Designation: E2943 – 15 (Reapproved 2021)

Standard Guide for Two-Sample Acceptance and Preference Testing With Consumers[1]

This standard is issued under the fixed designation E2943; 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

This guide is intended to be used by sensory consumer and marketing research professionals (referred to as the "researcher" or "research professional") as an aid to understanding issues associated with and to conducting two-sample acceptance and preference tests with consumers. This guide includes a general summary of considerations and practices for conducting hedonic tests followed by specific considerations and practices for both acceptance and preference testing, including pros and cons of each method. Final sections consider the incorporation of both acceptance and preference testing into the research plan and discuss potential lack of linkage in output/results between them. A flowchart outlining summary of these methods and references for further reading are also included.

1. Scope

1.1 This guide covers acceptance and preference measures when each is used in an unbranded, two-sample, product test. Each measure, acceptance, and preference, may be used alone or together in a single test or separated by time. This guide covers how to establish a product's hedonic or choice status based on sensory attributes alone, rather than brand, positioning, imagery, packaging, pricing, emotional-cultural responses, or other nonsensory aspects of the product. The most commonly used measures of acceptance and preference will be covered, that is, product liking overall as measured by the nine-point hedonic scale and preference measured by choice, either two-alternative forced choice or two-alternative with a "no preference" option.

1.2 Three of the biggest challenges in measuring a product's hedonic (overall liking or acceptability) or choice status (preference selection) are determining how many respondents and who to include in the respondent sample, setting up the questioning sequence, and interpreting the data to make product decisions.

1.3 This guide covers:

1.3.1 Definition of each type of measure,

1.3.2 Discussion of the advantages and disadvantages of each,

1.3.3 When to use each,

1.3.4 Practical considerations in test execution,

1.3.5 Risks associated with each,

1.3.6 Relationship between the two when administered in the same test, and

1.3.7 Recommended interpretations of results for product decisions.

1.4 The intended audience for this guide is the sensory consumer professional or marketing research professional ("the researcher") who is designing, executing, and interpreting data from product tests with acceptance or choice measures, or both.

1.5 Only two-sample product tests will be covered in this guide. However, the issues and recommended practices raised in this guide often apply to multi-sample tests as well. Detailed coverage of execution tactics, optional types of scales, various approaches to data analysis, and extensive discussions of the reliability and validity of these measures are all outside of the scope of this guide.

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

1.7 *This standard does not purport to address all of the safety concerns, if any, associated with its use. It is the responsibility of the user of this standard to establish appropriate safety, health, and environmental practices and determine the applicability of regulatory limitations prior to use.*

1.8 *This international standard was developed in accordance with internationally recognized principles on standardization established in the Decision on Principles for the*

[1] This guide is under the jurisdiction of ASTM Committee E18 on Sensory Evaluation and is the direct responsibility of Subcommittee E18.04 on Fundamentals of Sensory.

Current edition approved Jan. 1, 2021. Published April 2021. Originally approved in 2014. Last previous edition approved in 2015 as E2943 – 15. DOI: 10.1520/E2943-15R21.

Development of International Standards, Guides and Recommendations issued by the World Trade Organization Technical Barriers to Trade (TBT) Committee.

2. Referenced Documents

2.1 *ASTM Standards:*[2]

E253 Terminology Relating to Sensory Evaluation of Materials and Products

E456 Terminology Relating to Quality and Statistics

E1871 Guide for Serving Protocol for Sensory Evaluation of Foods and Beverages

E1958 Guide for Sensory Claim Substantiation

E2263 Test Method for Paired Preference Test

E2299 Guide for Sensory Evaluation of Products by Children and Minors

3. Terminology

3.1 *Definitions:*

3.1.1 For definitions of terms relating to sensory analysis, see Terminology E253.

3.1.2 For terms relating to statistics, see Terminology E456.

3.2 *Definitions of Terms Specific to This Standard:*

3.2.1 *α (alpha) risk, n*—probability of concluding that a difference in liking or preference exists, when, in reality, one does not.

