Designation: E 1958 – 98

AMERICAN SOCIETY FOR TESTING AND MATERIALS
100 Barr Harbor Dr., West Conshohocken, PA 19428
Reprinted from the Annual Book of ASTM Standards. Copyright ASTM

Standard Guide for Sensory Claim Substantiation[1]

This standard is issued under the fixed designation E 1958; 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

No format or standard for testing related to claim substantiation can be considered without a frame of reference for where that format or standard would fit within the legal framework that surrounds the topic. Tests are performed for three basic reasons:

(1) To determine how a product compares to another, usually a competitor or earlier version of itself;

(2) To provide the ability for marketing to use positive references in their presentation of the product to the consumer through advertising or packaging; and,

(3) To determine if a product actually performs within the scope of its intended use.

Whenever a claim is strong, it will be scrutinized closely by competition, and if found inconsistent with a competitor's test data, it could well be challenged in one or more forums. It may be challenged at the National Advertising Division of the Council of the Better Business Bureau, Inc./National Advertising Review Board (NAD/NARB), one or more networks, or in any of a variety of courts. No single test design or standard test will prevent a challenge. The criteria used by each of the potential forums are not identical and are constantly in a state of evolution. What was sufficient five or ten years ago probably would not be acceptable today and what will be required ten years from now is pure conjecture. What can be counted on is that as advocates of their client's positions, attorneys will defend tests they do while questioning, with great detail, every aspect of a competitor's protocol in the attempt to sway the arbiter to agree that they are in the right. So what is one to do? How can a standard be helpful?

This guide demonstrates what a group of professionals, skilled in the art of testing, considers reasonable. This represents a more effective method for both the defendant and the 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 such a departure surely will be scrutinized. Because of the necessity of such departures, the word" should" is used in this guide where 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 you find weaknesses, they should be corrected, since your competition surely will point them out. But what is reasonable? There is no specific answer to that question. What is reasonable will depend on the company making the claim and its posture toward advertising. Some companies are aggressive; 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. It will be affected by market pressures, such as timing, and of course, testing budgets, and the internal dynamics of a company's marketing and legal/regulatory approval departments. You can be certain that your competitor will consider your test unreasonable. This consideration is a given and does not matter. What does matter is that the forum reviewing your test considers it more reasonable than your competitor's challenge.

1. Scope

1.1 This guide covers reasonable practices for designing and implementing sensory tests, which validate claims pertaining only to the sensory characteristics of a product. A claim is a statement about a product, which highlights its advantages, sensory attributes or differences compared to itself or other products to enhance its marketability. Attribute, performance, and hedonic claims, both comparative and noncomparative, are covered. This guide includes broad principles covering selecting and recruiting representative consumer samples, selecting

[1] This guide is under the jurisdiction of ASTM Committee E-18 on Sensory Evaluation of Materials and Products and is the direct responsibility of Subcommittee E18.05 on Sensory Applications—General.

Current edition approved June 10, 1998. Published September 1998.

and preparing products, constructing product rating forms, test execution, and statistical handling of data. This guide was developed by expert practitioners in the field. The intent this guide is to disseminate good testing practices. Validation of claims should be made more manageable if the essence of this guide is followed.

Table of Contents

2. Referenced Documents

2.1 *ASTM Standards:*

E 253 Terminology Relating to Sensory Evaluation of Materials and Products[2]

3. Terminology

3.1 *Definitions*—Terms used in this guide are in accordance with Terminology E 253.

4. Basis of Claim Classification

4.1 A vital step in the substantiation of an advertising claim is the explicit statement of what the claim will be, or what one hopes it will be, prior to actual testing. Providing such a statement to all parties involved in the substantiation process, such as, marketing, marketing research, legal, consumer testing, sensory evaluation, research suppliers, etc., allows a maximum degree of focus in terms of corporate resources, the selection of appropriate test methods, and perhaps most importantly, maximizes the chances of making a reliable business decision about the claim to be made based on the results of substantiation research. It is important, therefore, for all involved parties to meet and agree (perhaps several times) prior to executing substantiation research, in order to communicate objectives and collaborate to provide the best possible results.

4.2 To develop clear statements of claims at an early stage and to develop a rational plan for testing, familiarity with the general classification of advertising claims is important. This familiarity also will facilitate the process of selecting appropriate consumer and sensory testing methods, since there are many tools available to the consumer/sensory testing professional. Each of these tools will answer specific questions and may support one type of claim but not another. The consumer/sensory testing function, therefore, provides an important source of information and experience in this regard, and as such, will provide much of the definition of testing methodology.

4.3 Advertising claims can be divided broadly into two classifications: comparative and noncomparative. The distinction between the two is whether a comparison is being made relative to an existing product, either the advertiser's or the competitor's, or to itself. A discussion of each of these classifications follows.

4.4 Comparative claims deal with comparisons between two or more products. The basis for comparison can be within the same brand, between two brands, or between a brand and the other products in the category.

4.4.1 Comparative claims generally take one of two forms: parity or superiority. Each is further subclassified into two important areas of application: hedonic and attribute/perception. Hedonics broadly applies to the questions of degree of liking and preference (overall, or on a specific attribute); and, attribute/perception applies to questions of perceived intensity or degree in specific product attributes. In superiority claims, combinations of the above can sometimes be found, where superiority is claimed on liking for specific attributes.

4.4.2 *Parity Claims*—Parity claims deal with claiming an equivalent level of performance relative to another brand. In general, parity claims are made relative to a market/category leader. Within parity claims, two additional classes exist: equality claims and unsurpassed claims (see examples below). In equality claims, two products are claimed to be equal in some factor. In unsurpassed claims, the claim is made that the other product is not better/higher in some way. From a statistical standpoint, parity claims may be somewhat more difficult to support than superiority claims. The appropriate null hypothesis must be considered carefully, for example, failure to find a significant difference does not necessarily mean that two products are identical, particularly for the equality claims. This hypothesis will be discussed further in the section on statistical methods. Examples of equality/parity claims include the following types.

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

4.4.2.2 *Attribute/Perception*:

[2] *Annual Book of ASTM Standards*, Vol 15.07.

"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."

4.4.2.3 *Overall Equality*:

"We're just the same, except for the price."
"You'll never know the difference between us and brand X."

4.4.3 Examples of unsurpassed claims include the following types.

4.4.3.1 *Hedonic*:

"No other product is better than our product."
"No other product is more liked for butter flavor."

4.4.3.2 *Attribute/Perception*:

"No other cake is moister 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."

4.4.4 *Superiority Claims*—Superiority claims deal with claiming a higher level of performance relative to another brand. Superiority claims can be against competitive brands ("cleans better than brand Z") or against an earlier formula of the brand ("now more cleaning power than before"). From a statistical standpoint, it can be easier to support a claim of superiority than one of parity, assuming that the superiority actually exists. This is because the null hypothesis is clear (that the two products are the same), and rejecting the null hypothesis indicates that the two products are different in at least one way. Examples of superiority claims include the following types.

