Designation: E1958 – 16a

Standard Guide for Sensory Claim Substantiation[1]

This standard is issued under the fixed designation E1958; 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.

INTRODUCTION

Formats or standards for testing related to sensory claim substantiation cannot be considered without a frame of reference of where that format or standard would fit within the legal framework that surrounds the topic. Product sensory claims tests are performed for three basic reasons: (*1*) *Comparison of Products*—Determines how one product compares to another, usually a competitor or earlier version of itself. (*2*) *Substantiation of Claims*—Enables marketing personnel to use positive references through advertising or packaging, or both, in the presentation of the product to the consumer. (*3*) *Test Performance*—Ascertains and establishes the tested product performance within the scope of its intended use.

The risk associated with each claim is assessed when considering claims substantiation. Compelling and aggressive claims are sure to be scrutinized closely by competitive firms, and if inconsistencies are found through competitive test data, the claims could be challenged in one or more of the following venues: *(1)* National Advertising Division (NAD) of the Advertising Self-Regulatory Council (ASRC), *(2)* one or more media, such as print, broadcast, or electronic media, *(3)* Consumer Advocacy Organizations, and *(4)* Civil or Federal courts. No single test design or standard test will prevent challenges. The criteria used by each of the potential forums are not identical and are constantly evolving. With the introduction of new technologies coupled with changing consumer demands, testing processes and protocols that were sufficient five or ten years ago may not hold up under today's criteria and scrutiny. Testing requirements of the future can only be a matter for speculation. The one constant is that, as advocates of their clients' positions, attorneys will defend their clients' testing processes and protocol while questioning with great detail every aspect of their competitor's protocol in the attempt to sway the arbiter to agree that their clients are in the right. Legal counsel should be part of any team developing claim substantiation.

This guide demonstrates what a group of professionals who are skilled in the science of testing consider appropriate from a scientific and technical standpoint, and represents an effective method for both defendant and challenger to determine the viability of a sensory claim. The key word is "appropriate." If a particular aspect of a test, or method, is not appropriate for a specific application, it should not be used. Care should be taken to clearly define the reasons and data supporting a deviation from the standard, as any departure invites scrutiny. Since departures are inevitable, the word "should" is used in this guide to indicate when other techniques may have applications in certain unusual circumstances. Whenever a test protocol has been completed, it should be critiqued for weaknesses, including whether experts in the relevant field would consider the research objectively designed, conducted, and analyzed, using procedures that give accurate and reliable results. If weaknesses are found, corrective action should be taken, since the competition may point out any weakness or discrepancy and challenge the study.

While the scientific and technical community identifies the appropriateness of a research method used to support a sensory claim, the legal community evaluates substantiation for legal claims using "reasonableness" as the criterion. With the importance of having a legal "reasonable basis" for a claim, the question remains, "What is reasonable?" Unfortunately, there is no specific answer to that legal question, as it will depend on the type of claim, product application and use, applicable regulations where the product is sold, and other factors. These considerations, market pressures (such as timing), and testing budgets can influence and impact the protocols to support a specific claim. This guide provides principles and considerations that need to be addressed for good sensory and consumer testing practices.

1. Scope

1.1 This guide covers reasonable practices for designing and implementing sensory tests that validate claims pertaining only to the sensory or perceptual attributes, or both, of a product. This guide was developed for use in the United States and must be adapted to the laws and regulations for advertisement claim substantiation for any other country. A claim is a statement about a product that highlights its advantages, sensory or perceptual attributes, or product changes or differences compared to other products in order to enhance its marketability. Attribute, performance, and hedonic claims, both comparative and non-comparative, are covered. This guide includes broad principles covering selecting and recruiting representative consumer samples, selecting and preparing products, constructing product rating forms, test execution, and statistical handling of data. The objective of this guide is to disseminate good sensory and consumer testing practices. Validation of claims should be made more defendable if the essence of this guide is followed.

Table of Contents

2. Referenced Documents

2.1 *ASTM Standards:*[2]

E253 Terminology Relating to Sensory Evaluation of Materials and Products

E1885 Test Method for Sensory Analysis—Triangle Test

E2164 Test Method for Directional Difference Test

2.2 *ASTM Publications:*[3]

ASTM Manual 13 Descriptive Analysis Testing for Sensory Evaluation

ASTM Manual 26 Sensory Testing Methods: Second Edition

STP 913 Physical Requirement Guidelines for Sensory Evaluation Laboratories

3. Terminology

3.1 *Definitions*—Terms used in this guide are in accordance with Terminology E253. Additional terms are as follows:

3.1.1 *attribute difference rating test*—this test also determines if one or more specific attributes differ between two samples. The intensities of the attributes are measured on rating scales showing several degrees of intensity. One or more specific attributes of the product that relate to the claim are rated. Samples are presented, and the panelists' task is to evaluate and assign each test sample an intensity to reflect the amount of the designated attribute(s).

3.1.2 *attribute difference tests*—in these test methods, the attribute of interest is defined prior to testing, and the panelists are trained to be able to identify the attribute in question and select or rate the relative intensity of that attribute. It is not necessary to evaluate every occurring attribute, only the attributes being addressed in the claim.

3.1.3 *ceiling effects*—this typically occurs when the majority of the scores occur toward the top of a rating scale. When the products are well-liked, there is not a sufficient amount of scale available to the respondents to differentiate the products. Variation in rating scores is compressed, making mean-based statistical tests misleading. Therefore, analysis should be performed using a more robust statistical model that does not have distributional requirements and is less prone to outlier influence such as multinomial logistic regression.

3.1.4 *central location testing (CLT)*—method of testing that provides maximum control over product preparation and usage. Central location testing assures that the participant actually evaluated the product in question and provides his or her

[1] This guide is under the jurisdiction of ASTM Committee E18 on Sensory Evaluationand is the direct responsibility of Subcommittee E18.05 on Sensory Applications--General.

Current edition approved Oct. 1, 2016. Published October 2016. Originally approved in 1998. Last previous edition approved in 2016 as E1958 – 16. DOI: 10.1520/E1958-16A.

[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 ASTM International Headquarters, 100 Barr Harbor Drive, PO Box C700, West Conshohocken, PA 19428-2959.

own opinion immediately following evaluation, rather than relying on past usage or recollection of a CLT.

3.1.5 *comparative claims*—designed to compare similarities and differences between two or more products. The basis for comparison can be within the same brand, between two brands, or between a brand and other products in the category.

3.1.6 *context/contrast effect*—flavor/texture of one sample can have an influence on the perceived flavor/texture of each subsequent sample.

3.1.7 *directional difference test*—this test method is used when determining whether one sample has more of a particular sensory characteristic than another. Two samples are presented, either simultaneously or sequentially, and the respondent chooses one of the samples as having a higher level of the specified characteristics.

3.1.8 *equality claims*—in equality claims, two products are claimed to be equal in one or more particular feature.

3.1.9 *experimental error*—variability between the panelist. This error can be accounted for by using more than one panelist to test each sample.

3.1.10 *home use testing (HUT)*—refers to tests that allow respondents to use the products in a more natural environment, rather than the controlled environment.

3.1.11 *measurement error*—repeatability within the individual panelist. This error can be accounted for by having each panelist test a particular sample more than once.

3.1.12 *monadic or single product tests*—product tests where only one product is experienced and rated.

3.1.13 *parity claims*—parity claims are claims that rank equivalent levels of performance or liking when comparing a particular product to another product. In general, parity claims are made relative to a market/category leader. Within parity claims, two additional classes exist: equality claims and unsurpassed claims.

3.1.14 *pattern effect*—any pattern in order will be detected quickly.

3.1.15 *positional bias*—respondents may be more sensitive to differences in specific samples in a series, such as the first or last sample.

3.1.16 *product variability*—batch-to-batch variation. This error can be accounted for by testing multiple and representative batches of a product.

3.1.17 *self-administered questionnaire*—questionnaires independently completed by the respondent are referred to as self-administered.

3.1.18 *superiority claims*—a superiority claim is supported if a statistically significant proportion of the respondents prefer the advertiser's product.

3.1.19 *superiority claims*—superiority claims assert a higher level of performance or liking relative to another brand. Superiority claims can be opposed to competitive brands (for example, "cleans better than brand Z") or opposed to an earlier formula of the brand (for example, "now more cleaning power than before").

3.1.20 *unsurpassed claims*—in unsurpassed claims, the claim stated indicates that the product(s) selected for comparison is not better/higher (or greater than) in some way to the target product(s) for which the analysis is executed.

4. Basis of Claim Classification

4.1 A fundamental step in advertising claim substantiation is creating an explicit statement of the claim prior to actual testing. The statement is then forwarded to all parties concerned in the substantiation process. Concerned parties could include marketing, marketing research, legal, consumer testing, sensory evaluation, research suppliers, etc. The statement is essential as it can encourage collaboration in terms of corporate resources, confirms the selection of appropriate test methods, and has the potential to maximize the chance of making reliable business decisions about the proposed claim, pending the results of substantiation research. Collaboration among all involved parties prior to executing substantiation research is critical in achieving the best results. All involved parties should meet and agree (perhaps several times) prior to implementing the substantiation research.

4.2 Familiarity with the general classification of advertising claims is important in developing clear statements of claims at an early stage and for developing a rational plan for testing. This familiarity also facilitates the process of selecting appropriate testing methods, among the many types of methods available to the consumer/sensory science professional. Each method answers specific questions and may support one type of claim but not another. Therefore, the consumer/sensory science function provides an important source of information and experience in claim substantiation and will provide much of the definition of testing methodology. There are multiple ways to support claims depending on the characteristics of the claim. Two approaches are consumer based and trained panel based evaluations.

4.3 Advertising claims can be divided into two fundamental classifications: Comparative and Non-Comparative. The distinction between the two classifications is whether a comparison is made relative to an existing product (advertiser's or competitor's) or to itself.

4.4 *Comparative Claims* are designed to compare similarities and differences between two or more products. The basis for comparison can be within the same brand, between two brands, or between a brand and other products in the category.

4.4.1 Comparative claims generally take one of two forms: parity or superiority. Parity and superiority are further subclassified into two central areas of application: hedonic and attribute/perception. Hedonics broadly concern measuring the degree of liking and preference—either liking overall or liking that is limited to one or more specific attributes. Attribute/perception claims apply to intensity when measuring one or more specific product attributes.

4.4.2 *Parity Claims*—Parity claims are claims that rank equivalent levels of performance or liking when comparing a particular product to another product. In general, parity claims are made relative to a market/category leader. Within parity claims, two additional classes exist: equality claims and unsurpassed claims.

4.4.3 *Equality Claims*—In equality claims, two products are claimed to be equal in one or more particular feature:

4.4.3.1 *Hedonic*—"Tastes as good as brand X."

4.4.3.2 *Attribute/Perception:* "Our product reduces odors as much as brand X."

"Our product lasts as long as brand X."

"Our cake is as moist as the leading brand."

4.4.3.3 *Overall Equality:* "We're just the same, except for the price."

"You'll never know the difference between us and brand X."

4.4.4 *Unsurpassed Claims*—In unsurpassed claims, the claim stated indicates that the product(s) selected for comparison is not better/higher (or greater than) in some way to the target product(s) for which the analysis is executed. Examples of unsurpassed claims include the following types:

4.4.4.1 *Hedonic:* "No other product is better than our product."

"No other product is more liked for butter flavor."

4.4.4.2 *Attribute/Perception:* "No other cake is more moist than ours."

"No other product has more butter flavor than ours."

"No other product reduces odors more than our product."

"No other product lasts longer than our product."

"No other product is thicker than our product."

"No other product cleans faster than our product."

4.4.5 *Superiority Claims*—Superiority claims assert a higher level of performance or liking relative to another brand. Superiority claims can be opposed to competitive brands (for example, "cleans better than brand Z") or opposed to an earlier formula of the brand (for example, "now more cleaning power than before"). Examples of superiority claims include:

4.4.5.1 *Hedonic:* "Our product tastes better than brand X."

"Our product tastes better than any other."

"Our product is preferred over any other brand."

4.4.5.2 *Attribute/Perception:* "Our cake is more moist than any other."

"Reduces odors more than brand X."

"Lasts longer than any other product."

"Thicker than brand X."

"Cleans faster than any other product."

4.4.5.3 In superiority claims, combinations of hedonic claims and attribute/perception claims can sometimes be found, when superiority claims are established based on overall liking and for specific attributes (for example, "Our hosiery is preferred over Brand X for overall liking and it offers more support and comfort.").

4.4.5.4 From a statistical perspective, it can be easier to support a claim of superiority than one of parity, assuming that the superiority actually exists. This fact about hypothesis-testing will be discussed further in the section on statistical methods.

4.5 *Non-comparative/Communications Claims*—The objective of the non-comparative/communications claim is to convey something specific about the product, usually a product benefit or difference, and in general, does not seek to provide comparative claims relative to other products. For example, the statement "provides long-lasting flavor" or "smells strong for one month" tells us something about the product, but not in a comparative sense relative to an existing product. These types of claims are common in new product types, but also are used to bring attention to specific product benefits. Examples of non-comparative/communications claims include the following types.

4.5.1 *Hedonic:*

"Tastes great."

"Makes your laundry outdoor-fresh."

"Leaves a long-lasting freshness you will like."

4.5.2 *Attribute/Performance:* "Removes odors for 60 days."

"Leaves glass streak-free."

"Leaves no residue on surfaces."

"Works fast."

NOTE 1—In the above attribute examples, some of these could be approached either as a non-comparative claim, since no other product is mentioned, or as a comparative claim versus an appropriate standard (streak-free glass, residue-free surface, odor-free room).

4.6 *Selecting the Appropriate Ad Claims Test*—Product claims made in print or on radio, TV, or the Internet require valid data that supports the intended claim. As with most sensory testing, it is necessary to first identify the project and test objectives for the study. The claim statement should indicate whether the claim is based on consumer or laboratory sensory methods or, in fact, some instrumental or chemical test. Sensory claims for preference or liking ("preferred over the leading brand" or "better than the competition") require consumer tests with the preference or liking questions to support the claim. Claims about product attribute(s) or performance can be based on data from consumers, who are asked about the specific attribute, or from laboratory sensory tests designed to measure the specific attribute(s). In some cases, both types of testing (consumer and laboratory) can be used together to support the same claim. The ad claims team needs to determine the type of claim, the claim statement, the target population, and the aspect(s) of the product that is the focus of the claim. Only then can the test to support the claim generate data with the right focus and weight to support the claim.