3.2.1.1 *Discussion*—Also known as Type I error or significance level.

3.2.2 *β (beta) risk, n*—probability of concluding that no difference in liking or preference exists, when, in reality, one does.

3.2.2.1 *Discussion*—Also known as Type II error.

3.2.3 *hedonic continuum, n*—hypothesized underlying continuous dimension measured by acceptance scales.

3.2.3.1 *Discussion*—It is presumed to run from strong disliking through a neutral region and onto strong liking.

3.2.4 *labeled affective magnitude scale, n*—labeled magnitude scale (LMS) is a hybrid scaling technique using a verbally labeled line with quasi-logarithmic spacing between each label and the scale consists of a vertical line, which is marked with verbal anchors describing different intensities (for example, "weak," "strong").

3.2.4.1 *Discussion*—Typically, subjects are instructed to place a mark on the line where their perceived intensity of sensation lies, with the upper limit of the scale being the strongest imaginable sensation (**1**).[3]

3.2.5 *Likert scale, n*—attitude scales that can be constructed in an "agree-disagree" format (**2**).

3.2.5.1 *Discussion*—The Likert-type scale calls for a graded response to each statement. The response is usually expressed in terms of the following five categories: strongly agree (SA), agree (A), undecided (U), disagree (D), and strongly disagree (SD). The individual statements are either clearly favorable or clearly unfavorable (**2 and 3**).

3.2.6 P_{max}, *n*—used in forced choice preference measures; a test sensitivity parameter established before testing and used along with the selected values of α and β to determine the number of respondents needed in a study.

3.2.6.1 *Discussion*—P_{max} is the proportion of common responses that the researcher wants the test to be able to detect with a probability of 1 – β. For example, if a researcher wants to have a 90 % confidence level of detecting a 60:40 split in preference, then P_{max} = 60 % and β = 0.10.

3.2.7 *risk, n*—possible consequences to the researcher's client when the test leads to an incorrect conclusion.

3.2.7.1 *Discussion*—Risk around decisions made based on research test results can be grouped into two types, loosely called a "false positive" (when the test detects a difference that does not exist) and a "false negative" when the study does not detect a true difference. In the case of a false positive, the company spends development time and resources on an alternative that does not deliver the intended effect. In the case of a false negative, the product developer or the company will miss a product opportunity and waste resources developing alternatives.

3.2.8 *sequential monadic, adj*—refers to the presentation or ordering in which respondents evaluate products or stimuli.

3.2.8.1 *Discussion*—In a sequential monadic test, the respondent is presented with one product at a time to evaluate.

3.2.9 *sign test, n*—statistical hypothesis test that can be used to compare two samples or a sample with a standard.

3.2.9.1 *Discussion*—No assumption is made about the shape or parameters of the population frequency distribution with the sign test and only the sign of the difference is considered.

3.2.10 *student's t test, n*—statistical hypothesis test used to compare the means of two samples or a sample mean to a standard value.

3.2.10.1 *Discussion*—It is appropriate when the measure of interest is normally distributed in small samples and, more generally, for continuous, unbounded, symmetric measurements when the sample size is larger. Assumptions include no ties in the data.

3.2.11 *Type I error, n—see* alpha risk.

3.2.12 *Type II error, n—see* beta risk.

3.2.13 *Wilcoxon-Mann-Whitney test, WMW, n*—rank-based independent sampling alternative to the student's *t*-test that is appropriate when the data are measured on a common continuous scale that is not normally distributed.

3.2.13.1 *Discussion*—In these situations, it can be more efficient (increased statistical power to find a difference at a given sample size) than a student's *t*-test. Like the students *t*-test, it requires the assumption that the data have no ties.

4. Summary of Guide

4.1 This guide covers the similarities and differences between acceptance and preference measures when used alone and together in a two-sample test (see Fig. 1). The two measures provide different information about respondents'

[2] For referenced ASTM standards, visit the ASTM website, www.astm.org, or contact ASTM Customer Service at service@astm.org. For *Annual Book of ASTM Standards* volume information, refer to the standard's Document Summary page on the ASTM website.