4.4.4.1 *Hedonic*:

"Our product tastes better than brand X."
"Our product tastes better than any other."
"Our product is preferred over any other brand."

4.4.4.2 *Attribute/Perception*:

"Our cake is moister 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."

4.5 *Noncomparative/Communications Claims*—This type of claim seeks to communicate something, usually a product benefit or difference, about the product, 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 noncomparative/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 noncomparative 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 The desired claim should precede the test and should not be based solely on a previous outcome that may be fortuitous and not replicable. Unless the test has been designed to explore subgroup analyses specified in advance in the test protocol and the subgroup sample size provides adequate power for such analyses, claims for the subgroup cannot be supported from the test alone. This will prevent a statistically significant yet random event, which is more likely to occur as more statistical tests are conducted, from being mistaken for a real effect; however, if a subgroup result is promising, the test may be repeated with a sample of new members of that subgroup. This sample should be at least as large as that of the initial test and the data from both tests need to support the desired claim.

CONSUMER BASED AFFECTIVE TESTING

5. Defining the Target Population

5.1 *Screening*:

5.1.1 Claims generally apply to the category user population. Sampling from any population other than the general usership, such as purchasers (who are a subset of users), requires 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. Adults with children and pet owners are classic examples of such dichotomies. For example, "Choosy mothers choose Jif™,"[3] is a claim specific to the purchaser and not necessarily 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 recency is not applicable, as for seasonal products or those with a long purchase-repeat cycle, identifying target consumers based upon positive future category usage intent is acceptable. The category should be defined in a way that justifies the selection of competitive products, for example, raisin bran rather than ready to eat cereal. Respondents should not be restricted to exclusive category usage, for example, only eat raisin bran; they also may use alternative products in related categories. 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 usership claims, respondents may be recruited by screening for brand usage; however, this screening should be conducted in a manner that does not allow the respondents to guess what brands are being tested. This can be accomplished by mentioning a number of brands with the brand or brands of interest embedded in the response along with a larger set of brands. Brand usage and frequency of use data also can be collected to help validate the sample composition. Product users can be defined by their response to the

[3] Jif™ is a registered trademark of Proctor and Gamble.

question, “What one brand of this product type do you use more often than any other?” or, “What brands have you used in the last (insert time period appropriate for category)?” If frequency of use is an issue, then the subject also may be asked how often they use the product or how many times they have purchased the product within a certain time frame (see 6.9 on Questionnaire Design).

5.2 *Sampling Techniques*:

5.2.1 Most claims situations refer to product performance as perceived by purchasers or consumers. These situations require sampling, which is projectable to the target population, as described below. Some objective claims, for example, this product has more . . ., can be substantiated by descriptive analysis by a trained panel. These panels are by design screened and trained to provide the highest possible level of descriptive sensory capability, and are not intended to represent typical consumers (see 9.3 on Descriptive Tests).

5.2.2 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 10.9 on Data Collection and Analysis).

5.2.3 The demographics of the test population should match those of the target. These demographics may include the population profile in terms of age, gender, and geography. Respondents also may be screened for their product usage pattern and the sampling density should reflect the geographic distribution of this group.

5.2.4 Use of quotas is helpful to achieve a match between a sample and the target population. Representation of age and sex 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.

5.2.5 If screening is deemed necessary for business reasons, the criteria must be stated in the test protocol and should be as objective as possible. Records must be kept, which indicate why potential subjects are rejected. Screening criteria should not be telegraphed to potential subjects. Subjects should be asked the traditional security screening questions about whether family members work in advertising or marketing or other related industries, including that of the test product.

5.2.6 A single sex sample or otherwise constrained demographic sample only should be employed when consistent with the stated claim and normal product usage. For example, certain products may be used primarily by women or the elderly.

5.2.7 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, not acceptable for claims substantiation.

5.2.8 Caution should be taken to insure that these files are not riddled with samplers, people who may say they use the product(s) being tested to take advantage of paid evaluation, or may not reflect the users’ latest buying habits. It should be verified that respondents have been recruited expressly for the test and have not participated in any consumer test within the past three months or any test within the category for at least six months.

5.2.9 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, factors, such as water hardness, humidity, average ambient temperature, etc., may affect product performance. If there is evidence that such factors do affect product performance, they should be taken into account in selecting test markets.

5.2.9.1 Preference claims have a greater 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.

5.2.9.2 When geographic region is suspected to be a factor relevant to a claim, the geography of subjects should be consistent with the scope of the claim. A national claim should be based on a sample representing all census regions (northeast, southeast, central, and west). A minimum of two markets in each of the four regions should be included. Regional claims should represent at least four markets, which are geographically dispersed across the region.

5.2.10 Use of more than the minimum number of markets is recommended because the sample is more representative, thereby enhancing projectability; and, the impact of (and validity of examining) results in any individual market is minimized.

5.2.11 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. Care should be taken to guard against bias in terms of social and economic groups. Having more than one test site in a city or metropolitan area is helpful in this regard. Sampling bias also can be minimized 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 intercepted.

5.2.12 A concern in selecting markets for the test is that the sample, in total, should represent adequately 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 users. A mix of large and small urban/metro, as well as rural markets is desirable.

5.2.13 It is useful to view the criteria for market selection as factors in an experimental design. After determining the factors, which need to be taken into account, 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 12 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, also is beneficial in convincing others that the test sample is a valid approximation of a probability sample.

5.2.14 Once a target population is defined and is represented adequately by sampling, results from the total sample, and not its subdivisions or subgroups, are what is critical to making a claim. It is not completely unexpected that results among some subgroup would not correspond to overall results. 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. For these reasons, given appropriate sampling from the target population, examination of subgroups is not a sound analytical practice for claims substantiation (see Section 13 on Statistical Analysis).

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 rands 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.

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, which comes in a drink mix and as a reconstituted liquid, both products would have to be tested in their drink mix forms, following the specific directions for preparation given on each product. 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 from high volume stores in the general location of the site of the test site, for example, representative medium-to-large chain supermarkets for food products, or large drug stores for over-the-counter pharmaceuticals. Purchasing products within a 50-mile radius of the test site is recommended. For other test methods, where product is distributed from one location directly to the consumers, samples also should be purchased from high volume stores, even though the 50-mile radius does not apply to each consumer.

5.4.2 The manufacturer's product also should go through the normal distribution chain prior to testing. 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, 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 product sourcing information be documented.

5.4.4 Store bought competitive products should be in the standard size package with the highest unit volume or in similar size, or both, to the test product; however, 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 nonhomogeneous during handling, such 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, which 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 utmost discretion. It may be feasible to replace the handle on a razor, but grinding of cereals may alter product beyond the point where the competitive assessment is credible.

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

5.6.1 If the manufacturer's product is not yet on the market, it should represent commercial production and either be typical retail age of competitive products or expected age due to the

manufacturer's distribution at the time of testing. The competitive product should be selected to represent average retail age at the time of testing. If 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 prevent 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 enhance various product characteristics.

5.7.3 All samples should be tested blind and with three-digit random codes. The respondents should have no leading or biasing information about the products that they are testing nor 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 to be more representative of the consumer experience.

