Designation: ~~E1958 – 07$^{\varepsilon 1}$~~ E1958 – 12

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 («) indicates an editorial change since the last revision or reapproval.

~~[1] NOTE—Editorially corrected 3.1.13 in February 2008.~~

INTRODUCTION

Formats or standards for testing related to 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. 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.

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 of the Council of the Better Business Bureau, Inc. (NAD), *(2)* National Advertising Review Board (NARB), *(3)* one or more media, such as print, broadcast, or electronic media, *(4)* Consumer Advocacy Organizations, and *(5)* 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. Conversely, it can only be speculated about the testing requirements of the future. 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.

This guide demonstrates what a group of professionals who are skilled in the science of testing consider reasonable, and represents an effective method for both defendant and challenger to determine the viability of a claim. The keyword is "reasonable." If a particular aspect of a test is not reasonable 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 application in certain unusual circumstances. Whenever a test protocol has been completed, it should be critiqued for weaknesses in reasonability. If weaknesses are found, corrective action should be taken, since the competition may point out any weakness or discrepancy and challenge the "reasonableness" of the study.

With the importance of "reasonableness," the question remains, "What is reasonable?" Unfortunately, there is no specific answer to that question. The measure of "reasonable" depends on the company making the claim and its approach toward advertising. Some companies are aggressive;

[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 ~~April 1, 2007~~Oct. 15, 2012. Published ~~May 2007~~December 2012. Originally approved in 1998. Last previous edition approved in ~~2006~~2007 as ~~E1958 – 06.~~E1958 – 07^{1}. DOI: ~~10.1520/E1958-07E01.~~10.1520/E1958-12.

others are conservative. It will depend on the nature of the claim and the status of the competitor, the magnitude of the advertising campaign, and the frequency of the advertisement's exposure. Market pressures (such as timing), testing budgets, and the internal dynamics of a company's marketing and legal/regulatory approval departments also affect the interpretation of "reasonable." Competition will consider most tests unreasonable; therefore, it is more important to focus on whether the review board considers the test more reasonable than the competitor's challenge.

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 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.

[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.

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 sub-classified 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.

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

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 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 9).

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 13).

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 (**~~4~~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.

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

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 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 **(82)**. 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.