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

subjective responses to products and should be deployed to meet different research or business objectives. Acceptance measures are recommended when there is a need to obtain information on intensity of liking/disliking and determine the relative hedonic status of two products. Preference measures are recommended when there is a need to obtain information on choice behavior or determine an ordinal relationship between two products. Correct sampling of respondents is critical in both types of test. The researcher shall carefully prepare the research learning plan and thoroughly review the pros and cons of the specific research design chosen (that is, measuring acceptance, measuring preference, measuring both) against the decision risks associated with each measurement. Acceptance and preference measures, while imperfect, continue to be extremely useful in managing the risk in developing and delivering new products to the marketplace.

5. Significance and Use

5.1 Acceptance and preference are the key measurements taken in consumer product testing as either a new product idea is developed into testable prototypes or existing products are evaluated for potential improvements, cost reductions, or other business reasons. Developing products that are preferred overall, or liked as well as, or better, on average, compared to a standard or a competitor, among a defined target consumer group, is usually the main goal of the product development process. Thus, it is necessary to test the consumer acceptability or the preference of a product or prototype compared to other prototypes or potential products, a standard product, or other products in the market. The researcher, with input from her/his stakeholders, has the responsibility to choose appropriate comparison products and scaling or test methods to evaluate them. In the case of a new-to-the-world product, there may or may not be a relevant product for comparison. In this case, a benchmark score or rating may be used to determine acceptability. A product or prototype that is acceptable to the target consumer is one that meets a minimum criterion for liking, and a product that is preferred over an existing product has the potential to be chosen more often than the less-preferred product by the consumer in the marketplace, when all other factors are equal.

5.2 The external validity (the extent to which the results of a study can be generalized) of both acceptance and preference measures to manage decision risk at all stages of the development cycle is dependent on the ability of the researcher to generalize the results from the respondent sample to the target population at large. This depends both upon the sample of respondents and the way the test is constructed. Within the context of a single test, acceptance measures tell the relative hedonic status of the two samples, quantitatively, as well as where on the hedonic continuum each of the samples falls, that is, "disliked," "neutral," or "liked." In contrast, preference measures tell the relative choice status of two samples within a specific respondent group. Results from these measures can and will vary from test to test depending on the number and type of respondents serving in each test, the size and nature of the sensory differences between the two samples, the method of executing the test, and any error present in the test. The identification, control, measurement, and tracking of variables that may influence results across tests (for example, production location, sample age, and storage conditions) are the responsibility of the researcher.

5.3 While measures of acceptance and preference are both subjective responses to products, and can be somewhat related, they provide different information. A product may be "acceptable" but still not be preferred by the consumer over other alternatives, and conversely, a product may be preferred over another but still not be acceptable to the consumer. These two terms, therefore, should not be used interchangeably. When a bipolar hedonic scale with multipoint options is used, the researcher should specifically refer to "liking," "acceptance," or "hedonic ratings." When preference measures are used, the researcher should refer to, "preference," "product selection," or "choice." Research professionals themselves should be precise in their usage of the terms "acceptance" and "liking," to refer only to scaling of liking. These researchers should use the terms "preference" and "choice" to refer to two ("Prefer A" or "Prefer B") or three-choice ("Prefer A" or "Prefer B" or "No Preference") response options given in a preference test. In addition to having different meanings, the two measures also do not always provide similar results. This guide will cover the similarities and differences in information each provides, some guidelines around implementation, and interpretation of findings. This guide will thus give users an understanding of the issues at hand when planning, designing, implementing, and interpreting results from acceptance and preference tests with consumers.

5.4 While both measures are commonly used to provide information for product development decisions and evaluating a product's competitive status, it is important to remember that pricing, positioning, competitive options, product availability, and other marketplace factors also impact a product's success.