5.7.5 The order of presentation also must be considered. 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; and

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

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

6.1 Monadic designs are those in which a product is rated on a stand alone basis. Comparative designs are those in which two or more products are presented to the same respondents to compare them to each other.

6.2 Noncomparative 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.

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

6.4 Since monadic testing is not the most direct method for making comparisons, it is not 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 (larger differences or larger sample size are required for significance). It is critical that the groups be matched adequately.

6.5 *Data Collection Strategies*:

6.5.1 Qualitative research, such as focus groups, are not acceptable for claims support since their findings are not projectable.

6.5.2 Both central location (CLT) and home use (HUT) methods potentially are acceptable, depending on the specifics of the category and usage. CLTs include all locations other than respondents' homes, including sensory facilities, mall facilities, field sites/supplier's premises, halls/community centers, etc. Each has some benefits and limitations.

6.6 *Central Location Testing*—This method of testing provides maximum control over product preparation and usage. This method assures that the target consumer actually uses the product and provides his or her own opinion then and there rather than relying on recollection. Blind testing often precludes the need to repackage product. CLTs can provide sensitive (head-to-head) comparisons, isolate specific attributes, such as color or flavor, and accommodate complex protocols. They are appropriate for parity and superiority claims.

6.6.1 Key limitations are that central location tests usually involve a single experience with small amounts of product under conditions, which may not closely duplicate ordinary 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, which can be controlled, are potential for respondents to overhear one

another and testing at times of day, which are inappropriate for the product, for example, breakfast cereal in the evening. Where these issues outweigh the limitations inherent in in-home testing, home use testing can be considered.

6.6.2 Respondents can be intercepted or pre-recruited (useful when testing is targeted to a specific time of day or where incidence is low). Tests which require special equipment may not be feasible in malls and lend themselves to pre-recruiting.

6.7 *Home Use Testing*—This method of testing allows for product usage under more typical, but not truly normal, conditions. Respondents can try the products when and how they normally would, and there is opportunity for repeated experience. They are useful when an overall evaluation of products cannot be conducted appropriately in a central location environment.

6.7.1 When attempting to decide if a given claim requires the use of a HUT to be substantiated, what must be determined is if the claim is context, or setting dependent, or both. For example, if a company claimed their air freshener kept a person's home smelling like freshly-cut flowers for 30 days, it is clear a CLT could not adequately represent the context of the use implied by the claim; therefore, a HUT would appear to be a more robust assessment of the claim.

6.7.2 A second issue related to the context, or setting requirement, or both, of a study must be grounded in fact. For example, it would be inappropriate to say that all products of an intimate nature, that is, toiletries, feminine care products, shower gels, must be tested in a HUT due to the way that consumers use them. First, these products legitimately could be evaluated in a CLT if the goal of the research is to gather information on salient, non-use performance, characteristics of the products. For example, it would be entirely appropriate to test toilet paper in a CLT if the objective of the study were to gather information regarding the "look and feel" of the tissue, outside of the context of use. Second, 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 of an intimate nature outside its normal environment artificially influence consumers' subsequent behaviors and evaluation, before a global statement regarding the preferred use of HUTs for a given product type is made. Further, these previous statements are not limited to products of an intimate nature, whose operational definition has yet to be defined clearly based on consumer terminology alone. They are just as relevant to all product categories that involve consumer evaluation gathered in an artificial test environment.

6.7.3 Lastly, besides examining the influence of study context, or setting, or both, when deciding on if using a HUT is warranted over another research approach, the issue of realistic product performance and generalization of study results to a target population must be examined. Certain product categories, that is, moisturizing creams, lotions, acne preparations, may require usage over an extended period of time to evaluate adequately product performance. In such instances, HUTs may be the most feasible method for providing realistic performance that is able to generalize to the target population as a whole.

6.7.4 Key limitations of home use include lack of control, and therefore nonuniformity of preparation and usage, lack of assurance that the respondent actually used the product, and in some instances, reliance on respondents' ability to recall. Family and friends may influence the response. In a HUT, even without direct questions, the influence of some attributes on others (halo effect) can be exacerbated. In addition, to ensure that respondents are rating the intended product, HUT requires sequential product placement. This design has limited sensitivity, relative to a paired design. As a result, in some product categories, HUTs are not suitable for parity claims.

6.8 *Interviewing Techniques*:

6.8.1 *Telephone*:

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 test, but rather voicing their opinion of a product's performance during actual use or over a period of time.

Telephone interviews can serve as a means of collecting data and opinions after respondents have been exposed to a stimulus, for example, calling respondents during/after placement of a product in their homes.

6.8.2 *Self Administered*:

6.8.2.1 Questionnaires completed by the respondent are referred to as self-administered.

6.8.2.2 Self-administration as a data collection method can be used in a variety of types of test methods, that is, respondents can complete a questionnaire in a mall facility, any other central location, or their homes. Responses to even the first question can be affect responses to later ones. Caution should be taken using claims based on questions beyond the first because the influence of earlier questions cannot be eliminated. In addition, when samples are presented in a monadic sequential testing order, bias of the questions asked of the first product may affect the ratings of second and subsequent samples.

6.8.2.3 In short, the most confidence can be placed in the responses to the first question of the first product evaluated and claim based on such data are the most strongly supported. Less confidence can be placed in data obtained from later questions and for products in the later positions. Researchers must be aware of these biasing effects and the potential corresponding weakness in supporting specific claims.

6.8.2.4 Care should be taken in the design of the study questionnaire to ensure that it is understandable by the target population, is simple and structured in a logical, unbiased manner. When the questionnaire does not meet these criteria, another data collection technique, for example, one-on-one, should be implemented.

6.8.2.5 Open-ended questions should not be used for comparative claim substantiation.

6.8.2.6 Trained panel tests (see Section 7) use self-administered questionnaires since respondents are trained and judgments are objective as opposed to hedonic.

6.8.3 *One-on-One Interviewing Techniques*:

6.8.3.1 These approaches involve eliciting answers/opinions from a single respondent via an interviewer.

6.8.3.2 Interviewers, who have been trained according generally to accepted procedures, (for example, Marketing Research Association guidelines), will record responses to questions after respondents are exposed to a stimulus, or asked a question.

6.8.3.3 The major potential disadvantage with this technique is interviewer bias, and variation between interviewers, particularly when the study is conducted in multiple locations, which usually is the case for claims substantiation studies. Interviewers should be practiced thoroughly, and 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 multiple code numbers for test products to better mask their identity and make trends more difficult for interviewers to pick up.

6.8.3.4 If the questionnaire has several questions, a one-on-one format is preferred since it will prevent respondents from reading ahead or going back, which may bias their answers to other questions.

6.8.3.5 When a claim substantiation study questionnaire involves skip patterns, the one-on-one format is recommended over self-administered, unless computerized interviewing software is used to ensure correct skips.

6.9 *Type of Questions*:

6.9.1 *Preference*—The preference question, to establish a choice between two alternative products, is the most direct way to establish superiority or parity, given adequate sample size (see 6.10.6.1, 6.14 on Test Design, and 8.11 on Data Analysis).

