Chapter 13

Preference Testing

Abstract Preference testing refers to consumer tests in which the consumer is given a choice and asked to indicate their most liked product, usually from a pair. Although these tests appear straightforward and simple, several complications are encountered in the methods, notably how to treat replicated data and how to analyze data that include a "no-preference" option as a response. Additional methods are discussed including ranking more than two products, choosing both the best and worst from a group, and rating the degree of preference.

The number of judges that are involved in a study may be such that rather unimportant differences may receive undue attention. It is quite possible to produce statistically significant differences in preference for product which have little practical value by simple increasing the number o judges that are utilized.

—H. G. Schutz (1971).

Contents

13.1 Introduction—Consumer Sensory Evaluation

Consumer sensory evaluation is usually performed toward the end of the product development or reformulation cycle. At this time the alternative product prototypes have usually been narrowed down to a manageable subset through the use of analytical sensory tests. Frequently, the sensory testing is followed by additional testing done through market research. The big difference between consumer sensory and marketing research testing is that the sensory test is generally conducted with coded, not branded, products while market research is most frequently done with branded products (van Trijp and Schifferstein, 1995). Also, in consumer sensory analysis the investigator is interested

H.T. Lawless, H. Heymann, *Sensory Evaluation of Food*, Food Science Text Series,
DOI 10.1007/978-1-4419-6488-5_13, © Springer Science+Business Media, LLC 2010

in whether the consumers like the product, prefer it over another product, or find the product acceptable based on its sensory characteristics. The consumer sensory specialist often has no interest in purchase intent, effect of branding, and/or cost factors. Thus, a product will not necessarily be financially successful just because it had high hedonic scores (was well liked) or because it was preferred over another product. Success in the marketplace is also affected by price, market image, packaging, niche, etc. However, a product that does not score well in a consumer acceptance test will probably fail despite great marketing.

Sensory techniques are widely used to assess the reactions of the public to a variety of stimuli, even environmental annoyances (Berglund et al., 1975). Historically, acceptance testing of foods with consumers represented an important departure from earlier methods based on the opinions of expert tasters or the assignment of quality scores by panels looking for product defects (Caul, 1957; Jellinek, 1964; Sidel et al., 1981). The growth of acceptance testing helped to foster the logical separation of the analytical sensory techniques from affective tests, a distinction that was lacking in the earlier traditions of expert grading and quality testing. Acceptability information is extremely useful. For example it can be combined with other sensory analyses, knowledge of consumer expectations, and product formulation constraints in determining the optimal design of food products (Bech et al., 1994; Moskowitz, 1983).

In foods and consumer products, there are two main approaches to consumer sensory testing, the measurement of preference and the measurement of acceptance (Jellinek, 1964). In preference measurement the consumer panelist has a choice. One product is to be chosen over one or more other products. In the measurement of acceptance or liking the consumer panelists rate their liking for the product on a scale. Acceptance measurements can be done on single products and do not require a comparison to another product. An efficient procedure is to determine consumers' acceptance scores in a multi-product test and then to determine their preferences indirectly from the scores. Both of these types of testing are called hedonic or affective tests. The term hedonic refers to pleasure. The goal of both types of tests is to assess the appeal of a product to a consumer on a sensory basis, i.e., to get the consumer's reaction on the basis of appearance, aroma, taste, flavor, mouthfeel, and texture. For non-food products, other sensory factors may come into play as well as the perceived efficacy or performance of a product. Other factors related to the product appeal are discussed in the next chapters such as the appropriateness of a product for its intended use and the consumer's satisfaction with a product (performance relative to expectations). This chapter will deal with simple preference choice. Historically, the term preference test has also been used to refer to surveys of people's likes and dislikes based on presenting lists of food names (Cardello and Schutz, 2006). Such data are generally scaled and thus could be called an acceptance test as the term is used in the next chapter. In this chapter we will use the term preference only to refer to experiments in which a choice is made between two or more alternatives.

The key to a successful consumer test is finding the right participants. Persons in the test must be representative of the population to which the results will be generalized. They should be users of the product and probably frequent users. The sensory specialist will often negotiate with the client (the person requesting the test and who will use the results and conclusions) as to just how frequently the person must use the product to qualify for the test. Obviously, no trained panelists are used for such a test as they approach the product in a different frame of mind from the average consumer. Sometimes employees may be used in the early stages of testing, but they must be users of the product category being tested. Further information on qualifying and screening consumer test participants is found in Chapter 15, but this principle applies equally to the next three chapters on preference, acceptability scaling, and consumer field tests.

