Designation: E 1958 – 98

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 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

[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.

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

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*:

"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

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

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 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