6.9.2 *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, one at a time, it measures degrees of acceptance differences and direction of liking, and it indirectly can measure preference(s) between products. 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 subjects. 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.

6.9.3 *Attribute/Diagnostic*—There are four types of attribute/diagnostic questions in general use: hedonic and preference questions about individual product attributes, such as sweetness, which measure degree of liking of the level of sweetness of a product (hedonic scale) and preference between the sweetness levels of two products; just right scales, which measure the appropriateness of the individual attribute level, for example, too sweet, just right or not sweet enough; intensity scales, which 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).

6.9.4 It would be inappropriate to use “just right” scales to support an intensity claim for a specific product attribute. Intensity claims need to be validated by using intensity scales. For example, the claim“ more butter flavor than Brand X”, only should be supported by significant difference in butter flavor using an appropriate scale for the intensity of butter flavor.

6.10 *Questionnaire Design*:

6.10.1 *Components*—Generally, there are four major components in a consumer questionnaire: Instructions to Respondents; General/Overall Questions; Specific Attribute Questions; and Classification or Demographic Questions. In addition, instructions to the interviewers are necessary in the case of interviewer-administered questionnaires.

6.10.2 Once the type of response, for example, acceptance, preference, diagnostics, and the attributes and attribute terms have been selected, attention should be given to the questionnaire format. The format of the questionnaire is determined by:

6.10.2.1 The components of the questionnaire, for example, instructions, general/overall questions, specific questions, demographics), and,

6.10.2.2 The organization of the various components.

6.10.3 Although there is not one perfect questionnaire format, this section focuses on several considerations for structuring a questionnaire format. In general, a well-designed questionnaire has the following characteristics:

6.10.3.1 Includes key components (questions) relevant to the claim;

6.10.3.2 Excludes questions not needed to support the claim. Precludes any potential biasing effect of any question on any other;

6.10.3.3 Provides sufficient explanations and clarity to the consumers or its use;

6.10.3.4 Looks organized and professional;

6.10.3.5 Is easy to decode; and,

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

6.10.4 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 (10–20) should be used for this purpose; however, company employees not related to the project and untrained in sensory testing also can be asked to participate in the assessment of the questionnaire, but not to participate in the study.

6.10.5 *Instructions*—If the questionnaire is self-administered and no orientation, verbal delivery of instructions, is given to respondents, the written instructions need to be complete and clear. If the questionnaire is interviewer-administered, or an orientation is given, or both, the written instructions only need 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 procedure recommended. In general, written instructions should include the following items.

6.10.5.1 The type of product and number of products to be evaluated.

6.10.5.2 The task manipulation procedure to be followed by

consumers, for example, bite, chew, rub, compress, wipe, apply.

6.10.5.3 Special directions in handling/using product, if required.

6.10.5.4 An indication of the overall flow or components of the questionnaire.

6.10.5.5 Examples of the rating technique or questionnaire usage, if required and only for complex techniques or questionnaires.

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

6.10.5.7 Although not recommended, if a complex or lengthy questionnaire is to be used, brief instruction statements ought to be given at the beginning of each questionnaire section.

6.10.6 *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. Examples of general/overall questions include:

6.10.6.1 Overall acceptance/liking;

6.10.6.2 Acceptance/liking of broad sensory dimensions, for example, with attributes; and,

6.10.6.3 Overall preference.

6.10.6.4 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.

6.10.6.5 *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. Product ratings that are fair and reflective of actual consumer response are essential in a claims test.

6.10.6.6 In general, questions asked first are judged to be free of influences or biases that may be present in 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 to 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.

6.10.6.7 *Total Text Context and Presentation Matters*—When setting up a claims support study, the number of products, the method of presenting these products, and the type of questionnaire 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.

6.10.6.8 Single product studies yield products ratings 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 from other products and other questions. Presentation and sampling of all the products in a pretest warm-up session can mitigate some of the position, order, and carry-over effects in a multi-product test. Finally, 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 questionnaires, item order matters less since all questionnaire available for review at any time and potentially can influence all other items.

6.10.7 *Recommendation Regarding Where to Position Questions*:

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

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

(b) *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.

6.10.7.2 The key claims question also could be presented at the end of the set, if all attributes need to be judged as in a personal care product such as shampoo, or a household product, such as dish detergent. Individual items can be influenced by others since the respondent can read and review the self-administered questionnaire at will.

6.10.8 *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 the respondent, 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 only can 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 product equally. In any multi-product test, placement of the key question must be consistent from product to product.

6.10.9 *Two-Sample Comparative Tests*—These tests, where preference or ranking data 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. These measures, therefore, will be free of the influence of other attribute question that may be asked, and thus, will be able to withstand scrutiny.

6.10.10 *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 include the following.

6.10.10.1 Attribute hedonic/liking questions;

6.10.10.2 Attribute intensity or attribute diagnostic questions; and,

6.10.10.3 Attribute preference.

6.10.10.4 The attribute hedonic/liking questions collect liking information on specific attributes, for example, liking of the herb combination, sweetness level, absorbency, comfort, hair shine. The attribute diagnostic questions collect information on the perceived intensity/level of that attribute, for example, intensity/level of fruitiness, saltiness, 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" likely are to be highlighted by challengers. Attribute preferences can be determined by questions, such as, "which do you prefer for ...."

6.10.10.5 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. Bear in mind, however, that asking about an attribute in more than one way increases the risk of results, which could be viewed as inconsistent, for example, a difference in preference without a difference in liking.

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

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

(b) The same anchors and positioning of the anchors in the hedonic scales should be used;

(c) The anchors for the diagnostic questions should be placed in the same positions for all questions; and,

(d) If both attribute liking and diagnostic questions are used, the format and position of both questions should be kept constant throughout 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 through out the questionnaire.

6.10.11 *Selection of Scale*—Once the type of consumer responses have been identified, for example, liking intensity, appropriateness, the type of scale is selected. As in the measure of other sensory responses, different types of scales can be selected.

6.10.12 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. The two types of measurement data that can be obtained for attributes are rating and ranking.

6.11 *Classification or Demographic Questions*—These questions are critical to demonstrating congruence between the target population and the target sample. Standard questions include age, sex, income range, frequency/heaviness of use, use of related product formats, for example, home-made versus ready to eat, and brand used most often. Race may be asked or recorded by observation to help compare the respondent sample to the target population. 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.

6.12 *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 every action 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 pretested, and that interviewers are thoroughly briefed and practiced before beginning data collection.

6.13 *Claim Substantiation with Trained Panels*—Trained panels are used when claiming your product has" more" or "less" of a specific attribute compared to the original formula or another product. Attributes must be objectively measurable (more butter flavor) as opposed to subjective (better butter flavor).

6.13.1 Trained panelists are specialists. Caution should be taken because of their high level of experience with the specific product category, the degree of sensitivity may exceed the "claim expectations" and not reflect end users' perceptions.