The sensory test is to some degree an artificial situation and may not always be a good predictor of consumer behavior in the real world regarding purchase and/or consumption (Lucas and Bellisle, 1987; Sidel et al., 1972; Tuorila et al., 1994; Vickers and Mullan, 1997). However, purchase decisions and even consumption involve many other influences than the simple sensory appeal of a product. In spite of this shortcoming, preference and acceptance tests can provide important information about the relative sensory appeal of a group of products that need blind-labeled consumer testing at some phase in the product development scenario. That is, the objective is to find the product among a group that has to be best potential for success, on a sensory basis. So in spite of the

limitations of the methods imposed by the artificial context of the testing situation, they are still quite valuable to the sensory specialist and the clients who request the product test.

The following three chapters will discuss preference testing, acceptability testing, and then consumer field tests and questionnaire design, respectively. This chapter will focus on the simple paired preference test. In spite of its appeal and its apparent simplicity, sensory and market researchers have added additional variations that complicate the test and analysis. The two main variations involve replication and the offering of a "no-preference" option (or both). We will recommend as a good practice the simplest version of a paired preference test, but the other more complicated versions and ranking are also discussed in this chapter. Some worked examples are provided in the chapter itself and in the appendices that follow.

13.2 Preference Tests: Overview

13.2.1 The Basic Comparison

Preference tests are choices involving comparisons between two products or among several products. If there are two products, this is known as a paired preference test. It is the simplest (and most popular) type of test that looks at the appeal of products to consumers. Paired preference tests are some of the oldest sensory tests. A publication in 1952 described a mail panel maintained (for the preceding 20 years!) by the Kroger Company food retailers that would receive pairs of products for comparison in the mail along with a questionnaire (Garnatz, 1952). Paired tests are popular in part because of their simplicity, because they mimic what consumers do when purchasing (choosing among alternatives), and because some people believe they are more sensitive than scaled acceptance. We have seen no hard data to substantiate this latter belief although intuitively it is possible that two products might receive the same score on a scale, but one might be slightly preferred to the other. However, it is also possible that a product could win in a choice test, but still be unappealing on its own (like an election where you do not like either candidate but vote based on the lesser of two evils). This is one shortcoming of a preference test that it gives you no absolute information on the overall appeal of a product. Acceptance testing with a scale is designed to do just that.

13.2.2 Variations

Variations on preference testing involve choosing a preferred product from multiple alternatives. One version of this is ranking, in which products are ordered from best liked to least liked. Another version gives products in small groups (usually three) and the consumer is asked which one is liked best and which one is worst. This is known as best–worst scaling, because the resulting data can be placed on a scale, even though the task itself involves a choice and not a response on a scale. Best–worst and paired tests are both special cases of ranking, i.e., you can think of a paired test as a ranking of only two products. Both ranking and best–worst scaling are discussed in this chapter in later sections.

Other important variations on the preference test involve the use of a "no-preference option" and the replication of the test on the same persons. The no-preference option provides more information, but complicates the statistical analysis. It is generally avoided by product researchers although in advertising claim substantiation, it may be required for legal reasons (ASTM, 2008). Replication is not common in preference tests. However, recent research has shown that replication will enhance the consumer's discrimination among products in an acceptance test. Also, replication can provide evidence as to whether there are stable segments of consumers who prefer different versions of a product. The primary goal of a preference test is to find a "winner," i.e., that product which has significantly higher appeal to consumers than other versions in the test.

There have been studies on the efficacy of the paired preference test with illiterate and semi-literate consumers (Coetzee, 1996; Coetzee and Taylor, 1996). These consumers, many of whom could not read or write, could reliably perform paired preference tests given verbal instructions and using paired symbols (the same symbol but one outlined and the other solid). When one of the authors tested the methods in a different country the method worked well with illiterate consumers (Coetzee, 1996). Paired preference tests using color codes have also been successfully used

(Beckman et al., 1984) and could be used with illiterate or semi-literate consumers. Preference tests are also suitable for young children as the task is straightforward and easily understood (Engen, 1974; Kimmel et al., 1994; Schmidt and Beauchamp, 1988; Schraidt, 1991).