5. Consumer Based Affective Testing

5.1 *Sampling:*

5.1.1 Claims refer to product performance or product liking by purchasers or consumers. Hedonic claims should always apply to the user population. Sampling from any population other than the users to whom the claim is focused, such as purchasers, may require a qualified claim to limit its generality. The test protocol should state clearly whether a claim is being made for the purchasers or the ultimate consumer of a product, or both, when the distinction exists. Classic illustrations would include adults with children and pet owners. For example, "Choosy mothers choose Jif[4]" is a claim specific to the purchaser and not the consumer. It is evident that the claim itself has a role in defining the target population.

5.1.2 Screening based upon recent category usage is recommended to identify target consumers. If recent category usage is not applicable (such as with seasonal products or products

[4] Trademark Jif is a registered trademark of the J. M. Smucker Company.

with long purchase-repeat cycles), identifying target consumers based upon positive future category usage intent is acceptable. The category should be defined in such a way that validates the selection of competitive products, (for example, "raisin bran" rather than "ready to eat cereal"). Respondents should not be restricted to exclusive category usage (such as eating only raisin bran), but also may use alternative products in related categories like corn flakes or bran flakes. Respondents also should not be restricted to heavy users, which are a subset of users and would require a qualified claim.

5.1.3 For category usage claims, respondents may be recruited by screening for brand usage, but care should be taken during screening to ensure respondents are unable to guess which brands are targeted for testing. Screeners can mention a large list of brands with the brand or brands of interest embedded in the questionnaire. Brand usage and frequency of use data also can be collected to help validate the target population. Product users can be defined by their responses to several questions, including:

5.1.3.1 "What one brand of this product type do you use most often?"

5.1.3.2 "What brands have you used in the last (insert time period appropriate for category)?"

5.1.3.3 If frequency of use is an issue, then the respondent also may be asked how often they use the product or how many times they have purchased the product within a specific time frame. More discussion can be found in 6.9.

5.2 *Sampling Techniques:*

5.2.1 The type of claim should be kept in mind when determining sample size. For example, parity claims may require more respondents than superiority claims (see 6.6) and some objective claims, (for example, "this product has more...") can be substantiated through descriptive analysis by a trained panel (see Section 10).

5.2.2 The demographics of the test sample should match those of the target population (that is, about whom the claim is being made). The demographics may include the population in terms of age, gender, and geography. Respondents also may be screened for their product usage patterns and the sampling density should reflect the geographic distribution of this group.

5.2.3 Using quotas is helpful to achieve a match between a test population and the intended target population. Representation of age and gender should match the target population and reflect the age distribution of users within each gender. Demographic information must be collected to demonstrate the validity of the sample.

5.2.4 Recruiting criteria of the test population must be stated in the test protocol and should be as objective as possible. Records must be kept indicating why potential respondents were rejected from the study. Screening criteria should not be revealed to potential respondents, and the standard security screening questions (for example, whether family members work in advertising or marketing or other related fields, including that of the test product) should be included.

5.2.5 A constrained demographic sample such as a single gender sample should be employed when it is consistent with the stated claim and normal product usage. For example, primarily women or the elderly may use specific products.

5.2.6 Names of potential test participants may be available from outside companies who sell marketing information. In many cases, a company may maintain its own database on product users. In most cases, these databases are maintained using good research technique; however, use of databases may not approximate a probability sample, and therefore, in certain instances, would not be acceptable for claims substantiation.

5.2.7 If potential respondents are selected based on an existing database, caution should be taken to ensure that the database is accurate. Oftentimes, databases include potential respondents who claim they use the product(s) being tested to take advantage of paid evaluation, or they may not reflect the users' latest buying habits. It is recommended that respondents be screened specifically for this test to ensure they represent the intended user and have not participated in consumer tests within the past three months or tests within the category for the past six months.

5.2.8 The geographic balance required for substantiating a claim is a function of the nature of the claim. Perception of laundry whiteness, pain relief, and other perceptual claims based on the functional performance of a product are unlikely to have a specific geographic dependence. However, when hedonic testing is conducted with a product used at home under widely varying conditions, for example, testing detergents in home, factors such as water hardness, humidity, average ambient temperature, and so forth may affect product performance and preference for the product. If there is evidence that such factors do affect product performance, they should be taken into consideration when selecting test markets.

5.2.8.1 Preference claims have a potential for geographical and demographic dependencies. Preference may vary by region or by socioeconomic factors, such as urban versus suburban versus rural. The evidence for or against such dependencies could come from patterns in product sales, or usage, or both.

5.2.8.2 When geographic region is assumed to be a factor relevant to a claim, the geography of respondents should be consistent with the scope of the claim. A national claim should be based on a sample representing major geographic regions (North, East, Midwest, South, West, etc). A minimum of two markets in each of the four regions should be included. Regional claims should represent at least four markets that are geographically dispersed across the region.

5.2.9 In general, simple or stratified random (quota) sampling methods may be employed. It is incumbent on the claimant to ensure that the random sample is not biased or meaningfully different from a probability sample; that is, all members of the target population or a strata within the population should be guaranteed an equal probability of being selected for the test. Guard against bias in terms of social and economic groups by having more than one test site in a city or metropolitan area. Minimize sampling bias by conducting interviews across a wide range of days of the week and times of day and by varying the location where potential respondents are recruited.

5.2.10 Be cautious when selecting markets and insure that the test adequately represents the people residing in the

geographic territory on which the claim is based. In categories with strong geographic differences in market share, the total market share should be approximated by representing high, low, and average share markets in the study. Regional sample sizes may vary, reflecting their contributions in terms of number, but not heaviness of usage. A mix of large and small urban/metro, as well as rural markets, is desirable.

5.2.11 The criteria for market selection may be viewed as a factor in an experimental design. After determining the necessary factors, a list of potential markets should be developed for each level of each factor. For example, a list of high, medium, and low share markets can be developed for each of four census regions, resulting in twelve cells. One market can be selected at random from each cell, representing each region at each level of brand development. Random selection of markets and test locations within markets is also beneficial in assuring others that the test sample is a valid approximation of a probability sample.

5.2.12 Once a target population is defined and is represented adequately by sampling, results from the total sample (not its subdivisions or subgroups) are the critical factor in making a claim. Results among some subgroup may not correspond to overall results because sample sizes in subgroups are smaller, and therefore, not as statistically reliable. Moreover, since there is risk of false positives and false negatives in testing any hypothesis, analysis of multiple subgroups will increase the overall error rate. Therefore, given appropriate sampling from the target population, examination of subgroups is not a sound analytical practice for claims substantiation (see Section 14).

5.2.13 For products to be ingested (food or beverage), respondents should not be allowed to participate if they have any food allergies, regardless if the allergen is expected to be present in the samples or not. A list of ingredients should be made available to the testing agency or any respondent who requests a copy.

5.3 *Selection of Products:*

5.3.1 If a test is being conducted to support a competitive claim that is not brand-specific (for example, versus "other leading brands"), then the competitive brands should be the two brands with the highest national market share. If the market is highly fractionated, such that the top two national brands control less than 50 % of the market, then more competitors must be included in the test. Either the three leading national brands or any brand that is among the top two in the four major geographic regions of the country must be tested. Unless the product is tested against brands representing at least 85 % of the national market, it is recommended that claims should be made against specific brands in lieu of general superlative claims. Eighty-five percent (85 %) of the market is defined as all products within said category, including the brand making the claim.

5.3.2 Competitive brands should be in the same market segment as the brand for which the claim is being made. If a brand straddles market segments, then products most similar in a reasonable competitive context should be used.

5.3.3 When competing products are sold in more than one form, the products being tested must be of the same form or in the form most relevant to the claim. If a powdered drink mix is being compared with a competitor's product that also comes in a powdered drink mix and as a reconstituted liquid, both brands would have to be tested in their reconstituted from powdered forms. The specific directions for preparation given on each product must be followed. If there is substantial crossover use of different forms, a claim involving different forms may be desired. The forms tested must be stated explicitly as part of the claim, for example, "instant tastes as good as ready-made."

5.4 *Sampling of Products When Both Products are Currently on the Market:*

5.4.1 For central location consumer tests, commercial products to be used for competitive claims testing should be purchased at the end of the distribution chain to ensure the product is representative of the product the consumer would purchase. Some products are made at different or multiple manufacturing sites. In those instances, the product should be purchased from a distribution center that services the particular test areas.

5.4.2 For other test methods in which the test product is manufactured at one location, samples can be purchased from any high volume store. Products should be sourced at the same time from the same store(s) in each local testing area. Products should reflect the choice available to local consumers. Care should be taken to include a variety of production sites and dates that typically are found on the retail shelf.

5.4.3 In cases where competitive products are not sold in the same stores (for example, fast food restaurants and private label products) test products should be sourced as close in time as possible from locations that reflect choices available to local consumers. It is important that the geographic identity of samples match that of local test participants. This way, if national products manufactured in more than one site have been formulated differently to appeal to regional differences in sensory preferences, appropriate products will be tested against relevant regional competitors. It is critical that all information regarding product sourcing be documented.

5.4.4 Competitive products should be purchased in the standard size package with the highest unit volume or in similar size, or both, to the test product. Trial size and club-store oversized product packages should not be used unless the package meets the specific target of the claim.

5.4.5 Every effort should be made to obtain competitive products of representative freshness found in the marketplace. All products in the test should be of typical age. A freshly-made product should not be compared against a product nearing its expiration date.

5.5 *Handling of Products When Both Products are Currently on the Market:*

5.5.1 After procurement but prior to testing, handling, length of storage, and storage conditions of all products must be identical and consistent with normal consumer practice.

5.5.2 Competitive samples must not show any signs of mishandling or abuse. If products become non-homogeneous during handling, in that they cannot be returned to their original state (precipitates may be returned to solution, but fractured pieces cannot be made whole), then test samples should be remedied for such defects. For example, the last

serving or two from a box of cereal that may have a disproportionate share of fines should be discarded or screened.

5.5.3 To minimize the likelihood of product recognition by respondents, manufacturers sometimes try to "blind" the competitive product. Manipulations beyond labeling the original package should be approached with extreme caution. Repackaging of product would need to be supported by instrumental and sensory tests demonstrating no impact on the product. Any alteration of the product itself to minimize recognition could potentially impact acceptability and should be applied with the utmost discretion. It may be feasible to remove a product from its identifying package, but altering the structure of a product, such as grinding cereals to mask their shape, may change a product beyond the point where the competitive assessment is credible. When a product is instantly recognizable by its appearance, shape, or design, then cognitive factors due to brand recognition or previous experience with the product may contribute to the ratings obtained in the study.

5.6 *Sampling of Product Not Yet on the Market:*

5.6.1 If the manufacturer's product is not yet on the market at the time of testing, the product should represent commercial production, and either be typical retail age of competitive products or expected age of the product when the cycle of the manufacturer's distribution is observed. The competitive product should be selected to represent average retail age at the time of testing. If a suitable product is not available in the test city, the product should be sourced from a nearby location.

5.6.2 To ensure that the claimed benefit of the new product results from the product itself and not from special handling during limited scale production, it is desirable, but may not always be practical, for the new product to have been made at the production facility. A new product, therefore, should be made at its intended manufacturing site, preferably on the same equipment and under normal operating conditions that will be used to manufacture the product. If pilot plant material must be used for claim support, then supplemental testing, for example, discrimination test for similarity, must be conducted to demonstrate that the claim benefits extend to material made at the production facility.

5.7 *Sample Preparation/Test Protocol:*

5.7.1 To minimize bias, it is essential that all samples for testing are prepared and served in a manner that will have limited impact on the perception of the products and in a manner that treats all of the products fairly.

5.7.2 For claims substantiation tests in particular, samples should be prepared and served under reasonably realistic conditions, that is, in a manner consistent with normal consumer practice. Samples should not be prepared in any fashion that would mask or alter various product characteristics.

5.7.3 All samples should be tested blind and with unbiased codes, such as three-digit codes. The respondents should have no leading or biasing information about the products that they are testing or about the overall objective of the study.

5.7.4 A decision must be made regarding the manner in which the samples will be presented to the respondents. For example, the samples can be served as pairs or one at a time (monadic presentation). Differences among samples are more likely to be detected when two or more samples are presented together; however, monadic presentation generally is considered more representative of the consumer experience.

5.7.5 The order of sample presentation must also be considered prior to testing and this must be designated according to a statistical design. Various psychological factors can influence judgment, for example, the impact for which the following order effects must be accounted:

5.7.5.1 *Context/Contrast Effect*—The flavor/texture of one sample can have an influence on the perceived flavor/texture of each subsequent sample.

5.7.5.2 *Positional Bias*—Respondents may be more sensitive to differences in specific samples in a series, such as the first or last sample.

5.7.5.3 *Pattern Effect*—Any pattern in order will be detected quickly.

5.7.5.4 *Ceiling Effects*—This typically occurs when the majority of the scores occur towards the top of a rating scale. When the products are well-liked, there is not a sufficient amount of scale available to the respondents to differentiate the products. Variation in rating scores is compressed, making mean-based statistical tests misleading. Therefore, analysis should be performed using a more robust statistical model that does not have distributional requirements and is less prone to outlier influence such as multinomial logistic regression.

5.7.6 It is essential to balance the order of presentation to distribute these effects across all products.

5.7.7 The test and questionnaire should be designed to be free of all forms of bias. Bias during testing may come from the samples, the test protocol, including the questionnaire, or the test environment, or a combination thereof. Other sections of this guide discuss these issues.

6. Test Design—Consumer Testing

6.1 Monadic designs are those in which a single product is rated by respondents at a time.

6.1.1 Sequential monadic designs require each respondent to evaluate products one at a time and in consecutive order.

6.1.2 Protomonadic tests consist of providing one product, obtaining ratings of that product on a variety of attributes, removing the first product, and replacing with a second product. No monadic ratings are obtained on the second product; instead, a paired-comparison test is conducted.

6.2 Comparative test designs are those in which two or more products are presented to the same respondents to compare the products to each other.

6.3 Comparative claims imply, but are not limited to, comparative designs, where each respondent evaluates two or more products. For comparative claims, paired comparisons are used most frequently. Simultaneous presentation provides the most direct comparison of the products. In some situations, sequential presentation may be needed that introduces execution and sensitivity issues, so there should be a rationale for choosing a sequential (monadic) presentation.

6.3.1 In cases where there are multiple products to be compared, the respondents may be able to evaluate all of the products (complete balanced block design) or a subset of products (an incomplete block design) or only a single product (monadic design). When the products are evaluated in subsets,

overlapping product blocks may be constructed using techniques such as Balanced Incomplete Blocks (BIBS) and Partially Balanced Incomplete Blocks (PBIBS). These Incomplete Block designs may require specialized analysis procedures to construct the correct averages, as outlined in Cochran and Cox (1)[5] and other statistical references.