6.13.2 Trained panels, discussed in Section 9, are selected for their abilities and trained to discriminate differences, or describe product's sensory properties without regard to personal preferences, or both. Trained panels are intended, therefore, to provide information that more closely resembles that of an analytical tool.

6.13.3 Trained panels also are different from "experts" that are drawn from personnel in the company or outside 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. For information on the appropriate uses of trained panels in claim substantiation, see 9.4.2.

6.14 *Preference Questions*—A procedure of asking preference questions is not easily arrived. Generally, it is accepted that the best way to ask the preference question is to ask the respondent which of the products tested they preferred, either Product 319 or Product 452, with out any reference to the degree of preference the respondent might have had. The difficulty is whether to offer a no preference choice and in what form such a choice, if offered, should be presented. This area has been hotly debated for years and likely will continue to be the subject of discussions in the future. Currently, the NAD, those television networks that have a preferred form and many of the 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 generally is accepted, it is not without its shortcomings. Some scientists have suggested that respondents offered a "no preference" choice will choose that option as a way of avoiding making a choice and that it is this process that inflates the number of "no preferences."

6.14.1 It is recommended that a more preferred form would be not to offer the asked "no preference," but to accept "no preferences" when they are voiced by the respondents. Care should be taken not to imply to the respondent that they must

make a choice or to otherwise force them to guess in any way. If questionnaires are being administered face to face, the interviewer can ask if the respondent is sure that they want to record a "no preference" but no pressure is to be exerted to try to force the respondent to change their position. If a respondent asks whether they can respond with a "no preference," an interviewer should reply as follows: "If it is your opinion that you cannot make a choice you many answer "no preference."

6.14.2 It is important that users of this guide remember that the above recommended method currently is not consistent with the standards that may be applied by the various forums to who the data might have to be presented and that a risk exists that opponents questioning the validity of a claim based on the above procedure may have conducted testing using the more generally accepted form. 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 Central location consumer tests often are conducted in mall facilities, particularly for intercept recruitment, or at the premises of the research supplier or interviewing service (for pre-recruited respondents). Occasionally, a third party location, such as a hotel, may be used. The venue should not have signs or other cues, which indicate the sponsor of the test. Testing conducted at the manufacturer's facilities is never acceptable for claims substantiation.

7.2 When 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. Even local claims should sample more than a single point. Sampling strategies have been discussed in 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 by curious parties, for example, completely repackaged 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 enough to avoid distraction of respondents and interviewers. Ventilation should be adequate to prevent odors from the preparation area to be detectable in the interviewing area, for example, if something is accidentally burnt. 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, which 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, or both. 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 or use small appliances. Water supply is necessary for most food and 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 your test.

8. Test Execution Dealing with Testing Agencies

8.1 Complete written instructions with verbal confirmation should be supplied to the contracted agency well in advance of the planned testing. These instructions will contain much of the content included in the test protocol. To assure data accuracy, these instructions must be complete.

8.2 *Timing*—Expected test date, time of day required, and length of time needed for each panelist to complete the test are foremost. These will help the agency determine test location, if more than one facility is needed and demands on personnel. Include time to complete the test and date you expect to receive results.

8.3 *Test Design*—Provide specifics of your test design for the agency to complete the test effectively, but not proprietary details. The agency needs to know the number of panelists, the number of products, and how to test the products. Test choices may include paired comparisons single product sampling, a sequential monadic design, or one of many other multiple product designs. Randomization of products is required to minimize position bias.

8.4 *Respondent Recruiting/Screening*—The agency needs to know who the respondent will be and the number of each to recruit. Specifics include gender, age range, regional habitat, category and brand usage, usage incidence needed, family size and ethnicity. Other instructions may be needed for targeted products. Test timing will be a factor in determining respondent availability to complete the test requirements.

8.4.1 The respondents must understand and accept their responsibilities in participating in the test. Sometimes, and informed consent form must be signed and kept as part of the test documentation (see Legal and Safety Guidelines).

8.4.2 Termination criteria for respondents not meeting testing needs must be planned and communicated. The need for the field agency to keep records of reason for each termination should be clearly specified in advance.

8.5 *Facility Requirements*—Test facility needs also must be communicated. Selection of a facility should reflect the needs for product preparation, length of time the respondents will be at the facility, and the number of respondents per session.

8.6 *Personnel Requirements*—So the agency can provide adequate personnel at the test site, they need to know what will be required of them for the test. Managing your product may require on site product preparation, special handling, or storage. Any of these could mean additional personnel are needed. Most claims support should include a double blind format if preparation is a part of the product presentation. Specifically, the preparer and the interviewee should be separate individuals to minimize product knowledge. A long interview may need twice as many interviewers and an extra person to supervise. Particularly sensitive male or female products may require special interviewing techniques and extra supervision. Large

respondent bases also require more personnel. Work through the interview yourself so you can communicate an estimate time per panelist to assure the agency assigns sufficient help.

8.7 *Product Requirements*—Consumer tests often require product shipment before the test date. The agency needs to know when it will be expected; how many people will be required to handle the product when it arrives; what kind of storage is needed, that is, ambient, air-conditioned, refrigerated or frozen; and, how long it needs to be under those storage conditions. Some products need to be shipped frozen then thawed and prepared. These products require advance planning for the agency.

8.7.1 Some products need assembly that may require a special skill. Be sure to include the time you expect this process to take in your instructions. Products requiring preparation usually also require serving equipment. You must specify amount of equipment, size needed, and cleaning instructions. Determine how the product is to be presented or displayed, and whether the interviewee or the panelist will serve the product.

8.7.2 Determine if the product needs special storage during the interview process, that is, kept at a specific temperature, kept out of the panelist eyesight, kept in the dark, etc. The panelist may need a rest period between samples. Each station may need to have a timer, a special light, and tape recorder. Special equipment often is supplied, but the agency needs to know it is a part of the test.

8.7.3 Communicate disposal instructions when the panelist has finished with the product. Determine if the product can be reused or if it must be discarded. Determine whether or not the product is secured and if it needs to be returned. Determine whether or not special disposal instructions are required.

8.8 *Interviewer Script*—The interviewer always should follow a script. The script should be without bias and flow smoothly. Interviewers need to be instructed to add nothing to the script. If the interview process is practiced by the project supervisor before test placement, no changes should need to be made in the field.

8.9 *Questionnaire Instructions*—Questionnaires can be voiced by the interviewer or self-administered by the panelist. You should instruct the agency of your choice. Each method will require a different amount of time to complete. Know what the timing is before turning it over to the agency. In either case, be sure instructions are on each questionnaire so each panelist gets the same detail in the same words. Slight nuances in instructions can change a panelist's perspective of the test or the product, creating yet another variable. Specify writing implement if it is important. For example, the panelist may need a No. 2 pencil if an electronic scanner is to be used to read the data. If a delayed response is needed, be sure to include timing in the instructions. A special technique may need to be taught to the panelist, for example, how to sniff a perfume or how to apply a new cream of cosmetic item. Any special instruction should be on each questionnaire.