Multiple paired preference tests form the basis for a new kind of threshold test, the consumer rejection threshold. Prescott et al. (2005) gave groups of wine consumers increasing levels of trichloroanisole in wine to try and find the level at which there was a consistent preference for wine without this "cork taint." This technique, discussed in Chapter 6, should find wide application in commodities in which the chemistry and origins of various taints and off-flavors are well understood (see Saliba et al., 2009, for another example).

13.2.3 Some Cautions

A common methodological problem comes from the temptation to add a preference test at the end of some other kind of sensory test. This should be avoided. It is very unwise to ask for preference choices after a difference test, for example. This is not recommended for a number of reasons. First, the participants in the two tests are not chosen on the same basis. In preference tests, the participants are users of the product, while in discrimination tests, panelists are screened for acuity, oriented to test procedures, and may even undergo rudimentary training. The discrimination panel is not representative of a consumer sample and it is usually not intended to be so. The emphasis instead is on providing a sensitive tool that will function as a safety net in detecting differences. Second, participants are in an analytic frame of mind for discrimination while they are in an integrative frame of mind (looking at the product as a whole, usually) and are reacting hedonically in preference tasks. Third, there is no good solution to the question of what to do with preference judgments from correct versus incorrect panelists in the discrimination test. Even if data are used only from those who got the test correct, some of them are most likely guessing (correct by chance). Exclusion of panelists from a consumer test on any other basis than their product usage is a poor practice. They have been selected based on their being a representative sample of the target group.

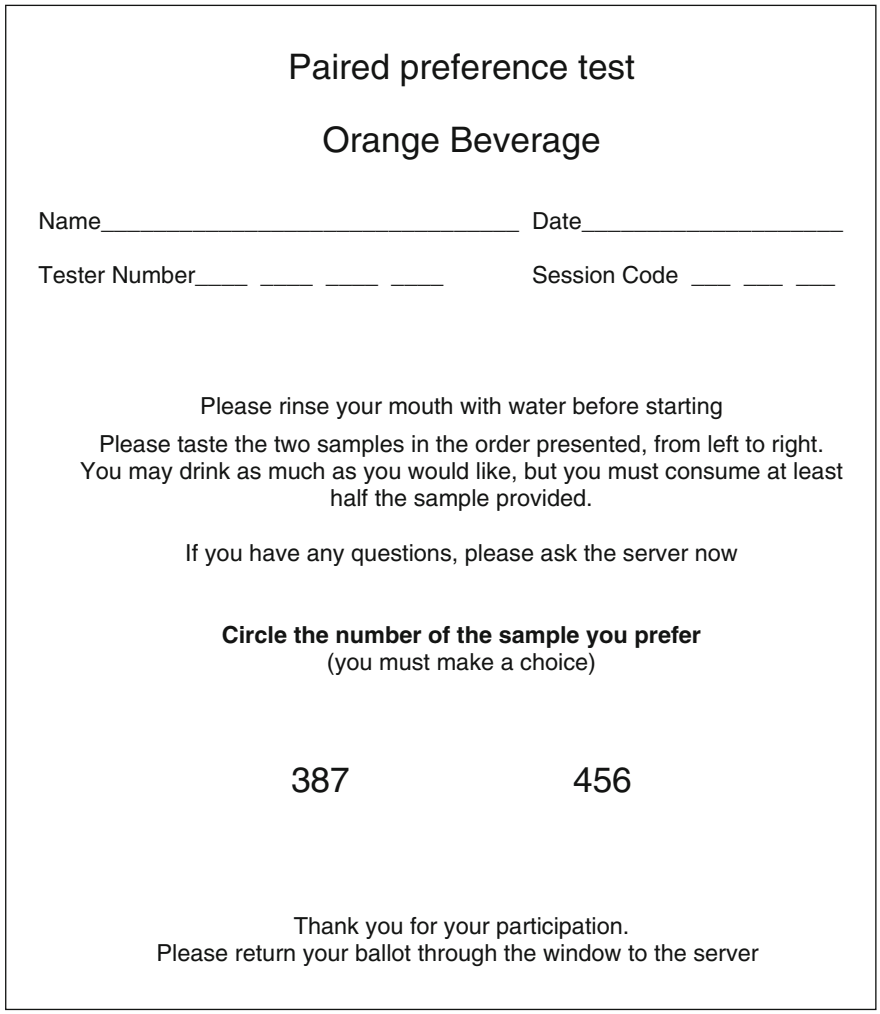
When doing a paired preference test keep in mind that the technique is designed to answer one and only one important question. Consumers are in an integrative frame of mind and are responding to the product as a whole. They are generally not analyzing it as to its individual attributes, although one or two salient characteristics may sometimes drive their decisions. However, these one or two salient characteristics may also cause other attributes to be viewed in a positive light, an example of what is called the "halo effect" (see Chapter 9). For these reasons, it is difficult to get consumers to accurately explain the basis for their choice. Although it is fairly common to ask for diagnostic information in a large and expensive multi-city consumer field test, one should recognize the "fuzziness" of this information that it can be difficult to interpret and that it may or may not be useful in any kind of important decision making.