6.4 Since monadic testing is not the most direct method for making comparisons, it is not always the most desirable approach. Nevertheless, sometimes it may be the only practical method to support comparative claims. For example, some products may require long periods of repeated usage to provide a consumer benefit, which can undermine the ability to make direct comparisons. In this case, product performance can be assessed by giving each product to a different group of consumers and conducting statistical analysis on the ratings. In monadic designs, respondents, as well as products, contribute to the total variation, rendering it less sensitive and larger differences or larger sample sizes are required for significance. It is critical that the groups be matched adequately.

6.5 Non-comparative claims may be supportable by either monadic or sequential-monadic test designs. While a monadic rating may provide a measure free from influences inherent in multi-product, sequential-monadic designs, either approach is sufficient to meet the "reasonable basis" required to make a claim.

6.5.1 Qualitative research, such as focus groups, is not acceptable for claims support since one cannot project their findings to a larger population of consumers.

6.5.2 Both central location (CLT) and home use (HUT) test methods can be acceptable, depending on the specifics of the category and usage. CLTs include all locations other than respondents' homes. These locations may include sensory facilities, mall facilities, field sites, supplier's premises, community centers, or others. Each type of location has some benefits and limitations that must be taken into consideration when projecting results.

6.6 *Data Collection Strategies:*

6.6.1 *Central Location Testing (CLT)*—This method of testing provides maximum control over product preparation and usage. Central location testing assures that the participant actually evaluated the product in question and provides his or her own opinion immediately following evaluation, rather than relying on past usage or recollection. Blind testing often precludes the need to repackage product. In addition, CLTs can provide direct product comparisons, isolate specific attributes, such as color or crunchiness, vanilla flavor, and so forth, and accommodate complex evaluation protocols. They are appropriate for parity and superiority claims.

6.6.1.1 Key limitations are that central location tests usually involve a single product exposure with small amounts of product under conditions that may not closely duplicate typical usage. Questions about whether such exposure can exaggerate trivial differences or whether CLTs provide a basis for forming a preference have been raised. Other limitations that can be controlled are potential for respondents to overhear one another, and testing at times of day that are inappropriate for the product, for example, breakfast cereal in the evening. Where these issues outweigh the limitations inherent to central location testing, home use testing can be considered.

6.6.1.2 Respondents can be intercepted from a public area if they meet the screening criteria or they may be pre-recruited and scheduled for testing (useful when testing is targeted to a specific time of day or where incidence is low). Tests that require special equipment have limited shelf life, or shortened project schedules may not be feasible in mall or intercept type facilities, and are better handled with pre-recruiting.

6.6.2 *Home Use Testing*—The term "Home Use Test" (HUT) refers to tests that allow respondents to use the products in a more natural environment, rather than the controlled environment of a CLT. Since there is still experimenter intervention (product placement and questionnaires) the HUT is not a truly normal use environment; but, it comes closer to how consumers actually use and evaluate products. These HUTs allow for product use that is more typical of normal use conditions, as respondents typically use the products where, when, and how they normally would. HUTs are particularly useful when an overall evaluation of the product cannot be realistically conducted in a CLT environment, or when a feature or benefit must be experienced under normal usage conditions.

6.6.2.1 The choice to use a HUT rather than a CLT to substantiate or evaluate a claim should be determined by the nature of the claim, the amount, type, and length of usage of the product. A very narrow claim about a particular flavor of a pre-made product could very well be evaluated in a CLT, while in an overall claim for suitability would usually require more extended use in the home environment, as with an air freshener.

6.6.2.2 Even if the product as a whole may require an HUT, certain visual, tactile, aural, or olfactory properties of the product may be evaluated in a CLT when the objective is to evaluate salient, non-use characteristics of the products. As an example, respondents could evaluate the look or hand and skin feel, or a combination thereof, of products such as toilet paper and other toiletries, feminine care products, or the aroma of a product. If a claim is being made concerning the context or setting of the actual use, or both, it would still need to be proven on a case-by-case basis that testing a given product outside of the home use environment does not artificially influence consumer behavior or perception.

6.6.2.3 When deciding between a CLT or HUT, one needs to consider the issue of realistic product performance and the ability to generalize the study results to the population that is being targeted by the claim. Certain product categories, such as moisturizing creams, lotions, and acne preparations may require usage over an extended period of time for respondents to evaluate product performance realistically. In such instances, HUT may be the most feasible method for providing realistic

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

performance and evaluations that can be generalized to the population that is being targeted by the claim.

6.6.2.4 A key difference between a CLT and a HUT is the limited experimental control in the HUT. As the HUT provides a more realistic use and evaluation environment, the experimenter has less control over product preparation or use, and must rely on the respondent's ability to recall features of the product use. As in normal use, this recall may be influenced by comments received from family and friends, and a respondent's overall impression of a product may influence his/her recall of particular attributes (for example, halo effect). Often, the usage experience will require sequential product placement and usage in a sequential design. The products may be compared at the end (as in a paired comparison or ranking design) or evaluated after each use (as in sequential monadic design or blocked design). These sequential designs may not be appropriate when the product substantially changes the test environment, so the environment in which a second or later product is applied would not be comparable to the first. Example products could include drain cleaners, mold or mildew removers, or shoe polish. This effect may require that respondents use only a single product (for example, a "monadic" or unblocked ANOVA).

6.6.2.5 Certain test conditions may compensate for some of the issues mentioned in 6.6.2.4 by using simultaneous or split-sample designs. An example could be the case where a respondent cleans one-half of a surface with one product and the other half with a second product. Other cases might include shampoo on different sides of the head or skin cleansers or treatments on different sides of the body or face. Care must be taken so split-sample use is counterbalanced across respondents to avoid potential limitations due to handedness and other biases.

6.7 *Interviewing Techniques:*

6.7.1 *Self-Administered:*

6.7.1.1 Questionnaires independently completed by the respondent are referred to as self-administered. Responses can be collected on paper questionnaires, from on-site computerized questionnaires, or from questionnaires administered over the Internet. Paper copies have the advantage of keeping the original data in its real state for an indefinite period of time. Paper copies of questionnaires can be re-examined as needed if any questions about the data arise. Automated data collection and Internet questionnaires have the advantage of being a direct record of consumer rating, uninfluenced by any possible human bias. The biggest risk in data collection is in the home use environment due to the lack of control over who answers the questionnaire, and therefore, whom the information actually represents, whether collected from paper or automated questionnaires.

6.7.1.2 Self-administered questionnaires can be used in both CLTs and in HUTs. Trained panelists exclusively use self-administered questionnaires.

6.7.1.3 A self-administered questionnaire must be understandable by the respondents with minimal to no verbal instructions by a test administrator. The questionnaire is simple and structured in a logical and unbiased manner. When the questionnaires do not meet these criteria, one-on-one interviewing may be required.

6.7.2 *One-on-One Interviewing Techniques:*

6.7.2.1 One-on-one interviews involve eliciting answers or opinions, or both, from a single respondent through an interviewer, either face-to-face or via telephone.

6.7.2.2 Interviewer training and instruction with practice ensure consistent and flawless execution by all interviewers at all test sites. Instructions include spelling out all actions and their contingencies so that no decisions need to be made by the field agency or the interviewer. Interviewers are thoroughly briefed and practiced before beginning data collection. It is strongly recommended that instructions be tested.

6.7.2.3 Interviewers record respondent responses to questions after they are exposed to a stimulus. The stimulus could be asking a question or testing a product.

6.7.2.4 Interviewer bias can be a major concern and a potential disadvantage with this technique. Double blind testing, where neither the respondent nor the interviewer knows the identity of the sponsor or the products, is imperative. Interviewer bias can be further minimized by using unique code numbers for test products to better mask their identity and make trends more difficult for interviewers to discern.

6.7.2.5 If the questionnaire has several questions, the one-on-one interviewing format is preferred over self-administered questionnaires, since interviewing will prevent respondents from reading ahead or going back, which may influence their answers to other questions.

6.7.2.6 When a claim substantiation study questionnaire involves skipping questions based on the answers to previous questions, referred to as skip patterns, the one-on-one format is recommended over the self-administered format, unless computerized interviewing software is used to ensure correct skips.

6.7.3 *Telephone:*

6.7.3.1 Use of the telephone for claim substantiation support usually will be limited to studies where respondents are not immediately reacting to a stimulus, as they would in a taste, visual, or tactile evaluation, but rather voicing their opinion of a product's performance during actual use or over an extended period of time.

6.7.3.2 Responses can be collected over the telephone from a self-administered questionnaire completed during product usage, or interviewers can ask questions based on respondents' recall of their product experience.

6.8 *Type of Questions:*

6.8.1 *Rankings*—When respondents can compare blocks of more than two (groups of) products, the most direct way to establish superiority or parity within a group of products is through the use of ranking designs. In this case, a respondent is presented with a block of products, either simultaneously or sequentially, and asked to rank the products for preference or other attributes (see section on data analysis). In cases where the advertiser is comparing more products than can be accurately ranked, blocking designs, such as Balanced and Partially Balanced Incomplete Block designs (BIBs and PBIBs) may be used, according to Meilgaard et al **(2)**. Consult a statistician for assistance in constructing these designs.

6.8.2 *Preference*—The choice among two or more alternative products is the most direct way to establish superiority or parity, given adequate sample size.

6.8.3 *Acceptance*—The nine-point hedonic scale traditionally is used for sensory acceptance measurements because it is reliable, valid, and of practical value. In addition to measuring degree of liking of a single product or multiple products evaluated sequentially, it measures degree of differences in acceptance and direction of liking. A large enough difference in mean ratings on an acceptance scale might lead to the researcher making an inference about preference. The hedonic acceptance scale can be used with a wide variety of products and with minimal respondent instruction. Absolute levels of liking can change over time and between groups, but scalar differences between products are reproducible with different groups of respondents. Resulting data lends itself to powerful parametric statistics. Other structured, semi-structured, and numerical scales can be used effectively for acceptance testing. When using other scales, care should be taken that the distributions are relatively normal so parametric statistics can be used. If not, nonparametric statistics should be applied.

6.8.4 *Attribute/Diagnostic*—There are four types of attribute/diagnostic questions in general use: (*1*) hedonic, (*2*) preference, (*3*) just right, and (*4*) intensity.

6.8.4.1 Hedonic scales measure the degree of liking of the level of an individual attribute in a product (for example, measuring the degree of liking of the level of fragrance of a product).

6.8.4.2 Attribute preference scales present questions about individual product attributes, such as the fit of a pair of jeans and the preference between the fit of two products.

6.8.4.3 Just right scales measure the appropriateness of the individual attribute level, for example, too sweet, just right or not sweet enough.

6.8.4.4 Intensity scales measure the strength of an individual attribute, for example, no sweetness to extremely sweet, and questions measuring which product has more or less of a specific attribute(s). In consumer hedonic testing, the researcher must have information that demonstrates that consumers truly understand the meaning of the sensory attribute. For example, consumers may confuse "sourness" and "bitterness" or interpret "creaminess" to mean creamy flavor, creamy texture, or both.

6.8.4.5 It would be inappropriate to use "just right" scales to support an intensity claim for a specific product attribute. Intensity claims must be validated by using intensity scales, where "0" is the anchor for none of the attribute and a higher number such as "9," "11," "15," or "100" is the anchor for an extreme amount of the attribute. For example, the claim "more butter flavor than Brand X", shall only be supported by a significant difference in butter flavor using an appropriate scale for the intensity of butter flavor.

6.9 *Questionnaire Design:*

6.9.1 *Format:*

6.9.1.1 Once the type of response (for example, acceptance, preference, diagnostics, and specific attributes) and attribute terms have been selected, attention should be given to the questionnaire format.

6.9.1.2 Although there is not one perfect questionnaire format, the format of the questionnaire is determined by several considerations.

6.9.1.3 In general, a well-designed questionnaire has the following characteristics:

(1) Includes key components (questions) relevant to the claim,

(2) Excludes questions not needed to support the claim (this will preclude any potential biasing effect of any one question on any other question),

(3) Provides sufficient explanations and clarity to the consumer on its use,

(4) Looks organized and professional,

(5) Is easy to decode, and

(6) Is appropriate to its interviewing method (self- or interviewer-administered).

6.9.1.4 It is recommended that the final questionnaire be tested prior to its use in the claims test. If consumers do not understand a required task or do not comprehend a given attribute, the questionnaire can be modified prior to the quantitative test. Optimally, a small group of consumers (10 to 20) should be used for this purpose; however, company employees not related to the project and untrained in sensory testing can be asked to participate in the assessment of the questionnaire, but not to participate in the study.

6.9.2 *Components*—Generally, there are four major components in a consumer questionnaire:

6.9.2.1 Instructions for respondents or interviewers, or both (if using interviewer-administered questionnaires),

6.9.2.2 General/overall questions,

6.9.2.3 Specific attribute questions, and

6.9.2.4 Classification or demographic questions.

6.10 *Instructions to Respondents*—If the questionnaire is self-administered and no orientation or verbal instructions are given to respondents, the written instructions should be complete, concise, and clear. If the questionnaire is interviewer-administered or an orientation is given, or both, the written instructions only need to be a summary of the evaluation process and directions. Because many consumers do not take enough time to read and understand directions carefully, an orientation, together with brief written instructions, is the recommended procedure. In general, written instructions should be located at the beginning of the questionnaire and include the following items:

6.10.1 The type of product and number of products to be evaluated;

6.10.2 The task manipulation or procedure to be followed by consumers (for example, bite, chew, rub, compress, wipe, smell, apply);

6.10.3 Special directions in handling or using or removing product, or a combination thereof, if required;

6.10.4 An indication of the overall flow or components of the questionnaire;

6.10.5 Examples of the rating technique or questionnaire usage; and

6.10.6 Instructions as to what consumers should do after completion of a sample evaluation and the whole test.

6.11 *Instructions to Interviewers*—These instructions must be clear enough to ensure consistent and flawless execution by all interviewers in all test sites. Adequate instructions spell out all actions and their contingencies so that no decisions need to be made by the field agency or the interviewer. It is strongly recommended that instructions be tested, and that interviewers are thoroughly briefed and practiced before beginning data collection. Interviewers should read from a prepared script to ensure consistency across interviewers, test sites, and sessions.

6.12 *General/Overall Questions*—Under this category, there are the questions that address general or overall impression. Usually, these questions are the most important questions in the test and need to come first. In general, ask only the key questions related to the claim to minimize potential bias in asking additional questions irrelevant to the claim. Examples of general/overall questions include:

6.12.1 Overall acceptance or liking,

6.12.2 Acceptance/liking of broad sensory dimensions, and

6.12.3 Overall preference.

6.13 *Positioning of the Key Product Rating Question*—Product tests almost always have an overall question, such as overall liking, acceptance, ranking, or preference. Placement in the questionnaire for this overall measure is very important in a claim test.