8.10 *Data Recording and Verification*—Instructions also need to be clear for recording the data gathered. If voluntary statements are solicited from the panelist, determine how these statements need to be recorded. Determine how it would be best to receive the responses, either by comments by category or verbatim. Determine if the panelist needs a number so all questionnaires or products used can be linked to the specific person. Ensure that safeguards are in place to assure the agency is not creating data or interviewing panelists who did not appear.

8.10.1 Data verification can be accomplished by a third party observer. For claim support this may be wise, for any test each questionnaire needs to be checked by a second person to verify the right product is matched with the proper questionnaire, all questions have been answered, and nothing has been missed.

8.10.2 All data and supporting documents need to be held for seven years for legal purposes.

8.11 *Data Submission*—The agency also needs to know when you need your data and how to submit it. Determine whether or not it is desireable to have toplines by phone with a complete report later or if all data transmission need to be in writing; whether an interim report be faxed; whether or not there is a final date for the completed report; who receives the data; and how the data will be formatted. The data base must be compatible with the data analysis system. For claim support, the original questionnaire must be returned as a final verification.

8.11.1 Each test is different and each product has different testing requirements. Thorough preparation, familiarization with the product, the test design, and the agency handling the test will help assure quality results. Meeting the agency representative in person and discussing the instructions and sharing the requirements goes a long way toward making the process easier for everyone. Finally, have the product ready on time. Delays often cause confusion for the agency and their personnel, which can damage your test results in the end.

9. Laboratory Testing Methods Used for Claim Substantiation

9.1 Laboratory sensory methods that include difference and descriptive test methods are intended to determine if a difference exists in the sensory properties of products (see 7.1-7.3), how much a specific characteristic differs among products (see 7.2 through 7.3), and to characterize a product's sensory attributes (see 7.3). These methods are not intended to predict or reflect the ratings of consumers. These methods provide more objective data regarding what can be perceived by humans without regard for personal preference. Their application to claim support is intended to be used for noncomparative or communication claims, overall claims of increase, decrease or equality in a specific attribute(s), or claims for magnitude of difference between products. They also are appropriate in those cases where consumers may not have the ability to measure the attribute(s) of interest due to unfamiliarity with the product category or vocabulary associated with it. These methods are not appropriate for claims of preference or acceptability. The appropriate application of these methods to claim substantiation requires careful consideration of several factors. It is mandatory that panelists be trained to use the test method selected and familiar with the meaning of product attribute descriptors used in the test. 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. For a complete reference on descriptive testing see Manual 13. Manual 26 contains information on other types of trained panels including discrimination.

9.2 *Types of Tests*—Overall difference tests will 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, etc. Difference tests all have some variants, but each creates an arrangement of samples representing a problem for the panelists to solve. A choice of sample is made, and this choice can be designated as either correct or incorrect. The most common overall difference tests include the following.

9.2.1 *Triangle Test*—Three sample are presented either simultaneously or successively. Two are the same, representing a single sample, while the third represents a different sample. The panelist is required to pick out the latter, the different sample.

9.2.2 *Duo-Trio Test*—The basic set of samples is the same as in the triangle test, that is, two identical, one different. In this case, however, one of the identical samples is labeled as the "control." The panelist can be asked to pick the product that is different from the control or the panelist can be asked to pick the product that is the same as the control.

9.2.3 *Other Tests*—A listing and description of additional difference test methods can be found in Manual 26.

9.3 *Attribute Difference Rating Tests*—These tests define how a specific characteristic differs between two samples. These tests focus on one or more specific attribute(s) of concern. These attributes are defined prior to testing, and the panelists are trained so that they are able to identify the attribute in question and scale that attribute based on an appropriate standard. It is not necessary to evaluate every occurring attribute, only the attributes being addressed. These tests provide more specific information than the overall difference testing. The rating scale methods provide the subjects with a scale showing several degrees of intensity. One or more specific attributes of the product type 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.

9.3.1 *Directional Difference*—As one of the simplest and most popular tests, this test method is used when determining whether one sample has more of a particular sensory characteristic than another.

9.3.2 *Others*—Attribute rating tests may have two or more samples. These are appropriate test methods depending on the number of samples. Common applications of attribute rating tests can be found in Manual 26.

9.4 *Descriptive Tests*—A descriptive test is a complete, detailed, and objective characterization of a product's sensory attributes, measuring some or all of the parameters found in a product or material (visual, auditory, olfactory, kinesthetic, etc.), using screened, qualified panelists who have been specifically trained for this purpose. This method provides information on the perceived sensory attributes and the intensities or strength of each sensory attribute, identifying specific differences between products in quantitative terms. The perceived intensity scores then are recorded in relation to absolute or universal scales, which allow the evaluation and comparison of relative intensities among attributes within a product as well as among products tested.

9.4.1 Descriptive tests are appropriate for use when detailed information is required on individual characteristics of the product. Examples of some applications of descriptive testing are as follows:

9.4.1.1 Documenting a products sensory characteristics;

9.4.1.2 Correlating instrumental and chemical measurements with sensory responses;

9.4.1.3 Interpreting consumer product responses by measuring real sensory differences; and,

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

9.4.2.1 Trained descriptive panels are an important tool for understanding products. Furthermore, trained descriptive panels are useful when the relationship between the trained descriptive panelists' response and the consumers' response to the same attributes/products is established empirically. Specifically, descriptive panels are suitable for claims involving sensory attributes, or performance, or both, such as "long lasting fragrance," and "tastes less salty." Data from trained descriptive panels cannot be used for preference or acceptability claims.

9.4.2.2 The main advantage of the trained descriptive panel as a tool for providing data for claims support is its capacity to break down products to individual attributes for detailed analysis, and quantifying those attributes. Trained descriptive panels, thus, are sensitive tools for detection of both large and small product differences. This sensitivity and precision also is its limitation. Trained descriptive panelists may find product characteristics and detect differences that ordinary, or untrained, consumers cannot. In short, the sensitivity of a trained panel means that the generalizability of the findings is limited only to those with similar training and sensitivity.

9.4.2.3 If the claim in question is intended to be interpreted as representing consumer's experience, then such a claim is tenable only if the relationship between the trained panel's response to products and consumer's evaluation are known. The more descriptive and consumer data available 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 panels data or consumer data that stands alone. Standalone trained descriptive data do not allow inferences to be made about consumers unless the relationship has been established empirically.

9.4.2.4 Keep in mind that it is the ordinary consumer who is reading the claim and makes judgments about the product itself. At present, while a claim made in an advertisement or on a product package does not always specify who, that is, an ordinary consumer or a trained panelist, the data are derived from, it is made in such a way as to suggest to the consumer that the consumer or panelist also would experience the product in the same ways as stated in the claim.

9.4.2.5 To summarize, trained descriptive panelists can provide data to support attribute or performance claims only,

not preference or acceptability claims. Furthermore, trained descriptive panel data are most useful when the relationship between the trained descriptive panelists' responses and consumers' responses to the same attributes/products is established empirically.

9.5 *Correlation of Trained Panel and Consumer Data*:

9.5.1 Often, trained panel data needs to be correlated with consumer responses. 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 lump both together.

