremely”); they can also be numbers, pictures, physical stimuli, or a combination of these. While anchoring points are required to define a scale, the researcher shall consider whether to label any or all intermediate scale points. When selecting word phrases, ensure wording is neutral and balanced along the entire scale to reduce potential biases (see 7.8.6). Also, the chosen wording should convey the intended distances between anchoring points as much as possible.

7.4.3 The range of anchoring points needs to match the range of expected responses. If a rating scale is anchored in such a way that all samples are scored very close to the top (or bottom) of the scale, it may be difficult to discern differences between the samples. For example, if sugar candies are evaluated on sweetness with the top of the scale anchored as “sweet” instead of “very sweet,” it may become difficult to detect differences in sweetness because of the samples all receiving high scores.

7.4.4 Cultural differences need to be taken into account when designing anchoring points on a scale. A direct translation of anchoring points may not be adequate (**7**). In some cultures, the use of numbers as anchors may not be appropriate (**7**). In some other cultures, there may be reluctance to use negative response categories such as “dislike” on a rating scale (**9**).

7.5 *Relativity of Responses on a Scale*—All responses on a scale involve a comparison to an internal frame of reference. This frame of reference may consist of all the products that the assessor has experienced in this category in the past or it may consist of a direct comparison to a concrete product provided in the test (that is, a reference sample). Because all sensory responses are inherently relative, they are especially prone to biases (see 7.8).

7.6 *Scale Orientation*—The orientation of the rating scale, whether it reads from left to right, right to left, up to down, or down to up, needs to conform to the cultural norms of the country in which it is used so that it does not introduce any confusion to the assessors (**7**). Rating scales can be presented in either a horizontal or a vertical layout.

7.7 *Instructions and Training:*

7.7.1 Some scales are more likely to require assessor training. The need to train assessors on scale use is determined by the research objective, the test method, and corresponding best practice. Typically, hedonic, behavioral choice and attitude scales do not require training. Relative and intensity scales may or may not require assessor training depending on the pool of assessors.

7.7.2 Instructions need to be concise but detailed enough that assessors will understand what their task is. In the case of hedonic scales, general instructions are typically sufficient. “Liking,” “sweetness,” and “thickness” are relatively unambiguous terms. Terms such as “creamy,” “smooth,” and “fresh” may be less clear with respect to what they refer to, either in terms of the product itself, or with respect to the assessor’s experience. In general, the researcher shall include terms that are as unambiguous as possible, and the sensory scientist always needs to interpret data carefully.

7.7.3 In the case of descriptive analysis and quality control in which assessors rate the intensity of specific attributes, training is recommended to ensure that assessors understand the attribute being rated and the meaning of the anchoring points on the scale. Ambiguity of terms or attributes may cause assessor confusion when evaluating a perceived intensity.

7.7.4 Members of descriptive panels are trained on the concept (common definition/criteria) of the attribute being measured and sometimes calibrated on intensity along the scale depending on the descriptive method used.

7.8 *Response Biases:*

7.8.1 No matter the type of rating scale being used, researchers need to be aware that scale responses are influenced