6.13.1 In tests where only overall acceptance/liking or preference is asked, these questions come first by default. Asking multiple overall questions runs the risk of obtaining conflicting results; however, in a more complex questionnaire, for example, with attributes, the position of these questions has to be decided.

6.13.2 In general, questions asked first are assumed to be more free of influences or biases that may affect questions appearing later. The extent to which ratings truly represent product performance is critical if a claim is challenged. When claims are challenged, methodologies are scrutinized, question order and flow are reviewed, and a judgment is made about the extent to which the overall liking/acceptance/ranking/preference rating is free from other-item influences or biases. Questions appearing first will stand up to such scrutiny. In a claims test, more confidence will be placed in data obtained from first-asked questions.

6.13.3 *Recommendation Regarding Where to Position Questions:*

6.13.3.1 *Monadic or Single Product Tests*—Product tests where only one product is experienced and rated.

(1) One question presented at a time by paper, computer, or interviewer. The key question pertaining to the claim should be positioned first. It will be free of influences of other questions and most defensible under scrutiny.

(2) Multiple Questions—Self-Administered—When the questionnaire allows all the items to be read or reviewed, the key question should be placed in the most logically appropriate position. It should appear first if what is needed is the consumer overall and immediate hedonic reaction without consideration of attributes.

6.13.3.2 The key claims question could also be presented at the end of the set, an example would be if all attributes need to be judged or the product used in a specific fashion prior to making a decision. Examples are a personal care product such as shampoo, or a household product such as dish detergent. Other questions can influence individual items since the respondent can read and review the self-administered questionnaire at will.

6.13.3.3 *Multi-Product Tests*—When more than one sample is to be evaluated by a respondent in a monadic sequential presentation, after the first product is evaluated, subsequent ratings will be affected by earlier products seen and the attributes that have been rated. Products must be sequenced (balanced for order of presentation or randomized presentation) to minimize effects of sensory adaptation, fatigue, and contextual effects. The effects of the attributes can only be overcome by having the liking or acceptance question at the end of the questionnaire so that the influence of the attribute ratings affects all products equally. In any multi-product test, placement of the key question must be consistent from product to product.

6.13.3.4 *Two-Sample Comparative Tests*—These tests, where preference or ranking data are obtained, are special cases of multi-product tests. Comparative questions that are to serve as the key data to support a claim should appear first. Therefore, these measures will be free of the influence of other attribute questions that may be asked, and thus will be able to withstand scrutiny.

6.14 *Total Text Context and Presentation Matters*—When designing claims research, the number of products, evaluation methods, and questionnaire development should be considered. Some formats allow only one item to be presented at a time as in interviewer- or computer-administered questionnaires. Other formats allow all questions to be reviewed or considered as in a self-administered paper questionnaire.

6.14.1 Single product studies yield evaluations free of influences from other products. In multiple product tests, the first product experienced and the first question answered is the only rating free of influence and potential bias. Presentation and sampling of all the products in a pretest warm-up session can mitigate some of the position, order, and carryover effects in a multi-product test. The position of a key rating question among many is more important when a single question is presented at a time in a preplanned order. In self-administered paper questionnaires, item order matters less, since all questions are available for review at any time and potentially can influence all other items.

6.15 *Specific Attribute Questions*—If claims are to be based on the attributes, direct questions can be asked. It is important that they be asked alone or positioned first in the questionnaire to avoid potential bias. Attribute questions are of three types and include the following.

6.15.1 Attribute hedonic/liking questions,

6.15.2 Attribute intensity or attribute diagnostic questions, and

6.15.3 Attribute preference.

6.15.4 The attribute hedonic/liking questions collect liking information on specific attributes, for example, liking of the herb combination, sweetness level, absorbency, comfort, or hair shine.

6.15.5 The attribute diagnostic questions collect information on the perceived intensity/level of that attribute, for example, intensity/level of fruitiness, saltiness, and oiliness/warmness. Attribute diagnostic questions are asked using either an absolute intensity scale, for example, none to extreme or a just-about-right scale, for example, too low/just about right/too high. The latter is not very useful for claims support, and deviations from 100 % "just right" are likely to be highlighted by challengers. If the claim has to do with a specific amount of an attribute, then an intensity scale should be used.

6.15.6 Attribute preferences can be determined by questions, such as, "which do you prefer for (state attribute of interest)..."

6.15.7 These attribute questions are used either alone or in combination. When more than one is asked, for example, liking and intensity, the same attribute term should be used. The selection of these terms is critical. However, asking about an attribute in more than one way increases the risk of results that could be viewed as inconsistent, for example, a difference in preference without a difference in liking.

6.15.8 The format used for the attribute questions should allow consumers to properly understand and respond to these questions. To achieve this goal, some considerations include the following:

6.15.8.1 The same type of scale should be used throughout the questionnaire, for example, a nine-point hedonic scale for all attribute liking questions.

6.15.8.2 The same anchors and positioning of the anchors in the hedonic scales should be used.

6.15.8.3 The anchors for the diagnostic questions should be placed in the same positions for all questions.

6.15.8.4 If both attribute liking and diagnostic questions are used, the format and position of both questions should be kept constant all through the questionnaire, for example, both questions for the same attribute positioned side-by-side throughout the questionnaire, or attribute liking question followed by the attribute intensity question throughout the questionnaire.

6.15.9 *Selection of Scale*—The two types of measurement data that can be obtained for attributes are rating and ranking. The selection of a scale is made based on the advantages and disadvantages of each, the ease of its use by consumers, and the type of data to be collected.

6.16 *Classification or Demographic Questions*—These questions are critical to demonstrating congruence between the target population and the target sample. Standard questions include age, gender, ethnicity, income range, frequency/heaviness of use, use of related product formats, for example, homemade versus ready to eat, and brand used most often. Within the questionnaire, questions involving specific brands or product formats must come after product evaluation or there is risk that responses to these questions can impact respondents' behaviors. For example, after a respondent commits to a favorite brand, they may look for and choose that product in a preference test.

6.17 *Preference Questions*—A procedure for asking preference questions is not easily chosen. It generally is accepted that the most effective way to ask the preference question is to ask the respondents which of the products tested they preferred without any reference to the degree of preference that the respondent may have had. The question of offering a no preference choice is subject to various opinions. This area has been discussed for years and likely will continue to be the subject of discussion in the future. Currently, some television networks, and some courts have taken the position that respondents should be given the opportunity to respond directly to an asked "no preference" alternative in the questionnaire. While this approach is generally accepted, it is not without its shortcomings. It is possible that respondents offered a "no preference" choice will choose that option as a way to avoid making a choice but it is also possible that respondents equally prefer both products.

6.17.1 A "no preference" option should be included, because respondents may not always have a product preference. If there are a high number of no preferences for the product category or attribute, making a preference claim is risky; a statistical risk assessment should be conducted.

6.17.2 It is important that users of this guide remember that the above recommended method is one of many approaches currently suggested and opponents may question the validity of a claim based on the above procedure, because they may have conducted testing using a different approach. Also, it is possible that within a given section of industry, there may be a consensus on a particular test format and that preference would be given to that test design over others.

7. Test Location

7.1 When central location consumer tests are conducted in mall facilities, particularly for intercept recruitment, or at the premises of the research supplier or interviewing service (for pre-recruited respondents), a third party location, such as a hotel, may be used. The venue should not have signs or other cues that indicate the sponsor of the test. Testing conducted at the manufacturer's facilities is never acceptable for claims substantiation.

7.2 When the geographic region is suspected to be a factor relevant to a claim, national or regional claims tests should be conducted across a number of geographically dispersed locations. Local claims should sample more than a single site (see 5.2).

7.3 Test facilities must be staffed by an experienced and professional interviewing organization. To avoid bias and achieve double blind testing, the people who prepare the test products should not conduct interviews for any part of that study, unless products are blinded well enough that brand identities cannot be determined (for example, completely repackaged products as opposed to overwrapped). Field supervisors must not identify the test sponsors to any staff involved with the test, and preparers must not discuss the identity of the test products with the interviewers.

7.4 Preparation activities must not impact the interviewing process. The preparation areas must operate quietly to avoid distracting the respondents and interviewers. Ventilation should be adequate to prevent odors from the preparation area to be detectable in the interviewing area, for example, if a

personal care item has a fragrance, or a food item is accidentally burnt. In addition, ventilation systems should provide adequate turnover of air between samples, as well as between respondents to minimize inappropriate carryover. The preparation area must not be visible to respondents. With the exception of tobacco testing, smoking should be forbidden in the interviewing area.

7.5 The testing area should have separate interviewing stations that are sufficiently isolated to avoid voice or visual influence of ongoing interviews on each other.

7.6 Testing often requires refrigeration capacity or cooking facilities and other accoutrements found in most households. Lighting must be adequate to allow the full visual impact of the test products, unless the test calls for intentional masking of appearance.

7.7 Adequate electrical outlets will be needed to test the product. Water supply is necessary for most laundry, cleaning, and food or beverage preparation, skin testing, or personal care product usage.

7.8 The ability to provide good traffic flow is often overlooked. Rooms with a separate entrance and exit may help.

7.9 Each test has different facility requirements and the agency needs to know the specific requirements for the proposed test.

8. Test Execution by way of Test Agencies—Food and Non-Food Testing

8.1 Each test is unique in its requirements and execution. Thorough preparation for a study includes clear strategies for collecting data and benchmarks for planning, preparing, and completing the study. Supplying the product to the agency on the agreed upon date makes the study run smoothly. A well-run study maximizes the potential for accurate and functional study results. Meeting personally with an agency representative to discuss processes, study requirements, and reporting criteria makes it possible for the agency to serve the needs of the client effectively.

8.2 *Protocol Documentation* is supplied to all contracted agencies as early as possible prior to testing. Two weeks is recommended when possible. It is a detail of the study content, including procedural requirements such as screening requirements, storage of test product, supplies, and any other specific study constraints. Fundamental to all Protocol Documentation is communicating all procedural specifics and in as much detail as possible.

8.3 *Time Constraints* include planned or expected test date, the length of time required for each respondent to complete the test, the number of days required for the test, and the time of day. Time constraints help the agency determine the test location (or locations), the personnel required to execute the study, the amount of time to complete the study, and to negotiate the date the client should expect to receive the study results.

8.4 *Test Design* provides specific documentation to enable the agency to complete the test effectively. Proprietary details may or may not be necessary, but a confidentiality document is highly recommended. The agency needs to know the number of respondents required, the number of products to be tested, expanded directions, including temperature of product, application, or function (how the product(s) is to be employed, for example, applied to skin, eaten, or dispensed), and methods desired in documenting study results. Test choices may include paired comparisons, sequential monadic designs, or one of many other multiple product designs. Randomization of the products in the study to minimize position bias is required.

8.5 *Respondent Recruiting/Screening*—The agency will need instructions regarding the respondent's demographic information that may include, amount other things, age range, income, gender, ethnicity, category and brand usage, usage incidence, family size, and regional habitat. Food allergy status should also be documented when appropriate. Other instructions may be necessary for targeted claims. Test timing is factored in to determine respondent availability to meet and complete the test requirements.

8.5.1 The agency is responsible for confirming that the respondents understand and accept their responsibilities before they participate in the study. Informed consent and confidentiality agreements are signed and retained as part of the study.

8.5.2 Criteria for qualifying and terminating (non-qualifying) respondents is the responsibility of the manufacturer or client of the agency. The agency is responsible for maintaining records of qualifying and non-qualifying (terminated) respondents that include clear reasons respondents were terminated.

8.6 *Personnel Requirements*—The agency is responsible for having sufficient personnel at the test site to administer the test. Comprehensive instructions detailing the various roles individuals may have to perform to execute the study successfully are the client's responsibility. On-site product preparation, special handling, serving, storage, and other variables could require additional agency personnel.

8.6.1 Most claims supports require a double blind format if preparation is a part of the product presentation. It is desirable that the preparer and the interviewer be separate individuals to minimize product knowledge.

8.6.2 Complicated questionnaires may require additional personnel to conduct and additional staff to supervise the interviewers. Some sensitive products such as products specific to male or female consumers may require additional training of the interviewers or extra supervision, or both. The manufacturer, client, or agency should role-play the execution or the questionnaire several times to determine a reasonable estimate of the time required per respondent. This enables the testing agency to assign sufficient personnel and to assure the test is conducted appropriately and the respondent does not feel rushed. Self-administered interviews require role play as well, since determining a reasonable amount of the time to complete the questionnaire has to be considered in order for the agency to plan facility space and personnel.

8.7 *Product Requirements*—Consumer tests commonly require product shipment before the test date. The agency needs to know when product is expected to be shipped, expected arrival date, storage requirements (ambient, air-conditioned,

refrigerated, or frozen), the length of time the product needs to be under the prescribed storage conditions, and how the product must be handled once the agency receives the product, plus any special instructions. Assurances of product safety, such as microbial and allergen statements, should be supplied, if appropriate.

8.7.1 Advance planning by the agency is required if there are special instructions in product handling. For example, if a product must be shipped frozen, thawed, and then prepared for a study, the agency needs to schedule its personnel in order to follow their clients' directions. Also, if a product requires assembly that requires a specific skill set, the client must include the time expected for the assembly, provide assembly instructions to the agency, and any special skills required for assembly. Products requiring preparation may also require specialized equipment. The amount of equipment, size, cleaning instructions, temperature, lighting, noise level, and other critical factors are important to communicate to the agency as well. The instructions should clearly indicate how the product is to be presented or displayed, or both, and how the product will be served, as well as portion sizes and other controls that may be necessary.

8.7.2 After the respondent has finished with the product, the agency needs complete disposal instructions to protect the clients' proprietary information. Specify whether the product is secured and must be returned, whether the product can be reused, or how it must be discarded.

8.8 *Facility Requirements* reflect product handling requirements. Product preparation, length of time each panelist needs, type of interview, and various other factors determine facility requirements. Consideration for facilities for nonfood items includes fragrance testing booths and accessibility to home type appliances or rooms (for example, washing machines, stovetops, and toilets). The client is responsible for selecting an agency that meet all the requirements of the test and communicating those requirements adequately to the agency.

8.9 *Interviewer Scripts* maintain consistency in data collection. Deviation from the script could impact study results; therefore, the script must be followed verbatim–no additions, subtractions, or tonal variation.

8.9.1 Some agencies offer scriptwriting services and can work with the client to develop suitable scripts, as well as train agency personnel to execute the interview as stipulated. Whether the agency personnel execute the interview or the client brings in an interviewer, rehearsing the interview questions can protect the integrity of the study.