Choice tests and rankings indicate the direction of preferences for the product but are not designed to find the relative differences in preference among the products. In other words, the results give no indication of the size of the preference. However, it is possible to derive Thurstonian scale values from proportions, giving some indication of the magnitude of the difference on a group basis. For example, Engen (1974) compared the hedonic range of odor preference among adults to the range of likes and dislikes for children using indirect scale values based on multiple paired comparisons. Adults showed a wider range of preferences on this basis.

13.3 Simple Paired Preference Testing

13.3.1 Recommended Procedure

In paired preference tests the participant receives two coded samples. The two samples are presented to the panelist simultaneously and the panelist is asked to identify the sample that is preferred. Often, to simplify the data analysis and interpretation, the subject must make a choice (forced choice) although it is possible to include a no-preference option (discussed later in this chapter). Figure 13.1 shows a sample score sheet without the no-preference option. The sensory specialist should make sure that the consumer panelist understands the task described by the score sheet.


Paired preference test

Orange Beverage

Name_____ Date_____

Tester Number____ ____ ____ ____ Session Code ___ ___ ___

Please rinse your mouth with water before starting

Please taste the two samples in the order presented, from left to right. You may drink as much as you would like, but you must consume at least half the sample provided.

If you have any questions, please ask the server now

Circle the number of the sample you prefer
(you must make a choice)

387 456

Thank you for your participation.
Please return your ballot through the window to the server

Fig. 13.1 Ballot example for a paired preference test when a choice is forced.

The paired preference test has two possible serving sequences (AB, BA). These sequences should be randomized across panelists with an equal number of panelists receiving either sample A or sample B first.

The steps in setting up and conducting a paired preference test are shown in Table 13.1. It is always appropriate to confirm the test objectives with the end user of the test results. Testing conditions should also be made clear in terms of the amount being served, temperature and other aspects of the physical setup. It is wise to write down these conditions and the procedures for serving as a standard operating procedure (SOP), so that the staff conducting the test has a clear understanding of what to do. Of course, the ballot has to be prepared, random codes assigned to products, and a counterbalancing scheme set up for the alternating positions of the two products. Consumers must be recruited and screened so that they are suitable for the test; usually frequent users of the product are appropriate. In this test, consumers are forced to make a choice. Responding with "no preference" or equally preferred is not an option.

13.3.2 Statistical Basis

For paired preference methods the probability of selection of a specific product is one chance in two. The null hypothesis states that in the long run (across all possible replications and samples of people) when the

Table 13.1 Steps in conducting a paired preference test

1. Obtain samples and confirm test purpose, details, timetable, and consumer qualifications (e.g., frequency of product usage) with client.
2. Decide testing conditions (sample size, volume, temperature, etc.).
3. Write instructions to the panelists and construct ballot.
4. Recruit potential consumers.
5. Screen for product usage to qualify.
6. Set up counterbalanced orders (AB, BA).
7. Assign random three digit codes and label sample cups/plates.
8. Conduct test.
9. Analyze results.
10. Communicate results to client or end-user.

underlying population does not have a preference for one product over the other, consumers will pick each product an equal number of times. Thus the probability of the null hypothesis is $P_{pop} = 0.5$. Remember that P_{pop}, the proportion that we are making an inference about, refers to the proportion we would see prefer one product over another in the underlying population. It is not the proportion preferring that sample in our data. Mathematically, this may be written as Ho: $p(A) = p(B) = 1/2$. The test is two tailed since prior to the study the experimenter does not know which sample will be preferred by the consumer population. There is no right answer; it is a matter of opinion. The alternative hypothesis for the paired preference test is written as Ha: $p(A) \neq p(B)$. Three data analyses can be used based on the binomial, chi-square, or normal distributions, respectively.