6. Hedonic Testing—Steps in Planning and Conducting an Acceptance or Preference Test

6.1 *Decide on the Key Question to be Answered: Liking or Choice or Both*—Before planning and implementing a test, the researcher should determine what is needed to be learned from the research and what decisions will be made based on the outcome. The researcher would be wise to consider overall business strategies and the wider context of the project before test implementation. Additional considerations include stakeholder alignment, resource availability, and the actionability of potential outcomes. The researcher translates the stakeholder's desired learning into a testable hypothesis, defines the test object and decision criteria, and confirms the objective and criteria with stakeholders before collecting data. Both types of tests may be done at all project stages—to get a product's baseline measure early in development, to gauge progress later in development, or when a product is already in the market.

6.2 *Set Decision Criteria: Action Standards, Hypothesis Direction, Sample Size, and Risk Levels:*

6.2.1 *Action Standards*—The action standard determines whether the product meets the success criterion set in advance for success. In the case of acceptance testing, the action

standard is set based on the product of interest's hedonic score relative to that of the second product. In the case of preference testing, the action standard is set based on the product of interest's preference score relative to that of the second product. The type and direction of the primary question, on which the action standard is based, factor heavily into the setting of the action standard.

6.2.2 *Determine Type and Direction of Question*—In general, there are two classes of questions associated with these types of evaluations: difference (directional or nondirectional) and parity questions.

6.2.2.1 *Directional—One-sided Hypothesis Testing*—The test hypothesis is often that a new version of a product will be better liked or preferred compared to the current product or that a given brand of product will be better liked or preferred compared to another brand. These are examples of one-sided tests. Note that if the goal is not achieved (the new product is not better liked or preferred compared to the current product), it cannot be determined whether the new product is at parity or less liked or preferred compared to the current product. One-sided tests require fewer respondents and, thus, can be the most cost-effective approach to evaluating the hedonic status of two products in an acceptance or preference test when the goal is to outperform another product. However, if the goal is not achieved, the relative status of one product versus another cannot be determined.

6.2.2.2 *Nondirectional—Two-sided Hypothesis Testing*—The classical two-sided test is most appropriate when the business or researcher wishes to know "which product is liked better?" or "which product is preferred?" when, for example, it is possible a new product may be either less liked or preferred or more liked or preferred than a comparison product. The advantage of this type of test is that it allows for a finding on either side of parity. However, two-sided tests require a larger sample size to achieve the same power as a one-sided test.

6.2.2.3 *Parity*—Hedonic parity, "equivalence in liking or preference," "just as good as in (liking or preference)," are studies in which the objective is to demonstrate that the two products' hedonic status is the same. Hedonic parity does not include superiority. "Unsurpassed" tests are those in which the goal is to establish that the product of interest is not less liked or less preferred than a comparison product. The "unsurpassed" test objective is to obtain support that the test product is comparable, or, possibly even higher, in liking or preference versus another product. Parity or unsurpassed test results may be used to support communications to the consumer. Regardless of the end use of the data generated in a hedonic test or parity, the researcher will need to sample substantially more respondents than is needed in tests for difference. Estimated respondent sample sizes to yield results sufficiently robust to support parity in liking or preference are between 200 to 500, depending on the size of the differences between the two products, the standard against which the test result is measured, and the variance associated with the liking scores. See Test Method E2263 and Guide E1958 for more detailed information on sample size requirements in preference tests when support for parity is the test objective.

6.2.3 *Review Previous Testing Results and Evaluate Risk Levels Appropriate to Project's Objectives and Decision Risks*—The researcher evaluates risk by gathering information about the status of the project that includes this particular research, estimating the resource risk around the results, and the impact of a false positive ("α-risk") or a false negative ("β-risk") test result. Alpha risk is the risk that arises from falsely declaring two products to be different when they are truly at parity, while beta risk arises from falsely declaring two products to be at parity when they are truly different. As an example, finding a difference when products are actually at parity could impact the business by leading it to launch a product it believes has a competitive edge when in fact no competitive advantage exists. Similarly, failing to detect a difference when products are, in fact, different could lead the company to spend unnecessary development time and resources to improve a product further, when, in fact, it already is liked more than a standard. Further, using lack of significance in a preference test as the rationale for stating that parity in preference exists is not correct and can lead, in the short term, to launching an inferior product.