8.10 *Questionnaires* can be read by an interviewer or self-administered by the respondent either in paper form or by using a computer. Each method may require a different amount of time to complete, so the client must clarify with the agency which method is to be used and a reasonable estimate of time each respondent will need to complete the questionnaire should be determined.

8.10.1 Written instructions on each questionnaire must be consistent so each respondent receives the same stimulus. Slight nuances in instructions can influence the respondents' perspective of the study or of the product, creating another (albeit unintended) variable that can impact study results.

8.10.2 All special instructions must be on each questionnaire, such as the type of writing utensil (for example, No. 2 pencil if questionnaires will be electronically scanned). Techniques such as applying creams, lotions, or cosmetics, wait times if delayed responses are recorded, and any other special instruction should be incorporated into the questionnaire.

8.10.3 Should a computer be employed for the study questionnaire, then data management, formatting, and the method(s) of transferring data must be incorporated into the study documentation provided to the agency.

8.11 *Data Recording and Verification* instructions should be unambiguous to ensure the required data is gathered and retained. The client must communicate to the agency whether the answers to study questions and products tested must be linked to each respondent so the agency can incorporate identification methods. If voluntary statements are solicited, the method of recording those statements must be predetermined. Safeguards must be in place to ensure the agency is not creating data or padding data by interviewing people who did not participate in the study.

8.11.1 Third party observers can be used to verify links between products, questionnaires, and respondents. For claim support, it is wise to incorporate third party observers, even if the observers are only required to confirm all questions have been answered on the questionnaire or just to observe the study. Validating a minimum of 10 % of the cases by phone is standard practice in the industry. If there are anomalies in these 10 %, then up to 100 % of the data should be validated.

8.12 *Data Submission* guidelines are critical for the client to provide the agency. Examples of guidelines needed are: who receives the data, when the data must be submitted, whether interim reports are required, and how to format the data for submission (for example, Web based, written hard copy, E-mail, and so forth). The database used to collect data must be compatible with the data analysis system, and for claim support, the original questionnaires need to be returned to the client as final verification.

9. Documents to Retain in Sensory Claims Substantiation Research

9.1 *Reasons for Retaining Documents*—Good documentation is always important for robust scientific testing, with information clear enough to allow an independent person to duplicate the work. Documents, or records, can be either paper or electronic, or both. They may include some of, but are not limited to, the elements described below. Appropriate elements should be selected based on the type of testing or product category, or both. These records will allow an independent review and evaluation of the following: (*1*) adherence to the existing guidelines for claims substantiation testing; (*2*) the objectivity of the testing procedures; (*3*) the rigor of the implementation; and (*4*) the accuracy of the results obtained. The list below is also useful for elements to consider when planning study design, execution, and analyses.

9.2 The following documents may be asked for by legal reviewers or other stakeholders, including governmental agencies. Consult with your legal staff or company document

retention policy for additional guidance (for example, how long to retain the documents, where to retain the documents (company making the claim, research supplier, etc.)).

(1) Statement of desired claims – dated to show desired claims specified in advance of test implementation.

(2) Test plan, including:

(a) Specific test objective

(b) Action standard/decision criteria

(c) Reference documents guiding research design and approach

(d) Method

(e) Respondents: number, screening document, recruiting method, agency who did recruiting, database drawn from

(f) Products: number, brand, age/use-by-dates, where sourced; agency who purchased products; shipping documents; picture of products at site (photocopy of product labels/ingredient statements); product codes (for example, one lot or multiple lots); building codes

(g) Product Rotation

(h) Test Design

(i) Procedure: describe how respondents experienced products (for example, product usage instructions, average amount of time), and how data were collected; describe test environment;

(j) Respondent instructions, questionnaire(s)/ballot(s)

(3) Fielding/study placement instructions

(4) Interviewer script

(5) Screener

(6) Re-screener

(7) Informed consent

(8) Description of palate cleansers or other ancillary items (where relevant)

(9) Picture or product tray of the product and package as given to respondents

(10) Picture of testing environment/how product actually used in Central Location Test (CLT)

(11) Dates of Testing

(12) Location of Testing

(13) Description of how product was acquired, shipped, stored, handled, prepared, and disposed of during entire test process. For some products, documentation of product retrieval from the consumer.

(14) Documentation of product purchase and shipping

(15) Raw data: paper ballots or electronic raw dataset

(16) Data validation method (for example, if paper ballot data entered in a database)

(17) Data analysis method and results

(18) Final Report circulated internally or externally, or both (for example, product registration)

10. Laboratory Testing Methods

10.1 Laboratory sensory methods that include discrimination and descriptive test methods are intended to determine if a difference exists in the sensory properties of products, and in the case of descriptive methods, to describe and quantify those differences. These methods provide objective data regarding what humans can perceive without regard for personal preference, and are not appropriate for claims of preference or acceptability.

10.1.1 The laboratory sensory methods' application to claim support is intended to be used to communicate:

10.1.1.1 Product attributes,

10.1.1.2 Overall claims of increase, decrease, or equality in a specific attribute(s), and

10.1.1.3 Claims for magnitude of difference between products.

10.1.2 The appropriate application of these methods to claim substantiation requires careful consideration of these factors:

10.1.2.1 It is mandatory that panelists are trained and experienced in the use the selected test method.

10.1.2.2 Panelists must be familiar with the meaning of product attribute descriptors used in the test.

10.1.3 Lack of experience with the test method or misunderstanding about the meaning of attribute descriptors can contribute to inappropriate conclusions being made from the data. ASTM Manual 13 and ASTM Manual 26 contain information on the appropriate application and interpretation of laboratory panel data.

10.2 *Types of Tests:*

10.2.1 *Overall Difference/Discrimination Tests*—These tests determine if a perceptible sensory difference exists between samples. This difference can occur due to any number of reasons including ingredient differences, processing changes, packaging changes, and so forth. Common overall discrimination tests include the following:

10.2.1.1 *Triangle Test*—Three blind-coded samples are presented either simultaneously or successively to panelists. Two of the samples are the same, representing a single sample, while the third represents a different sample. The panelist is required to identify the different sample.

10.2.1.2 *Duo-Trio Test*—The basic set of samples is the same as in the triangle test, but one of the identical samples is labeled as the "reference." The panelist is asked to identify the coded sample that is either the same or different from the reference.

10.2.2 *Attribute Difference Tests*—In these test methods, the attribute of interest is defined prior to testing, and the panelists are trained to be able to identify the attribute in question and select or rate the relative intensity of that attribute. It is not necessary to evaluate every occurring attribute, only the attributes being addressed in the claim.

10.2.2.1 *Directional Difference Test*—This test method is used when determining whether one sample has more of a particular sensory characteristic than another. Two samples are presented, either simultaneously or sequentially, and the respondent chooses one of the samples as having a higher level of the specified characteristics.

10.2.2.2 *Attribute Difference Rating Test*—This test also determines if one or more specific attributes differ between two samples. The intensities of the attributes are measured on rating scales showing several degrees of intensity. One or more specific attributes of the product that relate to the claim are rated. Samples are presented, and the panelists' task is to evaluate and assign each test sample an intensity to reflect the amount of the designated attribute(s).

10.2.3 *Descriptive Analysis Test*—A descriptive test is a complete, detailed, and objective characterization of a product's sensory attributes, measuring some or all of the sensory parameters found in a product or material (visual, auditory, olfactory, kinesthetic, and so forth) using screened, qualified panelists who have been specifically trained for this purpose. This method provides information on perceived sensory attributes and the intensities or strength of each sensory attribute, thus identifying specific differences between products in quantitative terms. See ASTM Manual 13 for details on descriptive methodology.

10.3 *Advantages and Limitations of the Use of Trained Descriptive Panels in Claims Support Research:*

10.3.1 Laboratory panels are useful in objectively determining if and how differences in sensory characteristics are perceived when the claim is attribute or performance focused, not preference or acceptance based. Attributes must be objectively measurable (more butter flavor) as opposed to subjective (better butter flavor). For example, an overall difference test can demonstrate that there is no change to the sensory characteristics of a beverage due to a packaging change, thus allowing the claim "tastes the same as bottled." A trained descriptive panel can show that a specific formulation delivers a claimed benefit, for example, measuring the duration of a fragrance compared to another formulation, allowing the claim "now longer lasting."

10.3.2 Laboratory panels are sensitive tools for detection of both large and small product differences. This sensitivity and precision also is its limitation. Laboratory panelists may find product characteristics and detect differences that typical, untrained consumers cannot. Claims are designed for the consumer so that the consumer should expect to experience the product in the same ways as stated in the claim. If the consumer cannot perceive it, then the claim should not imply that the perceived difference is one they would notice.

10.3.3 If the claim in question is intended to be interpreted as representing consumer experience, then such a claim is tenable only if the relationship between the trained panel's response to products and consumers' evaluation is known. The more descriptive and consumer data converge, the more convincing the claim. In short, converging descriptive data and consumer data make a claim significantly less vulnerable to criticism compared to claims based on descriptive panel data or consumer data that stand alone.

10.3.4 In correlating descriptive and consumer panel data, care should be taken to ensure that there is reasonable translation of terms. For example, trained panel data separates basic taste sweetness from aromatic sweetness, whereas consumers would integrate both together. Correlations may not be possible in cases where consumers do not have the necessary skills to measure or evaluate the attribute(s) in question. For example, trained panel data may support a claim of "more saffron flavor," but most consumers would not be able to measure this claim. Correlations between the trained panel and the consumer may not be necessary in cases where the claim is used to bring public attention to an attribute that might be new or unique. For example, "We've got buzz in every bite."

10.3.5 Note that trained descriptive panelists are different from "experts" who are drawn from personnel who have extensive experience with the product or product category. Experts may or may not be able to express the perception of differences or descriptions regarding products in terms that can be referenced by standards or treated statistically, and are not appropriate for use with ad claim substantiation.

11. Test Design—Laboratory Testing

11.1 The primary goal of laboratory panels, including descriptive, discrimination, and attribute testing, is to provide an objective sensory evaluation of a product. For claims substantiation, evaluation usually focuses on just one or two product attributes rather than a full product description. These tests can be used to support claims about specific product attributes, such as "Ours' is thicker," "It's less sweet," "It has more cheese flavor."

11.2 The test design and questionnaires for laboratory tests should ensure that descriptive/difference data are gathered in an objective and systematic fashion. The test objective and hypothesis should be defined clearly prior to the start of testing. All test procedures should be focused on this objective, such that the test design answers the specific claim that is desired in a short and concise format.

11.3 Panelists used for claim substantiation should be very familiar and experienced with the test method. Descriptive panelists should have extensive training in the descriptive methodology, considerable experience evaluating products, and should be trained specifically on the product under study. References should be used during descriptive training relevant to the product and attribute(s) being evaluated. There should be some documentation of the experience level and type of training received.

11.4 Consider the source of panelists for claims substantiation. If a panel is used that routinely tests the product, there may be some potential for bias. If the panel is familiar with the product, the panel may inadvertently describe it differently; for example, score a particular margarine higher in dairy flavor because the panel is more familiar with that flavor, than if the panel had never seen this product before. If such bias is anticipated, a panel internal to the company is not recommended to substantiate a claim.

11.5 The test design should be reviewed with other members of the technical team and the legal department to ensure accountability for all potential pitfalls.

11.6 *Product Procurement:*

11.6.1 A laboratory panel must test representative product samples. A representative sample is best accomplished by testing replicate samples of each brand that have been obtained at several representative locations and from several different distribution venues. Sample procurement and handling should occur following a strict protocol. All such information should be documented carefully.

11.6.2 Samples should be selected and handled in the same rigorous manner described in Section 5.

11.7 *Experimental Design*—The exact statistical design will need to be determined on a case by case basis; however, the

following describes some of the more important issues that must be considered when a statistical plan is being designed.

11.7.1 Design of discrimination testing is dependent on the selected test method. The number of panelists is key to ensuring adequate power of the test, and should be fully discussed with statisticians and legal prior to conducting the test. The number of samples to be evaluated is dictated by the method; for example, three samples are evaluated in a triangle test. ASTM Manual 26 provides general guidelines for discrimination test designs. Triangle testing is covered in detail in Test Method E1885 and directional difference testing is covered in Test Method E2164.

11.7.2 Design of descriptive panels is covered in ASTM Manual 13. Replications are an essential part of descriptive panel testing and the number of subjects and replications should be determined prior to the test. Three primary types of variability must be accounted for in the design for claims substantiation. These include the following:

11.7.2.1 *Measurement Error*—Repeatability within the individual panelist. This error can be accounted for by having each panelist test a particular sample more than once.

11.7.2.2 *Experimental Error*—Variability between the panelists. This error can be accounted for by using more than one panelist to test each sample.

11.7.2.3 *Product Variability*—Batch-to-batch variation. This error can be accounted for by testing multiple and representative batches of a product.

11.7.3 The number of samples a descriptive panelist evaluates in a session is important, as too many samples could create sensory fatigue. These issues are not likely to be of much consequence in a claims test, due to the fact that the number of samples and the number of attributes being evaluated are usually quite limited.

11.8 *Data Collection:*

11.8.1 For discrimination tests, the type of data collected is determined by the test method.

11.8.2 For claims substantiation, descriptive panelists should individually evaluate each sample. A group consensus format should not be used with descriptive analysis, as it will be subject to questions regarding the potential of group bias.

11.8.3 It is essential to be explicit about the technique the panelists should use to evaluate the samples. During the data collection phase, the panel leader should ensure that the test protocol is strictly followed.

11.9 *Data Analysis* (see Section 14):

11.9.1 Any analyses of data should be reviewed with a statistician.

11.9.2 Data should be analyzed according to the statistical design. A typical analysis for descriptive data would be an initial calculation of means and statistical deviations. Next, analysis of variance is performed to determine significant effects. Finally, a multiple comparison technique, such as Tukey's HSD, is used to determine which samples differed significantly.

11.9.3 The analysis of a duo-trio is based on the probability that, if there is no detectable difference, the different sample will be selected by chance one-half of the time. Analysis of a triangle test is based on the probability that, if there is no detectable difference, the different sample will be selected by chance one-third of the time. Data are analyzed using the binomial or chi-square test.

11.9.4 Analysis of a paired comparison is based on the probability that, if there is no detectable difference, the different sample will be selected by chance one-half of the time. Data are analyzed using a binomial test.

12. Questionnaire Construction

12.1 Questionnaires used in discrimination tests are specified by the test method. The focus of the questionnaire is on the selection of a sample based on its difference from other samples, either for its overall difference or for its higher (or lower) intensity in a specific attribute.

12.2 The main objective of descriptive panel tests is to provide an accurate description of a product in terms of its perceived attributes and their intensities. Questionnaires for trained panels should ensure that data are collected with the goal of obtaining all the necessary information in a concise and easily understood format.