The binomial distribution allows the sensory specialist to determine whether the result of the study was due to chance alone or whether the panelists actually had a preference for one sample over the other. The following formula allows the exact probability of one specific outcome (but not the overall probability for the hypothesis test). This equation gives the probability of finding y judgments out of N possible, with a chance probability of one-half:

$$p_y = (1/2)^N \frac{N!}{(N-y)!y!} \tag{13.1}$$

where

N = total number of judgments

y = total number of preference judgments for the sample that was most preferred

p_y = probability of making the number of preference choices for the most preferred sample

In this formula, $N!$ describes the mathematical factorial function which is calculated as $N \cdot (N-1) \cdot (N-2) \ldots 2 \cdot 1$. For example, the exact probability of five out of eight people preferring one sample over another is $(1/2)^8 \bullet (8!)/(3!)(5!)$ or 56/256 or 0.219. Bear in mind that this is the probability of just one outcome (one term in a binomial expansion, see Appendix B) and two other considerations need to be taken into account in testing for the significance of preferences. The first of these is that we consider the probability of detecting an outcome this extreme *or more extreme* in testing for significance, so the other terms farther out in the tail of the distribution must also be added to the probability value. So in our example, we would also have to calculate the probability of getting 6 out of 8, 7 out of 8, and 8 out of 8 and add them all together. You can see that this becomes very cumbersome for large consumer tests and so this approach is rarely done by hand, although there are good statistical tables that use exact binomial probabilities (Roessler et al., 1978). The second consideration is that the test is two tailed, so once you have added all the necessary terms, the total probability in the tail should be doubled. Remember we have no a priori prediction that the test will go one way or the other and so the test is two tailed. These considerations lead to the calculated values shown in Table 13.2. The table gives the minimal values for statistical significance as a function of the number of people tested. If the obtained number preferring one product (the larger of the two, i.e., the majority choice) is equal to or exceeds the tabled value, there is a significant preference.

13.3.3 Worked Example

In a paired preference test using 45 consumer panelists, 24 panelists preferred sample A. In Table 13.2 we find that the table value for 45 consumer panelists with an alpha criterion of 5% is 30. This value is larger than 24 and therefore the consumer panelists did not have a preference for one sample over the other. Let us assume that in a different study, 35 of 50 consumer panelists

Table 13.2 Minimum value (X) required for a significant preference

N	X	N	X	N	X
20	15	60	39	100	61
21	16	61	39	105	64
22	17	62	40	110	66
23	17	63	40	115	69
24	18	64	41	120	72
25	18	65	41	125	74
26	19	66	42	130	77
27	20	67	43	135	80
28	20	68	43	140	83
29	21	69	44	145	85
30	21	70	44	150	88
31	22	71	45	155	91
32	23	72	45	160	93
33	23	73	46	165	96
34	24	74	46	170	99
35	24	75	47	175	101
36	25	76	48	180	104
37	25	77	48	185	107
38	26	78	49	190	110
39	27	79	49	195	112
40	27	80	50	200	115
41	28	81	50	225	128
42	28	82	51	250	142
43	29	83	51	275	155
44	29	84	52	300	168
45	30	85	53	325	181
46	31	86	53	350	194
47	31	87	54	375	207
48	32	88	54	400	221
49	32	89	55	425	234
50	33	90	55	450	247
51	34	91	56	475	260
52	34	92	56	500	273
53	35	93	57	550	299
54	35	94	57	600	325
55	36	95	58	650	351
56	36	96	59	700	377
57	37	97	59	800	429
58	37	98	60	900	480
59	38	99	60	1000	532

Notes: N is the total number of consumers
X is the minimum required in the larger of the two segments
Choice is forced
Values of X were calculated in Excel from the z-score approximation to the binomial distribution
Values of N and X not shown can be calculated from $X = 0.98\sqrt{N} + N/2 + 0.5$
Calculated values of X must be rounded up to the nearest whole integer
Tests with $N < 20$ are not recommended but critical values can be found by reference to the exact binomial (cumulative) probabilities in Table I
Values are based on the two-tailed Z-score of 1.96 for $\alpha = 0.05$
Critical minimum values for $\alpha = 0.01$ can be found in Table M

preferred sample A over sample B. In Table 13.2, we find that the table value for 50 panelists at alpha of 5% is 33. The obtained value of 35 is greater than this minimum and therefore the consumers had a significant preference for sample A over sample B.