6.2.4 *Set Sample Sizes Based on Direction and Risk Levels*—For both acceptance and preference tests, a sufficient sample size shall be used to ensure enough test power. Practically, the researcher will need to strike a balance between test power and the number of respondents one can afford to employ. Commercial software for such calculations includes, but is not limited to, SAS, SPSS, JMP, Stata and Minitab. Free calculations are available at http://statpages.org/ or http://www.stat.uiowa.edu/~rlenth/Power/index.html. Sample sizes for preference tests at different risk levels can be found in Test Method E2263.

6.3 *Plan Data Analysis*—It is critical to determine how the data will be analyzed before data collection as the method of analysis will impact power and variability calculations needed to determine sample size. It is best to outline the decision criteria as they relate to the specific measures used in the test in advance and gain the alignment amount stakeholders. Following this, the researcher should outline the possible outcomes of the test before the data are collected, as unexpected results will be challenged on many different levels: "Was the test executed properly?" "Was the right method/measure used?," and so forth.

6.4 *Define Respondent Sample*—For both acceptance and preference studies, it is important that the results from the samples respondents reflect the target market, current category, or brand users for the product. For both acceptance and preference testing, respondents should include those most relevant to the question under study: specific brand users, product category users, or targeted non-category users. This recommendation is particularly true when the research question is hedonic in nature. When the research question is functional, or performance related, it may be appropriate to use employees or non-target consumers to screen products for attributes such as "easy to open," "dispenses uniformly," "covers completely," and so forth.

TABLE 1 Types of Respondents

Respondent Sample Type	Recommended?	Rationale
Target users—Currently using the product, flavor/form users, would purchase/use again	Yes	Differences in hedonic responses among a sample of such respondents are most likely to reflect that of the population of target users, assuming that the sampling plan includes a sufficient number of respondents and the appropriate selection criteria have been applied.
"Convenience" sample—Category users who are positive toward the concept, and so forth, and positive to the flavor in the case of a food product	"Qualified yet," with associated risks	Liking or preference response likely to mirror that of the target consumer up to a point: if product differences are small, or there is sensory segmentation in the target group, hedonic responses might mislead the researcher.
"Convenience" sample—External respondents, not current users, not users of the category, or even rejectors of the category.	No	Liking or preference response to the two products may not mirror that of the target consumer.
Non-R&D and project team, for example, marketing, sales, and plant personnel	No	Bias toward own product.
Research and Development personnel, not on project team	No	Knowledgeable about project objectives, technical knowledge about product, bias toward own project.
Project team/stakeholders	No	Knowledgeable about project objectives, technical knowledge about product, bias toward own product.
Trained or experienced panelists used in discrimination or descriptive tests	No	Testing and training experiences lead this group of respondents to evaluate the products objectively rather than the subjective evaluations required in hedonic tests.

6.4.1 Target user selection criteria may be based on a number of criteria: demographics, geography, psychographics, proprietary segmentation information, or product usage behavior, or combinations thereof. For existing products or line extensions, a sample of current users of the product or brand is recommended to assess a product's suitability for the brand. Additionally, if the product is intended to attract competitive users or new users, then respondent samples from the group(s) is/are needed, since the study results can vary significantly across different subgroups of brand users within the category. Based on the degree of consumer segmentation within a category or the presence of a small number of competitors, the selection of respondents can greatly influence the study results, particularly for preference studies conducted with in-market products. It is generally accepted that loyal or heavy users of a product may recognize their product, even in an unbranded product test, and are biased toward rating it more favorably than the other product within the study. After the acceptance or preference measure is completed, the researcher can ask respondents to postulate the brand identity of the products. Clear documentation of respondent selection criterion is required so that this information is available for any subsequent related consumer studies.