12.3 There are several ways to construct a descriptive panel questionnaire; but more importantly, the questionnaire should be brief, including only the attributes necessary to establish or support the claim(s). A recommended procedure is to use questions from a more comprehensive, established evaluation ballot, and then elect specific attributes needed for the claims testing, thus eliminating or reducing any special training time.

12.4 If specific training is necessary, the training should be accomplished with relevant products or materials, or both, that reference the specific product under study. Select panelists with experience in evaluating similar products or attributes, or both, necessary to authenticate and defend the claim. Pilot testing should be conducted to detect any questionnaire or methodological deficiencies and to confirm applicability and accuracy.

13. Test Facility

13.1 *Environment:*

13.1.1 When selecting a test facility for a claims test, take into consideration such environmental aspects as color, lighting, and air control, specifically temperature and humidity. The evaluation area should also be free of distractions from other panelists, laboratory personnel, or general noise.

13.1.2 It is optimal to use natural colors when selecting the furnishings in the testing area. Make the walls of the evaluation area off-white to prevent unwanted effects of color on the sample product.

13.1.3 Most testing does not require special lighting. Shadow-free illumination at intensity typical of an office area is suitable for most studies. The exception is when visual attributes are being evaluated, requiring the specification and documentation of lighting conditions.

13.1.4 Ideally, the sensory testing area should be maintained at approximately 72°F, with a relative humidity between 45 and 55 %. Ventilation should be such that extraneous odors are eliminated, particularly if fragrances or aromas/flavors are being evaluated.

13.2 *Facility Design*—The facility design and overall space requirements depend on the number and nature of tests conducted and on the type of products. Different designs and layouts are illustrated in ASTM STP 913.

14. Statistical Analyses

14.1 *Paired-Preference Studies*—More than statistical criteria are involved in developing a sampling plan for product tests designed to support advertisement claims. It is widely recognized that attempting to collect a simple random sample is impractical and that cluster samples, for example, multiple city CLTs, with quotas are accepted alternatives. Sections 5.1 and 5.2 detail approaches to sampling to ensure they adequately approximate the population to which the claim is intended to apply. Instead, this section focuses on the analysis of the preference results, addressing the two forms of the claim, superiority and parity, under the assumption that the data sample can be treated as arising from a simple random sample.

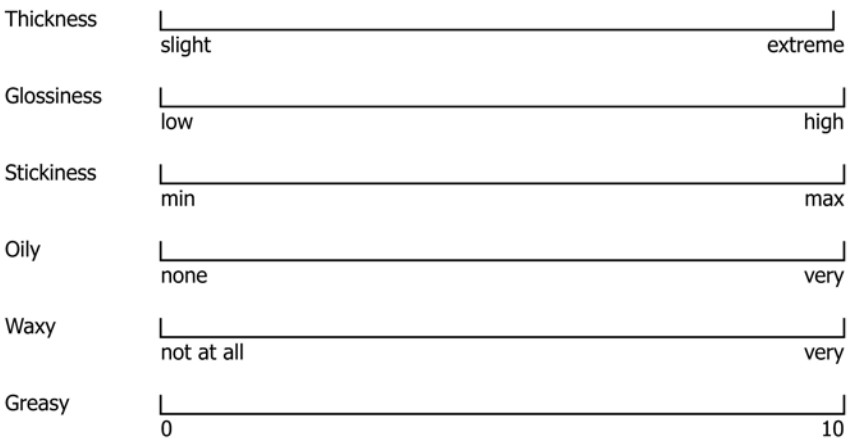
14.2 *Superiority Claims*—A superiority claim is supported if a statistically significant proportion of the respondents prefer the advertiser's product.

14.2.1 A binomial test can be used to analyze the data once the no preference votes are split equally between the two products. If an odd number of people expresses no preference, the extra no preference vote should be awarded to the competitor. Specifically, let n_1 be the number of people preferring the advertiser's product and n_0 be the number of people who expressed no preference. Then set $x = n_1 + \frac{n_0}{2}$ when n_0 is even and $x = n_1 + \frac{(n_0 - 1)}{2}$ when n_0 is odd. The number x can then be compared to the cutoff value for significance found in a binomial table for the desired significance level, such as Table 1, even though the exact distribution is not binomial (see Ennis and Ennis (**3**)). Using this procedure, the Type 1 error will never be greater than the nominal value. To use Table 1, let n be the total number of people participating in the study and compute x as above. For larger sample sizes than those shown in the table, the cut-off value can be determined directly from the binomial distribution. Regardless of the significance level of the test, if the percentage of people expressing no preference is 20 % or more, an unqualified preference claim should not be made (a strong technical rationale is needed to exceed this guideline). Other analytic approaches may be appropriate, but require justification. If the statistical hurdle is not passed from analysis of the total data, the advertiser can still make a preference claim, providing that the analysis, excluding the "no preferences," shows significance, and the advertisement includes a suitable reference to the fact that the claims is based on "those who expressed a preference."

14.2.1.1 *Example 1*—In a preference test among 204 consumers (n), 100 consumers choose the advertiser's product (n_1), 90 choose the competitor's product and there are 14 no preferences (n_0). Splitting the no preferences equally leads to 107 (x) counts for the advertiser's product. With a sample size of 204 (n), we determine from Table 1 that we require 115 preference judgments in favor of the advertiser to declare significance at the 95 % level. The advertiser has not met this requirement and cannot claim superiority based on this test.

14.2.1.2 *Example 2*—In a preference test among 221 consumers (n), 120 consumers choose the advertiser's product (n_1), 90 choose the competitor's product and there are 11 no preferences (n_0). Splitting the no preferences equally leads to 125 (x) counts for the advertiser's product and 96 for the competitor. Note that the left-over no preference count after equal splitting was assigned to the competitor. With a sample size of 221 (n), we determine from Table 1 that we require 124 preference judgments in favor of the advertiser to declare significance at the 95 % level. The advertiser has met this requirement and can claim superiority based on this test.

14.2.2 The ability to detect departures from parity, that is, 50:50 preferences, improves as the number of respondents increase. The number of respondents is under the control of the advertiser, and it is the advertiser who risks missing the opportunity to make a superiority claim when too few respondents participate in the test. As such, this guide does not specify a minimum number of respondents for a preference test to support a superiority claim. To help the advertiser select the number of respondents, Table 2 contains the minimum values



NOTE 1—These scales are for example only. Disparate scales, that is, some using numbers and some words, are not recommended. Consistent scale style is the norm.

FIG. 1 Examples of Scales

"Our sunscreen leaves less sticky residue on your skin."

---

1) Paired comparison method

In front of you are two samples. Please rub the first sample on your left arm, followed by rubbing the second sample on your right arm. You may reuse on a different portion of your skin as often as you wish.

Which samples leaves less sticky residue on your skin? [ ] 212 [ ] 791

Attribute rating

Please use sample #_____ and rate it on the scale below.

Amount of sticky residue

None at all large amount

0 10

Clean your hands with alcohol and cotton ball.

Now Please use sample #_____ and rate it on the scale below.

Amount of sticky residue

None at all large amount

0 10

FIG. 2 Claim: Nonfood Example

of the observed preference proportions required to support a superiority claim for various numbers of respondents. Also presented in Table 2 are two ways to assess the sensitivity of the test for various numbers of respondents. The third column in Table 2 contains the probabilities of detecting a 55:45 % preference split for the various numbers of respondents considered. The final column of the table contains the preference percent that has an 80 % likelihood of being detected for the various numbers of respondents considered. Both of the last columns demonstrate the advantage that larger sample sizes present to the advertiser. These pieces of information can be used together with an assessment of the testing resources available to the advertiser to decide how large of a test needs to be run (see 14.2.3).

14.2.3 In some product use situations such as home or central location use of absorbent or cleaning products, panelists may not experience the full range of the product use situations and will be unable to detect a difference between the products they are comparing or the products may appear so similar that they are unable to form a preference. In these cases where experience is stochastic rather than deterministic, the advertiser may elect to allocate the no-preference votes among the products being tested. The statistical analysis should reflect this allocation, as outlined in David (**4**) and later references. It is the responsibility of the advertiser to demonstrate that the techniques used for the allocation are appropriate, as in Braun (**5**). Consult a statistician for assistance.

14.3 *Parity Claims:*

14.3.1 Failure to conclude that a significant difference in preference exists between two products does not prove that two products are equally preferred. The failure to achieve statistical significance may result from using an insufficient number of respondents, thus yielding an insensitive test. Or failure to find a statistically significant preference may be due to sampling error—sampling procedures resulted in a group of respondents who exhibited no preference. Further, observing a preference percent slightly less than 50 % does not prove that parity does not exist. For superiority claims, the advertiser assumes the risk of an insensitive test; however, when a parity claim is desired, the competitors are at risk from insensitive tests. Larger numbers of respondents are preferred because they both protect the competitor and provide an advantage to the advertiser.

14.3.2 Section 4.4.2 describes parity claims and provides two classes under this category: Equality Claims, and Unsurpassed Claims. In an equality claim, "two products are claimed to be equal in one or more particular features." These features may include specific hedonic or attribute/perception measures or may be included in an overall measure. In an unsurpassed claim "the product(s) selected for comparison is not better/higher (or greater than) in some way to the target product(s) for which the analysis is executed." Specific hedonic or attribute/perception measures may be used.

14.3.3 *Equality Claims*—Since one cannot prove the null hypothesis, an equality claim must be specified in terms of an acceptable range of possible difference within which the products can be considered to be essentially equivalent. For paired preference and difference testing in which the instruction is to choose the product which is most preferred or has the most of some specified attribute (for example, sweetness or moistness), one may choose a 45 %:55 % split for the population as a limit on the meaning of equality. If *either* product equals or exceeds 55 % of the choices, then the products are not considered to be essentially equivalent or "equal" in the meaning of this section. Table 3 provides the required lower

TABLE 1 Counts for One-tailed Superiority Testing

NOTE 1—Minimum number of preference judgments for significant at α = 0.05. Add numbers from the first column to numbers from the first row to get sample size.

n	0	1	2	3	4	5	6	7	8	9
10	9	9	10	10	11	12	12	13	13	14
20	15	15	16	16	17	18	18	19	19	20
30	20	21	22	22	23	23	24	24	25	26
40	26	27	27	28	28	29	30	30	31	31
50	32	32	33	33	34	35	35	36	36	37
60	37	38	38	39	40	40	41	41	42	42
70	43	43	44	45	45	46	46	47	47	48
80	48	49	49	50	51	51	52	52	53	53
90	54	54	55	55	56	57	57	58	58	59
100	59	60	60	61	61	62	62	63	64	64
110	65	65	66	66	67	67	68	68	69	69
120	70	71	71	72	72	73	73	74	74	75
130	75	76	76	77	78	78	79	79	80	80
140	81	81	82	82	83	83	84	84	85	86
150	86	87	87	88	88	89	89	90	90	91
160	91	92	92	93	94	94	95	95	96	96
170	97	97	98	98	99	99	100	100	101	101
180	102	103	103	104	104	105	105	106	106	107
190	107	108	108	109	109	110	111	111	112	112
200	113	113	114	114	115	115	116	116	117	117
210	118	118	119	119	120	121	121	122	122	123
220	123	124	124	125	125	126	126	127	127	128
230	128	129	130	130	131	131	132	132	133	133
240	134	134	135	135	136	136	137	137	138	138
250	139	140	140	141	141	142	142	143	143	144
260	144	145	145	146	146	147	147	148	148	149
270	150	150	151	151	152	152	153	153	154	154
280	155	155	156	156	157	157	158	158	159	159
290	160	161	161	162	162	163	163	164	164	165
300	165	166	166	167	167	168	168	169	169	170
310	170	171	172	172	173	173	174	174	175	175
320	176	176	177	177	178	178	179	179	180	180
330	181	181	182	183	183	184	184	185	185	186
340	186	187	187	188	188	189	189	190	190	191
350	191	192	192	193	193	194	195	195	196	196
360	197	197	198	198	199	199	200	200	201	201
370	202	202	203	203	204	204	205	205	206	207
380	207	208	208	209	209	210	210	211	211	212
390	212	213	213	214	214	215	215	216	216	217
400	217	218	218	219	220	220	221	221	222	222
410	223	223	224	224	225	225	226	226	227	227
420	228	228	229	229	230	230	231	231	232	233
430	233	234	234	235	235	236	236	237	237	238
440	238	239	239	240	240	241	241	242	242	243
450	243	244	244	245	246	246	247	247	248	248
460	249	249	250	250	251	251	252	252	253	253
470	254	254	255	255	256	256	257	257	258	258
480	259	260	260	261	261	262	262	263	263	264
490	264	265	265	266	266	267	267	268	268	269
500	269	270	270	271	271	272	272	273	274	274
510	275	275	276	276	277	277	278	278	279	279
520	280	280	281	281	282	282	283	283	284	284
530	285	285	286	286	287	288	288	289	289	290

choice count to claim equality at the 95 % confidence level for sample sizes of 400 to 1995. Table 4 is the corresponding table for the 99 % confidence level. For example, a paired test with 1000 consumers is conducted in which consumers choose the sweetest sample and 470 chose one product and 530 chose the second product. Since the required lower choice count must fall between 476 and 500 inclusive and the experiment resulted in a value of 470, we cannot declare support for the equality hypothesis. However, if 481 consumers had chosen one of the products, we would have concluded in favor of the equality hypothesis.

TABLE 2 Performance Characteristics of a Preference Test for Superiority Significance Level: Alpha = 5 %

n	*Pc*[A]	Power[B]	80 % Detect[C]
100	58.2	25.8	62.3
200	55.8	40.8	58.7
300	54.7	53.5	57.1
400	54.1	63.9	56.2
500	53.7	72.4	55.5
600	53.4	79.1	55.1
700	53.1	84.3	54.7
800	52.9	88.3	54.4
900	52.7	91.3	54.1
1000	52.6	93.6	53.9

[A] *Pc* = minimum observed percent preference required to claim superiority at the 5 % level of significance.
[B] Power = likelihood of claiming superiority when the actual preference for the advertiser's product is 55 %.
[C] 80 % Detect = actual preference for the advertiser's product that has an 80 % likelihood of being detected.