9.5.2 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.

9.5.3 Correlations between the trained panel and the consumer may not be necessary in cases where the claim is used to bring to public attention an attribute that might be new or unique. For example, "We've got buzz in every bite."

10. Test Design

10.1 The primary goal of trained panels including descriptive, discrimination and attribute testing, is to provide an objective 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," etc.

10.2 The test design and questionnaires for trained panel tests should ensure that descriptive/difference data is gathered in an objective and systematic fashion. The test should be designed with the goal of obtaining the necessary information with a concise and easily understood format.

10.3 The test objective and hypothesis should be defined clearly prior to the start of testing. All test procedures should be aim directly at that objective. Plan the test design to answer the specific claim that is desired, while keeping the test short and to the point.

10.4 Panelists used for claim substantiation should be highly trained. They should have not only considerable experience evaluating products, but also should be trained specifically for the product under study. References should be used during training, as described in other sections of this guide. There should be some documentation of the experience level and type of training received.

10.5 Consider the source of panelists for claims substantiation. If you use a panel that routinely tests your product, there may be some potential for bias. If the panel is familiar with your product, they may inadvertently describe it differently, for example, score a margarine higher in dairy flavor because they are used to that flavor, than if they had never seen this product before. If an internal company panel is being used to make a claim, you may wish to validate the panel against an outside panel source to prevent the criticism of bias on the part of the inside employee panel.

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

10.7 *Sample Procurement*:

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

10.7.2 Samples should be selected and handled in the same rigorous manner described in he consumer test (see 5.3-5.6).

10.8 *Experimental Design*:

10.8.1 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.

10.8.2 Replications are an essential part of trained panel testing. Three primary types of variability must be accounted for in the design for claims substantiation. These include the following.

10.8.2.1 *Measurement Error - Repeatability Within the Individual Panelist*—This error can be accounted for by having each panelist test a particular sample more than once.

10.8.2.2 *Experimental error - variability between the panelists*—This error can be accounted for by using more than one panelist to test each sample.

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

10.8.3 The number of samples a panelist sees in a session is important, for example, too many samples could create fatigue. These issues are not likely to be of much consequence in a claims test since the number of samples and the number of questions about each sample usually are quite limited.

10.9 *Data Collection*:

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

10.9.2 It is essential to be explicit about the method 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.

10.10 *Data Analysis (see Section* 13*)*:

10.10.1 Any analyses of data should be reviewed with a statistician prior to proceeding.

10.10.2 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, that is Tukey's HSD, is used to determine which samples differed significantly.

10.10.3 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 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.

10.10.4 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 the time. Data is analyzed by the t-test.

11. Questionnaire Construction

11.1 The main objective of trained panel tests is to provide an accurate description of a product. Questionnaires for trained panels should ensure that data is collected with the goal of obtaining all the necessary information in a concise and easily understood format. There are numerous ways to ask the trained panelist questions.

11.2 The questionnaire should be brief, including only the attributes necessary to establish or support the claim(s). The questions should be drawn from a more comprehensive, established evaluation ballot, whenever possible, to reduce the necessity of special training.

11.3 If specific training is necessary, it should be accomplished with relevant products, or materials, or both, that reference the specific product under study. The panelists used for these evaluations should be previously established and validated using performance data available with similar products, or attributes, or both, as those to be supported. Pilot testing should be done to detect any questionnaire or methodological deficiencies and confirm applicability and accuracy.

11.4 *Methods*—Panelists should be well trained before executing any of the following tests:

11.4.1 *Paired Comparison*—A product test in which the panelist's task is to identify one of the two products presented as having more or less of a specific attribute;

11.4.2 *Attribute Rating*—Provides a score for each product that yields a measure of its location on a scale and a measure of the magnitude of the difference between products; and,

11.4.3 *Descriptive Analysis*—Provides detailed information on individual aspects of a product.

11.4.4 For the last two of the above methods, a variety of rating scales may be used as shown in Fig. 1. The panelist indicates an answer by placing a mark anywhere on the line.

12. Test Facility

12.1 *Environment*:

12.1.1 The design of a test facility should take into consideration such environmental aspects as color, lighting, and air control, including temperature and humidity.

12.1.2 The furnishings in the testing area should be natural colors, and the walls of the booths should be off-white to prevent unwanted effects on color of the sample product.

12.1.3 Most testing does not require special lighting. In general, booths should have even, shadow-free illumination at an intensity typical of an office area.

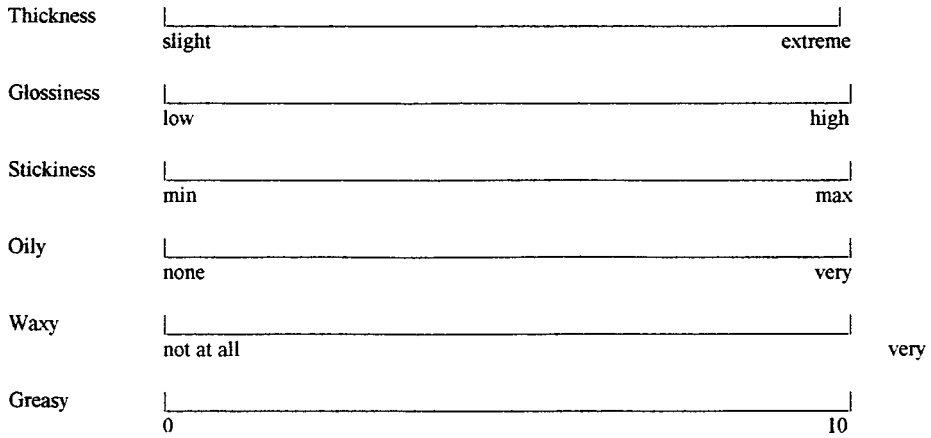
12.1.4 Ideally, the sensory testing area should be maintained at approximately 72°F, with a relative humidity between 45 and 55 %.

12.2 *Facility Design*—The facility design and overall space required depend on the number and types of tests and on the type of products. Different designs and layouts are collected in STP 913.

13. Statistical Analyses for Paired Preference and Trained Panel Data

13.1 *Paired-Preference Studies*—More than statistical criteria are involved in developing a sampling plan for product tests designed to support advertisement claims. It is recognized widely that attempting to collect a simple random sample is impractical and that cluster samples, for example, multiple city CLTs, with quotas are accepted alternatives. It is not the purpose of this section to address the appropriateness of this approach beyond stating that the demographics of the sample should be checked to ensure that 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.

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



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

13.2.1 The test statistic used to support a superiority claim is the following:

$$Z = \frac{P - 0.50}{0.50/\sqrt{n}} \quad (1)$$

where:

P = proportion of the respondents who prefer the advertiser's product, and

n = number of respondents.

If Z is greater than 1.645, the superiority claim is supported at the 5 % (one-tailed) level of significance. For the purpose of supporting an unqualified "preference" claim, "no preference" responses should not be allocated, whether in total or in part, to the advertiser desiring to claim the preferences. Regardless of the significance level of the preference, if the percentage of "no preferences" is 20 % or more, an unqualified preference claim shouldn't be made. In these cases, the preference claim should be made in terms of "those who expressed a preference." 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 claim is based on "those who expressed a preference."