6.4.2 *External Respondents: Minimum Respondent Requirement for Acceptance and Preference Testing*—It is highly recommended that respondents be recruited and selected from a population of target users for the products being tested. By doing so, the researcher should be able to generalize findings. While some debate exists as to the suitability of using employees to obtain products' hedonic information as a best practice, use of employees as respondents for either acceptance or preference testing is strongly discouraged as there may not be a meaningful relationship between employees' and external target users' responses to the tested products. "Convenience" samples (typically small samples of respondents drawn from one source, such as a church or a university that may not be users of the products, category acceptors, or even familiar with the product category) are recommended with reservations, only if they are concept positive and flavor positive if a food product is to be tested. These reservations are based on the common convenience sampling practice of obtaining small number of consumers (for example, less than 100) when using a local area source, coupled with the possibility that drawing respondents from a single area might not include consumers representing different sensory segments. Results from respondents drawn from a convenience sampling method may not represent consumers who are actual users. See Table 1, which outlines recommendations for obtaining different consumer samples.

6.4.3 Trained descriptive, discrimination panelists or frequently used internal panelists drawn from the technical areas of a company should not be used as respondents in an acceptance or preference test. Because of their training and analytical orientation and their knowledge of the product's technical features, these panelists are likely to respond to products different from untrained consumers. See Table 1, which lists the various types of respondent samples that might be considered for an acceptance or a preference test, recommended usage, and rationale.

6.4.4 For new product categories, it may be difficult to identify the criteria for selecting the target consumer. For new products, the researcher may want to select category acceptors who are also early adopters, consumers who actively seek and purchase new products in the category, or those that are positive to the idea or concept of the new product (concept acceptors).

6.5 *Record Product Information*—The researcher needs to record the product information on the package. Most researchers take a picture of the product or remove the label and photograph to the front label information, ingredients, and nutritional facts. The lot number and "use by" dates also need to be recorded. If the product is not on the market, then the formula or composition and information needed for retrieval of

the ingredients, processing, and manufacturing location should be recorded. Preparation or other usage instructions and carriers used should also be documented. These records will allow future researchers to compare results from the same product if needed.

6.6 *Develop Questionnaire*—Diagnostic information (intensity, just-about-right (JAR), "Check All That Apply" (CATA)), open-ended likes and dislikes, or other measures that help explain product performance may be included in both acceptance and preference tests. The recommended practice is to ask the overall liking or preference question first, before diagnostic questions, if the hedonic question is going to be used for decision-making. If a preference question is to be included, the option of including a "no preference" response shall be considered (see 8.5).

6.7 *Collect Data*—Present the proper set of products in a manner that ensures unbiased responses. Checks and balances need to be implemented to ensure that data collected provide actionable results. For unbranded testing, sensory information that allows a product's brand to be identified should be eliminated or reduced as much as possible. Likewise, the ages, the condition, and the handling of the samples being tested should be comparable. The method of sample presentation should be balanced to reduce order and context effects. See Practice E1871, Guide E1958, Test Method E2263, and *ASTM Manual 26* (**4**) for more complete descriptions of methods to manage or eliminate bias in sensory tests. Samples are typically served in sequential monadic fashion when conducting acceptance testing, while sequential monadic or simultaneous presentation are both common modes of sample presentation in preference testing. While a somewhat less sensitive determination of the relative hedonic status of two products may also be obtained via monadic testing (different respondent groups evaluate each of two samples), this guide has, as its focus, the more common sequential monadic presentation.

6.8 *Analyze Data and Interpret Results—Determine Whether Action Standard Has Been Met):*

6.8.1 *Data Analysis Information for Both Acceptance and Preference Measures*—The research plan for the specific analysis when both acceptance and preference are measured should specify in advance the alpha level, beta level, and direction (one-sided or two-sided) of the statistical tests. For preference tests, the plan should also include information on the number of common response (P_{max}) and the size of difference to be detected. For acceptance tests, the size of difference to be detected and the estimated variability in liking of both products should also be included. The results are compared with the decision criteria for interpretation.