14.3.4 *Unsurpassed Claims*—The difference between an unsurpassed claim and an equality claim is that in an unsurpassed claim an advertiser may include superiority (preferred or greater or less of some attribute) to establish the claim. This leads to the fact that an unsurpassed claim uses only one of the two limits that were used to define an equality hypothesis. For instance, in an equality preference claim, the preference probabilities must fall between 45 % and 55 %. In an unsurpassed preference claim the preference probabilities may fall above 45 % in favor of the advertiser's product. Sample size requirements for an unsurpassed claim are generally much lower than for an equality claim. Table 5 provides the minimum choice counts for the advertiser's product to make an unsurpassed claim at the 95 % confidence level for sample sizes of 100 to 895. Table 6 is the corresponding table for the 99 % confidence level. Notice that in Table 5 when sample sizes approach and exceed 300, experimental results in which the advertiser's product receives less than 50 % of the choices can support an unsurpassed claim because at this sample size it is still possible to reject the hypothesis that the advertiser's product is inferior to the competitor (that is, the population choice probability is less than 45 %). In fact, at a sample size of 800, the advertiser could obtain results such as 48 % (advertiser):52 % (competitor) and still claim to be unsurpassed by the competitor.

14.3.5 The required minimum of 300 respondents protects the competitor from parity claims resulting from insensitive tests. If the observed preference for the advertiser's product is at least 50 %, based on a 300 respondent test, then the competitor can be 95 % certain that the true preference for the advertiser's product is no lower than 45 %. Increasing the respondent base above 300 allows the advertiser to support an unsurpassed claim with observed preferences slightly less than 50 %, while still protecting the competitor (with 95 % certainty) that the true preference for the advertiser's product is not lower than 45 %. Table 7 contains the minimum preference percentages required to support an unsurpassed claim for various numbers of respondents that protect the competitor from the worst case 45 % preference with 95 % certainty. Table

TABLE 3 Results Required to Support an Equality Hypothesis at the 95 % Level[A]

NOTE 1—In a paired test, the observed lower choice count must fall between the table value and 0.5*n* inclusive to declare support for an equality hypothesis at the 95 % level.

n	count	*n*	count	*n*	count	*n*	count	*n*	count	*n*	count	*n*	count	*n*	count
400	196	600	290	800	384	1000	476	1200	569	1400	661	1600	753	1800	845
405	199	605	293	805	386	1005	479	1205	571	1405	663	1605	756	1805	848
410	201	610	295	810	388	1010	481	1210	573	1410	666	1610	758	1810	850
415	203	615	297	815	391	1015	483	1215	576	1415	668	1615	760	1815	852
420	206	620	300	820	393	1020	486	1220	578	1420	670	1620	762	1820	854
425	208	625	302	825	395	1025	488	1225	580	1425	673	1625	765	1825	857
430	210	630	304	830	398	1030	490	1230	583	1430	675	1630	767	1830	859
435	213	635	307	835	400	1035	493	1235	585	1435	677	1635	769	1835	861
440	215	640	309	840	402	1040	495	1240	587	1440	680	1640	772	1840	864
445	218	645	311	845	405	1045	497	1245	590	1445	682	1645	774	1845	866
450	220	650	314	850	407	1050	500	1250	592	1450	684	1650	776	1850	868
455	222	655	316	855	409	1055	502	1255	594	1455	686	1655	779	1855	870
460	225	660	318	860	411	1060	504	1260	597	1460	689	1660	781	1860	873
465	227	665	321	865	414	1065	506	1265	599	1465	691	1665	783	1865	875
470	229	670	323	870	416	1070	509	1270	601	1470	693	1670	785	1870	877
475	232	675	325	875	418	1075	511	1275	603	1475	696	1675	788	1875	880
480	234	680	328	880	421	1080	513	1280	606	1480	698	1680	790	1880	882
485	237	685	330	885	423	1085	516	1285	608	1485	700	1685	792	1885	884
490	239	690	332	890	425	1090	518	1290	610	1490	703	1690	795	1890	887
495	241	695	335	895	428	1095	520	1295	613	1495	705	1695	797	1895	889
500	244	700	337	900	430	1100	523	1300	615	1500	707	1700	799	1900	891
505	246	705	339	905	432	1105	525	1305	617	1505	709	1705	802	1905	893
510	248	710	342	910	435	1110	527	1310	620	1510	712	1710	804	1910	896
515	251	715	344	915	437	1115	530	1315	622	1515	714	1715	806	1915	898
520	253	720	346	920	439	1120	532	1320	624	1520	716	1720	808	1920	900
525	255	725	349	925	442	1125	534	1325	627	1525	719	1725	811	1925	903
530	258	730	351	930	444	1130	537	1330	629	1530	721	1730	813	1930	905
535	260	735	353	935	446	1135	539	1335	631	1535	723	1735	815	1935	907
540	262	740	356	940	449	1140	541	1340	633	1540	726	1740	818	1940	910
545	265	745	358	945	451	1145	543	1345	636	1545	728	1745	820	1945	912
550	267	750	360	950	453	1150	546	1350	638	1550	730	1750	822	1950	914
555	269	755	363	955	456	1155	548	1355	640	1555	733	1755	825	1955	916
560	272	760	365	960	458	1160	550	1360	643	1560	735	1760	827	1960	919
565	274	765	367	965	460	1165	553	1365	645	1565	737	1765	829	1965	921
570	276	770	370	970	462	1170	555	1370	647	1570	739	1770	831	1970	923
575	279	775	372	975	465	1175	557	1375	650	1575	742	1775	834	1975	926
580	281	780	374	980	467	1180	560	1380	652	1580	744	1780	836	1980	928
585	283	785	377	985	469	1185	562	1385	654	1585	746	1785	838	1985	930
590	286	790	379	990	472	1190	564	1390	657	1590	749	1790	841	1990	933
595	288	795	381	995	474	1195	567	1395	659	1595	751	1795	843	1995	935

[A] Table reproduced with permission from "Tables for Sensory Methods," The Institute for Perception, Richmond, Virginia, 2006.

7 also presents the likelihood that preference tests based on various numbers of respondents will support the unsurpassed claim when the true preference for the advertiser's product actually is at exactly parity, that is, $P = 50$ %. The final column of Table 7 shows how low the actual preference proportion may be, with 95 % certainty, when a 50 % preference result is observed in a study. The information in Table 7 illustrates the advantage of larger sample sizes for the advertiser.

14.3.6 The test statistic used to support unsurpassed claims is as follows:

$$Z = \frac{P - 0.45}{\sqrt{(0.45 \times 0.55/n)}} \quad (1)$$

where:

P = proportion of the respondents who prefer the advertiser's product plus the proportion of respondents that had "no preference," and

n = number of respondents.

14.3.7 If Z is greater than 1.645, the unsurpassed claim is supported at the 5 % (one-tailed) level of significance.

14.3.8 The equality claims tables are based on a one degree of freedom noncentral chi-square distribution. The noncentrality parameter is obtained by converting the equality specifications (55:45, 45:55) into Z scores and squaring them. The unsurpassed tables are based on Eq 1. The tables are constructed so that the counts guarantee that the Type 1 error is 5 % or less (Table 3 and Table 5) or 1 % or less (Table 4 and Table 6).

14.4 *Paired Comparison/Difference Studies*—The technique described in 14.3.2 – 14.3.8 is also used for analyzing data from a paired comparison or paired difference study. In a paired comparison study, each respondent is presented with two samples and is asked to select the sample that has more (or less) of the characteristic of interest. In a sense, a paired preference study is just a special case of a general paired comparison study in which the characteristic of interest is preference.

14.4.1 The same criteria used in the paired preference study for determining numbers of respondents and the number of correct answers needed to support either a superiority or a

TABLE 4 Results Required to Support an Equality Hypothesis at the 99 % Level[A]

NOTE 1—In a paired test, the observed lower choice count must fall between the table value and 0.5*n* inclusive to declare support for an equality hypothesis at the 99 % level.

n	count	*n*	count	*n*	count	*n*	count	*n*	count	*n*	count	*n*	count	*n*	count
700	346	900	440	1100	534	1300	627	1500	720	1700	813	1900	906	2100	999
705	348	905	443	1105	536	1305	630	1505	723	1705	816	1905	908	2105	1001
710	350	910	445	1110	539	1310	632	1510	725	1710	818	1910	911	2110	1003
715	353	915	447	1115	541	1315	634	1515	727	1715	820	1915	913	2115	1005
720	355	920	450	1120	543	1320	637	1520	730	1720	822	1920	915	2120	1008
725	357	925	452	1125	546	1325	639	1525	732	1725	825	1925	918	2125	1010
730	360	930	454	1130	548	1330	641	1530	734	1730	827	1930	920	2130	1012
735	362	935	457	1135	550	1335	644	1535	737	1735	829	1935	922	2135	1015
740	365	940	459	1140	553	1340	646	1540	739	1740	832	1940	924	2140	1017
745	367	945	461	1145	555	1345	648	1545	741	1745	834	1945	927	2145	1019
750	369	950	464	1150	557	1350	651	1550	744	1750	836	1950	929	2150	1022
755	372	955	466	1155	560	1355	653	1555	746	1755	839	1955	931	2155	1024
760	374	960	468	1160	562	1360	655	1560	748	1760	841	1960	934	2160	1026
765	376	965	471	1165	564	1365	658	1565	751	1765	843	1965	936	2165	1029
770	379	970	473	1170	567	1370	660	1570	753	1770	846	1970	938	2170	1031
775	381	975	475	1175	569	1375	662	1575	755	1775	848	1975	941	2175	1033
780	384	980	478	1180	571	1380	664	1580	758	1780	850	1980	943	2180	1036
785	386	985	480	1185	574	1385	667	1585	760	1785	853	1985	945	2185	1038
790	388	990	482	1190	576	1390	669	1590	762	1790	855	1990	948	2190	1040
795	391	995	485	1195	578	1395	671	1595	764	1795	857	1995	950	2195	1042
800	393	1000	487	1200	581	1400	674	1600	767	1800	860	2000	952	2200	1045
805	395	1005	489	1205	583	1405	676	1605	769	1805	862	2005	955	2205	1047
810	398	1010	492	1210	585	1410	678	1610	771	1810	864	2010	957	2210	1049
815	400	1015	494	1215	588	1415	681	1615	774	1815	867	2015	959	2215	1052
820	402	1020	496	1220	590	1420	683	1620	776	1820	869	2020	962	2220	1054
825	405	1025	499	1225	592	1425	685	1625	778	1825	871	2025	964	2225	1056
830	407	1030	501	1230	595	1430	688	1630	781	1830	874	2030	966	2230	1059
835	410	1035	503	1235	597	1435	690	1635	783	1835	876	2035	968	2235	1061
840	412	1040	506	1240	599	1440	692	1640	785	1840	878	2040	971	2240	1063
845	414	1045	508	1245	602	1445	695	1645	788	1845	880	2045	973	2245	1066
850	417	1050	510	1250	604	1450	697	1650	790	1850	883	2050	975	2250	1068
855	419	1055	513	1255	606	1455	699	1655	792	1855	885	2055	978	2255	1070
860	421	1060	515	1260	609	1460	702	1660	795	1860	887	2060	980	2260	1073
865	424	1065	518	1265	611	1465	704	1665	797	1865	890	2065	982	2265	1075
870	426	1070	520	1270	613	1470	706	1670	799	1870	892	2070	985	2270	1077
875	428	1075	522	1275	616	1475	709	1675	802	1875	894	2075	987	2275	1079
880	431	1080	525	1280	618	1480	711	1680	804	1880	897	2080	989	2280	1082
885	433	1085	527	1285	620	1485	713	1685	806	1885	899	2085	992	2285	1084
890	435	1090	529	1290	623	1490	716	1690	809	1890	901	2090	994	2290	1086
895	438	1095	532	1295	625	1495	718	1695	811	1895	904	2095	996	2295	1089

[A] Table reproduced with permission from "Tables for Sensory Methods," The Institute for Perception, Richmond, Virginia, 2006.

parity claim are also used in a paired comparison study. That is, Tables 2 and 3 can be used to analyze the data from a paired comparison study, substituting the characteristic of interest for "preference," where the term occurs in the tables.

14.5 *Analysis of Data from Scales:*

14.5.1 Data from acceptance tests, descriptive-panel studies collected using unstructured line scales, magnitude estimation, or category scales with at least five points are commonly analyzed as continuous data using parametric statistical methods such as analysis of variance. Analysis of variance is used to statistically compare the average ratings of the products in the test, one response at a time.

14.5.2 Both acceptance tests and descriptive analysis panels vary widely in the number of samples involved in the study and in how the samples are distributed to the respondents who participate in the study. These issues determine the form of the analysis of variance model that is appropriate for analyzing the data from the study (see Meilgaard et al. (2) or ASTM Manual 26). For complicated or irregular product-presentation schemes, it may be necessary to consult a statistician to determine the appropriate model to analyze the data.

15. Keywords

15.1 advertisement claim(s); claim substantiation; consumer testing; descriptive testing; sensory laboratory testing; sensory testing

TABLE 5 Results Required to Support an Unsurpassed Hypothesis at the 95 % Level[A]

NOTE 1—In a paired test to declare the advertiser's product unsurpassed at the 95 % level relative to a competitor, the choice count for the advertiser's product must equal or exceed the table counts at the sample sizes indicated.

n	count	*n*	count	*n*	count	*n*	count
100	54	300	150	500	244	700	337
105	56	305	152	505	246	705	339
110	59	310	154	510	248	710	342
115	61	315	157	515	251	715	344
120	63	320	159	520	253	720	346
125	66	325	162	525	255	725	349
130	68	330	164	530	258	730	351
135	71	335	166	535	260	735	353
140	73	340	169	540	263	740	356
145	76	345	171	545	265	745	358
150	78	350	173	550	267	750	360
155	80	355	176	555	270	755	363
160	83	360	178	560	272	760	365
165	85	365	180	565	274	765	367
170	88	370	183	570	277	770	370
175	90	375	185	575	279	775	372
180	92	380	187	580	281	780	374
185	95	385	190	585	284	785	377
190	97	390	192	590	286	790	379
195	100	395	195	595	288	795	381
200	102	400	197	600	291	800	384
205	104	405	199	605	293	805	386
210	107	410	202	610	295	810	388
215	109	415	204	615	298	815	391
220	112	420	206	620	300	820	393
225	114	425	209	625	302	825	395
230	116	430	211	630	305	830	398
235	119	435	213	635	307	835	400
240	121	440	216	640	309	840	402
245	124	445	218	645	312	845	405
250	126	450	220	650	314	850	407
255	128	455	223	655	316	855	409
260	131	460	225	660	319	860	411
265	133	465	227	665	321	865	414
270	135	470	230	670	323	870	416
275	138	475	232	675	326	875	418
280	140	480	234	680	328	880	421
285	143	485	237	685	330	885	423
290	145	490	239	690	332	890	425
295	147	495	241	695	335	895	428

[A] Table reproduced with permission from "Tables for Sensory Methods," The Institute for Perception, Richmond, Virginia, 2006.