13.2.2 The ability to detect departures from parity, that is, 50:50 preference, improves as the number of respondents increases. 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 1 contains the minimum values of the observed preference proportions required to support a superiority claim for various numbers of respondents. Also presented in Table 1 are two ways to assess the sensitivity of the test for various numbers of respondents. The third column in Table 1 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 number 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.

TABLE 1 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.

13.3 *Parity Claims*:

13.3.1 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. 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. A minimum sample size of 300 per cell, therefore, is required in a preference test being conducted to support a parity claim. Larger numbers of respondents are preferred because they both protect the competitor and provide an advantage to the advertiser.

13.3.2 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 a parity 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 2 contains the minimum preference percentages required to support a parity claim for various numbers of respondents that protect the competitor from the worst case 45 % preference with 95 % certainty. Table 2 also presents the likelihood that preference tests based on various numbers of respondents will support the parity claim when the true preference for the advertiser's product actually is at exactly parity, that is, $P = 50$ %. The final column of Table 2 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 2 illustrates the advantage of larger sample sizes for the advertiser.

13.3.3 The test statistic used to support parity claims is as follows:

TABLE 2 Performance Characteristics of a Preference Test for Parity 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 parity at the 5 % level of significance.

[B] Power = likelihood of claiming parity 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.

$$Z = \frac{P - 0.45}{\sqrt{(0.45 \times 0.55/n)}} \quad (2)$$

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.

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

13.4 *Paired Comparison/Difference Studies*—The technique described in 13.3.2 also is 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 characteristics of interest is preference.

13.4.1 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 parity claim also are used in a paired comparison study, that is, Table 1 and Table 2 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.

13.4.2 Another class of tests that can be used to support advertisement claims is the group of discrimination tests, for example, the duo-trio test and the triangle test. The tables can be used or the duo-trio test because, as with paired-preference tests and the more general paired-comparison tests, the expected proportion of correct responses in the absence of any real difference is 50 %; however, if a triangle test is used, then a different set of tables is needed because the expected proportion of correct responses in a triangle test when no detectable difference exists is only 33 %. The difference in the expected proportion of correct answers affects the number of respondents needed to achieve any given level of sensitivity in the test. Meilgaard et al present statistical tables for various sample sizes for the triangle test used to support either a superiority claim or a parity claim.

13.5 *Analysis of Data from Scales*:

13.5.1 Data from acceptance tests, descriptive-panel studies collected using unstructured line scales, magnitude estimation, or category scales with at least five points are 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.

13.5.2 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 or STP 434). For complicated or irregular product-presentation schemes, it may be necessary to consult a statistician to determine the appropriate model to use to analyze the data.

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 has grown into one of the largest voluntary standards development systems in the world. ASTM is a non-profit 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 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 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 standard must meet all applicable requirements of that standard.

X1.1.3 From the work of 131 standards-writing committees, ASTM has published more than 9000 standards each year. These standards and other related technical information are sold by ASTM throughout the world. An ASTM 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 E-18 is technical committee of ASTM. The purpose of Committee E-18 is to promote knowledge, stimulate research, and the develop principles and standards for the sensory evaluation of materials and products. Committee E-18 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 E-18, 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 E-18 Developed This Guide*:

X1.2.1 In November of 1990, Committee E-18 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 E-18 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 E-18, a general call at the main committee meeting and through ASTM publications had been made to all members that this committee was now ready to begin work. Anyone, members of E-18 or other interested parties, is 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 find themselves questioning or defending 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 decision 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 which 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 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 or 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 specifics. There is no panacea; ingenuity always will be required, and research always will need to be tailored to the situation at hand.

X1.7.3 Discussion of specific methodologies is not intended to limit the types of approached 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 for Parity and Superiority Determined?*

X1.8.1 The statistical criteria related to paired-preference claims have been developed through extensive discussions and consensus decisions of the task group participants. A paired-preference test becomes more sensitive as the number of respondents increases. "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*) meaningful difference (that is, the "Power" of the test); and (*3*) the level of risk that is deemed acceptable for 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.

X1.8.3 For superiority claims, the value selected for the level of risk that is deemed acceptable for concluding that a difference in preference exists when, in fact, it does not is the commonly used Type I error rate of $\alpha = 5\ \%$. No specific values for the smallest meaningful difference and the power of the test were presented for superiority claims because, as pointed out in the text, the advertiser is at risk if one chooses to run a test with a small number of respondents; however, the information given in Table 1 guides the advertiser in selecting an appropriate number of respondents focuses on a 55:45 % split in preference proportions and a power of 80 %. As with the Type I error rate of a –5 %, 80 % is a commonly used value for statistical power. The choice of a 55:45 % split in the preference proportions is the task-group participants' consensus value for the smallest meaningful difference.

X1.8.4 For parity claims, the value selected for the level of risk that is deemed acceptable for concluding that a difference in preference exists when, in fact, it does not is again $\alpha = 5\ \%$. The smallest meaningful difference in the preference proportions is formally set at the consensus value of 55:45 % as evidenced by the test statistic used to determine the validity of the parity claim. If the advertiser's product and the competitor's product were actually at parity, that is, 50:50 %

preference proportions, then the minimum sample size of $n=300$ gives the advertiser a 50 % likelihood of being able to make the parity claim, that is, Power=50 %, while at the same time protecting the competitor with a 95 % level of confidence that the true difference in preference proportions is not greater than the 55:45 % split. Table 2 gives the advertiser information that illustrates how sample sizes larger than the $n=300$ minimum increase his likelihood of being able to claim parity when it exists.

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

X1.9.1 When you want to demonstrate that the strength of one sensory characteristic (color, minty, sweet, shine, sticky) is more, less, or equal to that of a competitor.

X1.9.2 When you want to demonstrate that treatment with your products increases or decreases a specific perceived property (underarm odor, peanut flavor, dry skin).

X1.9.3 When you want to determine the size (strength/intensity) of the difference between a sample and a competitor or your current product (twice as crisp, half the residue, etc.).

X1.9.4 Descriptive analysis is not a good method if it is desireable 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 As few as 8 to 12 panelists participate in a descriptive panel because the level of training insures low variability in the data. This high level of training requires fewer individuals to show small differences between or among samples.

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

X1.12.1 Only the 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 Analysis Panel?*

X1.14.1 If the panel already has been trained to evaluate a product's flavor, texture, aroma, feel, or appearance and has been tested and validated for its ability to discriminate (detect) and describe the product attributes and intensity differences, the panel can be prepared for an advertisement claims attribute test in 1–2 h.

X1.14.2 If the panel has been trained and validated previously, the training for complex products with complex attributes may require significant training.

X1.14.3 If the panel is to be trained just once for this advertisement claim study, hours of training and practice per attribute, depending on the complexity of the product and the complexity of the attribute(s), may be required.