6.9 *Report and Communicate the Results—Derive a Message About Product's Relative Hedonic or Preference Status*—Once the mechanics of the test are complete and data are collected, analyzed, and reviewed, the researcher has the job of communicating what the results mean: which product is liked better; which product was selected more often over the other; and the evidence, if any, for consumer segments; limitations of generalizing to other respondent groups; and how the results compare to previous findings. Caution, however, should be taken in comparing results to prior findings as consumer response is often context dependent. For example, other products included in the research may influence ratings for the products of interest. Recommendations as to next steps, based on test findings as related to business strategy, should be included.

7. Acceptance Testing

7.1 *Definition of Acceptance Testing: Affective Continuum*—The nine-point hedonic scale is a bipolar scale with the same format as Likert scales. Three broad categories are represented: "like," "neutral," and "dislike." This type of hedonic scale is used when the primary goal of the research is to learn where two products fall on this hedonic continuum and the size of the hedonic differences between them. The nine-point hedonic scale provides degree and direction from the neutral point "neither like/nor dislike." The original nine-point hedonic scale was constructed empirically and, while the verbal anchors have been shown to have equal interval properties for the original stimuli (**5**), some researchers do not accept the equality of the categories (**6**).

7.2 *Set Decision Criteria: Action Standards, Hypothesis Direction, Sample Size, and Risk Levels:*

7.2.1 Use acceptance measures when there is a need to identify the two products' relative status on the hedonic continuum, that is, where on the scale each is rated, that is, whether consumers "like," "dislike," or are "neutral" toward each one of two products and when the interval relationship between the two samples needs to be quantified.

7.2.2 The hypothesis to be tested will state either that there is some difference in liking between the samples or that there is no difference in liking between the samples. The action standard will be based on whether the obtained results are consistent with the hypothesis at a prespecified probability level. It is typical to test at the 90 or 95 % confidence level.

7.2.3 The number of consumers to be included in the research will depend on several factors: (*1*) the consumer sample size used historically in the company, (*2*) the minimum size of the sensory difference in liking (in scale units) desired to be detected between the two products, and (*3*) the variability in liking ratings among the respondents. If consumer-liking data exist from previous testing of the same products, this historical data can be used to estimate the variability that is likely to be found in a consumer test of the same products (standard deviation/standard error). For many U.S. consumer products companies, sample sizes between 100 and 150 are common when the test hypothesis is to establish differences in liking. In acceptance tests, it is possible to gauge in advance the risk of missing a true difference in liking between two samples (beta) if one knows the size of the difference one wishes to detect (if, for example, one wishes to be able to detect a difference of 0.3 hedonic units on a 9-point hedonic scale) and knows the variance in liking ratings for the samples before conducting the test. As an example, 130 people are required to have an 80 % chance of detecting a 0.5 difference with 95 % confidence when using a 9-point hedonic scale with a standard deviation of 1 unit in a 2-sample test. For acceptance tests with the 9-point hedonic scale, a sample size of 112 respondents is

needed to detect a 10 % difference in the scale given the variability in the data in this meta-analysis (**7**).

7.3 *Plan Data Analysis*—Data analysis for a two-sample acceptance test is typically a dependent (related) samples *t*-test. For a finding of one product being liked more or less than another, the researcher only needs to set the confidence level in advance. See 6.2.2.3 for a discussion of parity.

7.4 *Define Respondent Sample*—See 6.4.

7.5 *Develop Questionnaire:*

7.5.1 *General Considerations*—The questionnaire for an acceptance test will consist of one or more liking scales for overall and possibly attribute ratings of acceptance, and could also include diagnostic scales such as intensity or just about right. Scale format options vary widely.

7.5.2 *Scale Format Options*—The nine-point hedonic category scale may be presented in either a horizontal or vertical layout, with categories labeled as follows; "9" Like Extremely, "8" Like Very Much, "7" Like Moderately, "6" Like Slightly, "5" Neither Like Nor Dislike, "4" Dislike Slightly, "3" Dislike Moderately, "2" Dislike Very Much, and "1" Dislike Extremely. The scaling numbers may or may not be included with the scale anchors. Other options include the hedonic scale as a line scale (usually 15 cm), labeled affective magnitude scale (**7-9**) or ratio scale (**10**). Each of these options has relative advantages and disadvantages, which vary depending on the research objective and respondent sample. If results will be compared across tests, it is important to use the same scale consistently. Extrapolating results from one scale to another is not recommended as end-point effects and other psychological issues make this imprecise at best and grossly incorrect at worst. The Office of Scale Research at Southern Illinois University can assist researchers with scale identification and usage. See http://scaleresearch.siu.edu/.