TABLE 6 Results Required to Support an Unsurpassed Hypothesis at the 99 % Level[A]

NOTE 1—In a paired test to declare the advertiser's product unsurpassed at the 99 % level relative to a competitor, the choice count for the advertiser's product must equal or exceed the table counts at the sample sizes indicated.

n	count	*n*	count	*n*	count	*n*	count
100	57	300	156	500	251	700	346
105	60	305	158	505	254	705	348
110	62	310	160	510	256	710	351
115	65	315	163	515	259	715	353
120	67	320	165	520	261	720	356
125	70	325	168	525	263	725	358
130	72	330	170	530	266	730	360
135	75	335	172	535	268	735	363
140	77	340	175	540	270	740	365
145	80	345	177	545	273	745	367
150	82	350	180	550	275	750	370
155	85	355	182	555	278	755	372
160	87	360	184	560	280	760	374
165	90	365	187	565	282	765	377
170	92	370	189	570	285	770	379
175	95	375	192	575	287	775	381
180	97	380	194	580	289	780	384
185	99	385	196	585	292	785	386
190	102	390	199	590	294	790	389
195	104	395	201	595	296	795	391
200	107	400	204	600	299	800	393
205	109	405	206	605	301	805	396
210	112	410	208	610	304	810	398
215	114	415	211	615	306	815	400
220	117	420	213	620	308	820	403
225	119	425	216	625	311	825	405
230	122	430	218	630	313	830	407
235	124	435	220	635	315	835	410
240	126	440	223	640	318	840	412
245	129	445	225	645	320	845	414
250	131	450	228	650	323	850	417
255	134	455	230	655	325	855	419
260	136	460	232	660	327	860	421
265	139	465	235	665	330	865	424
270	141	470	237	670	332	870	426
275	143	475	239	675	334	875	428
280	146	480	242	680	337	880	431
285	148	485	244	685	339	885	433
290	151	490	247	690	341	890	436
295	153	495	249	695	344	895	438

[A] Table reproduced with permission from "Tables for Sensory Methods," The Institute for Perception, Richmond, Virginia, 2006.

TABLE 7 Performance Characteristics of a Preference Test for Unsurpassed Significance Level: Alpha = 5 %

n	*Pc*[A]	Power[B]	LL95[C]
100	53.2	26.2	41.8
200	50.8	41.2	44.2
300	49.7	53.8	45.3
400	49.1	64.2	45.9
500	48.7	72.6	46.3
600	48.3	79.2	46.6
700	48.1	84.4	46.9
800	47.9	88.3	47.1
900	47.7	91.4	47.3
1000	47.6	93.6	47.4

[A] *Pc* = minimum percent preference required to claim to be unsurpassed at the 5 % level of significance.
[B] Power = likelihood of claiming to be unsurpassed when the actual preference for the advertiser's product is 50 %.
[C] LL95 = lower limit of a one-sided 95 % confidence interval that represents how low the actual percent preference may be when a 50 % preference proportion is observed in the study.

APPENDIX

(Nonmandatory Information)

X1. COMMONLY ASKED QUESTIONS ABOUT ASTM AND CLAIM SUBSTANTIATION

X1.1 *What is ASTM?*

X1.1.1 Since it was first organized in 1898, ASTM International has grown into one of the largest voluntary standards development systems in the world. ASTM International is a nonprofit organization which provides a forum for producers, users, ultimate consumers and those having a general interest (representatives of government and academia) to meet on common ground and write standards for materials, products, systems, and services. The purpose of ASTM according to its charter is "the development of standards on characteristics and performance of materials, products, systems and services, and the promotion of related knowledge."

X1.1.2 ASTM International believes that technically competent standards result when a full consensus of all concerned parties is achieved and rigorous due process procedures are followed. This philosophy and standards development system ensure technically competent standards have the highest credibility when critically examined and used as the basis for commercial, legal, or regulatory actions. ASTM International standards are developed and used voluntarily. Standards become legally binding only when a government body references them in regulations or when they are cited in a contract. Any item that is produced and marketed as conforming to an ASTM International standard must meet all applicable requirements of that standard.

X1.1.3 From the work of 131 standards-writing committees, ASTM International has published more than 12 000 standards each year. These standards and other related technical information are sold by ASTM International throughout the world. An ASTM International standard is subject to revision at anytime by the responsible technical committee and must be reviewed every five years, and if not revised, either reapproved or withdrawn.

X1.1.4 Committee E18 is a technical committee of ASTM International. The purpose of Committee E18 is to promote knowledge, stimulate research, and develop principles and standards for the sensory evaluation of materials and products. Committee E18 is comprised of nearly 300 industry and academia professionals-food scientists, sensory scientists, psychophysicists, statisticians, psychologists, and other professionals, representing the world's leading universities and Fortune 500 companies. These professionals are at the forefront of new product development technology, designing and applying the appropriate sensory methods for the evaluation of food, beverage, tobacco, household and personal care products, worldwide.

X1.1.5 This guide was recommended, developed and approved by the collective membership of ASTM Committee E18, individuals who are intimately involved with the design and analysis of studies to assess product performance, and who are responsible for the interpretation and communication of their research results to the business and professional communities. As a standard, the recommendations put forth in this document are subject to review by the Society at regular intervals, to assure up-to-date and accurate information.

X1.2 *Why ASTM Committee E18 Developed This Guide:*

X1.2.1 In November of 1990, Committee E18 held a discussion on the increased interest in sensory testing to support advertising claims. Although a number of individuals and groups had made recommendations on how to effectively conduct sensory tests for advertisement claims, there were many inconsistencies between groups.

X1.2.2 Because Committee E18 is composed of sensory professionals whose purpose is to write voluntary industry standards for this field, it seemed logical that they should attempt to review, combine, and filter individual and group recommendations into one document. Those contributing to this document represent both large and small corporations, academicians, and consultants in a wide variety of consumer products categories. The categories include but are not limited to food, beverage, cosmetics, health and beauty aids, and other related products.

X1.2.3 The goal is to provide a document that is straightforward, easy to understand, and implement. The members contributing to this guide bring together many years of experience in designing, implementing, and analyzing these types of tests. The intent is to provide a technically sound document that will be equitable for all including the advertiser, the challenger, and ultimately the consumer.

X1.3 *How Are the Members of This Committee Recruited?*

X1.3.1 After the subcommittee had been approved by the Executive Committee of E18, a general call at the main committee meeting and through ASTM publications was made to all members that this committee was now ready to begin work. Anyone, members of E18 or other interested parties, was invited to participate. The only criteria for members to receive "working documents" is that participants be "active" members, fully participating in both the decision and production processes. Members who do not wish to fully participate are welcome at any meeting to participate in the discussion and vote on issues. At each meeting the members are asked to

encourage anyone in their respective companies for input or to attend the meetings personally.

X1.4 *Who Is the Intended User for This Guide?*

X1.4.1 This guide is written for all those who are involved in evaluating products from a sensory perspective and supporting product claims based upon those evaluations. This encompasses anyone from those who set up product tests, to the end-users of those product test results.

X1.4.2 Within the industries devoted to developing new products or maintaining the competitive edge of existing products, the intended users include sensory evaluation and consumer research professionals, product formulators or developers, marketers, advertisers and copywriters, as well as the consumer advocates and legal professionals who may question or defend such claims.

X1.4.3 Based on the consensus of those in the forefront of current practice, this guide will direct the inexperienced practitioner or peripheral professional through the detailed heart of a complex process.

X1.5 *What Is the Intended Use of the Guide?*

X1.5.1 Claims research usually will be scrutinized by competitors who will critically evaluate all aspects of the methodology and findings. Research must be conducted in a scientifically sound manner, or a claim based upon it will be in jeopardy. Claims research requires expertise in several disciplines, including experimental design, sampling, and statistical data analysis. In addition, methodological expertise also is required because executional factors, and question content can affect the outcome of the research. This guide recommends best practices from a technical perspective based on the expertise and experience of research professionals.

X1.5.2 Ultimately, the advertising media, and in the case of disputes, arbiters, determine the adequacy of research as substantiation for a claim. This guide will not alter these roles. The intent is to assist and strengthen decisions by claimants, competitors, and those who need to evaluate research by identifying technically sound practices, which comprise valid research.

X1.5.3 As a set of guidelines, this guide is not intended to be prescriptive. In many cases, there may be more than one reasonable approach, and the pros and cons of each option must be weighed carefully to determine the best approach. This guide is an aid to judgment, and it is hoped that it will help those with a vested interest in claims substantiation research be knowledgeable about the subject.

X1.6 *What Are the Applications of This Guide?*

X1.6.1 This guide can help those considering advertising claims by discussing the key factors, which can impact the validity of claims substantiation research. As such, it can help readers decide whether to pursue a claims test and design valid research that will have the best chances of withstanding challenge. Another application is to help critically evaluate existing research. This application can be used in one's own research to decide whether it should be used to substantiate a claim or to evaluate others' research to decide whether a challenge is worth pursuing. Media clearance personnel, attorneys, and arbiters can use this guide to help develop positions on the adequacy of research in question.

X1.7 *What Are the Limitations of This Guide?*

X1.7.1 Unlike many physical tests for which ASTM standards have been written, the scope of this guide is too diverse for a uniform specification. It provides guidelines for practices, which comprise scientifically valid claims research. Since no single universal method is specified, claimed conformity with the guidelines cannot substitute for detailed description of the research methodology.

X1.7.2 This guide is not intended to serve as a template or "cookbook" for all situations. Each situation is unique and what is reasonable will be determined by the objectives of the test. There is no panacea; a rationale will be required, and research will always need to be tailored to the situation at hand.

X1.7.3 Discussion of specific methodologies is not intended to limit the types of approaches or methodologies, which could be used in claims substantiation research. Ultimately, any reasonable, methodologically sound approach should be considered for claims support. As in other fields of research, there are a number of issues upon which qualified practitioners do not agree. Where this is the case, the pros and cons of some alternatives are discussed.

X1.8 *How Are The Statistical Criteria Determined?*

X1.8.1 The statistical criteria have been developed through extensive discussions and consensus decisions of the task group participants. For example, a paired-preference test becomes more sensitive as the number of respondents increase. "Sensitivity" in a rigorous statistical sense is based on three criteria: (*1*) the smallest difference in preference proportions (that is, the advertiser's versus the competitor's) that is deemed to be meaningful; (*2*) the probability that the test will be significant when the difference between the preference proportions equals the meaningful difference (that is, the "power" of the test); and (*3*) the level of risk that is deemed acceptable for falsely concluding that a difference in preference exists when, in fact, it does not. Once values for these three criteria are selected, the number of respondents necessary to deliver that level of "sensitivity" can be computed using basic statistical techniques.

X1.8.2 For both superiority and parity claims, it has been decided to protect the competitor against adverse outcomes resulting from insensitive tests. The advertiser has control over the sensitivity of the test, and therefore, is free to increase the number of respondents to values that correspond to his selected levels of acceptable risk without compromising the fair levels chosen for the competitor by the task group.

X1.9 *When is Descriptive Analysis the Best Method to Use for Claim Support?*

X1.9.1 When desiring to demonstrate the strength of one sensory attribute (for example, color, minty, sweet, shine, sticky) is more, less, or equal to that of a competitor.

X1.9.2 When desiring to demonstrate that treatment with the product increases or decreases a specific perceived property (for example, underarm odor, peanut flavor, dry skin).

X1.9.3 Descriptive analysis is not a good method if it is desirable to know about or make a claim about liking, goodness, preference, or any other subjective consumer-type response.

X1.10 *How Does Descriptive Analysis Differ from Tests with Regular Consumers?*

X1.10.1 Descriptive panels are highly trained and behave more like analytical tools or instruments in that they only describe what attributes are perceived and how strong they are. There is no indication of preference or liking.

X1.11 *How Many People Participate in a Panel?*

X1.11.1 Descriptive panels can have smaller base sizes than studies using consumers because they use trained panelists. Trained panelists can show differences between or among samples after effective training and validation because attributes are clearly defined and panelists are familiar and practiced with attribute evaluations. These help reduce data variability. For actual numbers of panelists used in typical descriptive studies, the reader is referred to recent articles in refereed publications such as Journal of Sensory Studies or Food Quality and Preference.

X1.12 *How Many Attributes Are Evaluated by the Panel to Make an Advertising Claim?*

X1.12.1 Only the sensory attributes (terms, properties, characteristics) about which a claim is to be made should be rated for intensity by a panel.

X1.13 *Can a Descriptive Panel in One Geographic Area Test a Product That is Sold or Used in Another?*

X1.13.1 Any descriptive analysis panel, that has been properly trained, can test a product or sample from anywhere in the world. The panel does not represent some segment of the population, but rather, represents the ability of humans to discriminate (detect) and describe properties and their strength.

X1.14 *How Much Training is Necessary to Prepare a Descriptive Panel?*

X1.14.1 The amount of training depends on a number of factors, such as objectives and product complexity.

X1.14.2 If the panel has been trained and validated previously, the training for complex products with complex attributes will still require adequate training for the product type and research objectives. For further information, see ASTM Manual 13 “Descriptive Analysis Testing for Sensory Evaluation.”

X1.15 *What Other Means are Available to Ensure that a Claim is Defensible?*

X1.15.1 At one of the early meetings of the group, we were very fortunate to have Ron Smithies, then director of the National Advertising Division of the Council of Better Business Bureaus (NAD), attend and share some of his expertise on claim support. The NAD is now a self-regulatory unit of the Advertising Self-Regulatory Council (ASRC). He outlined a principle that said that there are three ways to support a claim: consumer data, instrumental data, and trained panel data. If an advertiser can support a claim with two of the three methods, they have a very strong case for the claim. Although instrumental data is not in the scope of this guide, this type of data can be very helpful in support of an advertisement claim. Serious consideration should be given to using two or even three of the types of data when designing studies.

REFERENCES

(1) Cochran, W.G., and Cox, G.M., *Experimental Designs*, 2nd Edition, Wiley, New York, NY, 1957. Reissues 1992 as part of Wiley Classics Library.

(2) Meilgaard, M., Civille, G. V., and Carr, B. T., *Sensory Evaluation Techniques*, 3rd Edition, CRC Press, New York, NY, 1999.

(3) Ennis, D. M. and Ennis, J. M., Accounting for no difference/preference responses or ties in choice experiments, *Food Quality and Preference*, Vol. 23, No. 1, 2011, pp. 13–17.

(4) David, H. A., *The Method of Paired Comparisons*, Charles Griffin & Company, London, 1988.

(5) Braun, V., Rogeaux M., Schneid, N., O’Mahony, M., and Rousseau, B., “Corroborating the 2-AFC and 2-AC Thurstonian Models Using Both a Model System and Sparkling Water,” *Food Quality and Preference*, Vol 15, No. 6, 2004, pp. 501-507.