7.5.3 *Number of Scale Points*—An odd number of categories, or scale points, with a "neutral" midpoint and a balanced number of categories on either side of the midpoint are typical of hedonic rating scales. Unbalanced scales will not fairly represent the range of hedonic responses consumers might have. More scale points provide the advantage of increased sensitivity in finding liking differences between two products. End-point avoidance means that an *N*-point scale is effectively an *N* minus two-point scale to the extent that respondents avoid using the end points. For example, a nine-point scale is often effectively seven-point scale, and a five-point scale is often effectively a three-point scale (**11**).

7.5.4 *Inclusion of Diagnostic Scales*—Although the liking rating is the primary response with acceptance scales, further diagnostic questions may be included in the questionnaire. Researchers frequently ask consumers to either (*1*) rate the intensity or liking of the product on specific attributes, or (*2*) indicate the extent to which the product is "Just About Right" (JAR) on specific attributes, or both, (that is, opportunity analysis). Both intensity attributes and JAR ratings are diagnostic. They are intended to provide the researcher with information to interpret the liking status and provide guidance as to how to improve it. JAR data are used to explain why products are liked or how the product can be improved or both. Note that the response to these questions may be biased by a halo effect as the respondent may be justifying their prior choices/ratings. For more information on JAR scales see *ASTM Manual 63* (**12**).

7.5.5 *Ask Acceptance before Diagnostic Questions*—The first question asked is generally thought to be the most unbiased. Placing the acceptance question first is recommended if that is the primary measure of interest. Placing the acceptance question after the attribute questions may change (usually lower) the mean overall liking ratings. The diagnostic questions should use consumer language and refer to attributes that consumers would typically notice. For example, asking about "glue lines" (in a cardboard package) in a consumer product is too technical, while asking how difficult it was to open the package is not. It is hypothesized that focusing on specific attributes before the overall acceptance question may prompt consumers to pay closer attention to certain product characteristics that they might otherwise ignore and, therefore, cause them to be more critical when answering later questions. In monadic sequential designs, the second-sample acceptance result may be influence by diagnostic questions asked in the first sample (**13**). This is one reason that the order of product evaluation is carefully balanced across samples.

7.6 *Collect Data*—See 6.7.

7.7 *Analyze Data and Interpret Results—Determine Whether Action Standard Has Been Met*—Once data have been collected and checked for correctness, the statistical analysis of the data may be done using the actual variability measures. Parametric analyses, such as a dependent *t*-test in a two-product, one-respondent group test, are typically done with acceptance data, although nonparametric alternatives such as sign or signed rank tests on the differences should be considered when the data fail the parametric assumptions. After the data have been collected, they should be reviewed to determine if the variability and distribution assumptions used in planning the test were met. If not, a prespecified action standard may not have the desired risk levels. Since the business goal, the analysis, and the desired risk levels determine the action standard, it may be necessary to adjust these to attain the desired properties. If the true variation in liking is not known, either the action standard or the desired risk levels can be set before the test is conducted, not both. This is because a measured variation in liking that is larger than that assumed pre-testing will result in either greater risk levels associated with a given action standard or a more stringent action standard to maintain the prespecified risk levels.

7.7.1 *Plot Data, Review Variability, and Measures of Central Tendency*—It is critical that the researcher examine not just the mean score or the summary liking or preference data from a test but also the distribution of responses and the relationship of these responses to characteristics of the panel sample, for example, segmentation. It is also good practice to determine how well the data meet the requirements of any statistical tests that will be performed. As an example, examine the skewness, kurtosis, and normality of the distributions for each of the products. If the acceptance ratings are bimodal for both products, the researcher can do a cluster analysis to determine what the mean liking is for each product for each cluster and to