13.3.4 Useful Statistical Approximations

Most sensory specialists use these simple lookup tables for finding the significance of a test outcome. Remember that the test is two tailed and that you cannot use the tables for the paired difference test, which are one tailed. If you do not have the tables handy or wish to calculate the probability for some value not in the table, you can also use the Z-score formula for proportions shown in Chapter 4 and Appendix B for discrimination tests. The binomial distribution begins to give values very near the normal distribution for large sample sizes, and since most consumer tests are large ($N > 100$), this approach is nearly mathematically equivalent. The following formula can be used to calculate the z-value associated with the results of a specific paired preference test (Stone and Sidel, 1978). The formula is based on the test for a difference of two proportions (the observed proportion preferring versus the expected number or one-half the sample, as follows:

$$z = \frac{(p_{\text{obs}} - p) - 1/2N}{\sqrt{pq/N}} = \frac{(X - Np) - 0.5}{\sqrt{Npq}} \quad (13.2a)$$

Or

$$z = \left[(X - N/2) - 0.5\right] \Big/ 0.5\sqrt{N} \quad (13.2b)$$

where

X = number of preference judgments for the most preferred sample
$p_{\text{obs}} = X/N$
N = total number of judgments (usually the number of panelists)
p = probability of choosing the most preferred sample by chance
$q = 1–p$ and for paired preference tests: $p = q = 1/2$

The probability associated with the paired preference test is two-tailed so that a Z-value of 1.96 is

appropriate for a two-tailed test with alpha equal to 0.05. The obtained *Z*-score must be larger than 1.96 for the result to be statistical significant. For other alpha levels, the *Z*-table in the Appendix, Table A, may be consulted.

Another approach is to use the chi-square distribution which allows the sensory scientist to compare a set of observed frequencies with a matching set of expected (hypothesized) frequencies. The chi-square statistic can be calculated from the following formula (Amerine and Roessler, 1983), which includes the number –0.5 as a continuity correction. The continuity correction is needed because the χ^2 distribution is continuous and the observed frequencies from preference tests are integers. It is not possible for one-half of a person to have a preference and so the statistical approximation can be off by as much as 1/2, maximally. As in any chi-square test against expected values we have five steps: 1) subtract the observed value from the expected value and take the absolute value, 2) subtract the continuity correction (0.5), 3) square this value, 4) divide by the expected value, 5) sum all the values from step 4.

$$\chi^2 = \frac{[|(O_1 - E_1)| - 0.5]^2}{E_1} + \frac{[|(O_2 - E_2)| - 0.5]^2}{E_2} \tag{13.3}$$

O_1 = observed number choices for sample 1
O_2 = observed number choices for sample 2
E_1 = expected number choices for sample 1 (in a paired test = *N*/2)
E_2 = expected number of choices for sample 2 (*N*/2 again)

The critical value for χ^2 with one degree of freedom is 3.84. The obtained value must exceed 3.84 for the test to be significant. Note that this is the square of 1.96, our critical *z*-score. The binomial *z*-score and chi-square test are actually mathematically equivalent as long as both use or both do not use the continuity correction (see Proof, Appendix B, Section B.6).

13.3.5 The Special Case of Equivalence Testing

A parity or equivalence demonstration may be important for situations such as advertising claims. Establishing an equivalence or parity situation for a preference test usually requires a much larger sample size (*N*) than a simple paired preference test for superiority. The theoretical basis for a statistical test of equivalence is given in Ennis and Ennis (2009) and tables derived from this theory are given in Ennis (2008). The theory states that a probability value can be obtained from an exact binomial or from a normal approximation as follows:

$$p = \Phi\left(\frac{|x| - \theta}{\sigma}\right) - \Phi\left(\frac{-|x| + \theta}{\sigma}\right) \tag{13.4}$$

where phi (Φ) symbolizes the cumulative normal distribution area (converting a *z*-score back to a *p*-value), theta (θ) is the half-interval, in this case 0.05, and sigma (σ) is the estimated standard error of the proportion (square root of *pq/N*). *x* in this case is the difference between the observed proportion and the null (subtract 0.5). Here is an example, modified from Ennis and Ennis (2009).

In a paired preference test of a soft drink, 295 of 600 consumers prefer one product and 305 choose the other. Using a boundary of 0.50±0.05 as the "equivalence" region (i.e., a true population proportion that lies between 45 and 55%), we can perform a simple test from Eq. (13.4).

First, we continuity correct the proportion of 295 to 294.5, giving a proportion of 0.4908. Subtracting 0.5 gives our *x* value of –0.0092. Then our standard error is estimated by

$$\sigma = \sqrt{\frac{0.45(0.55)}{600}} = 0.02031$$

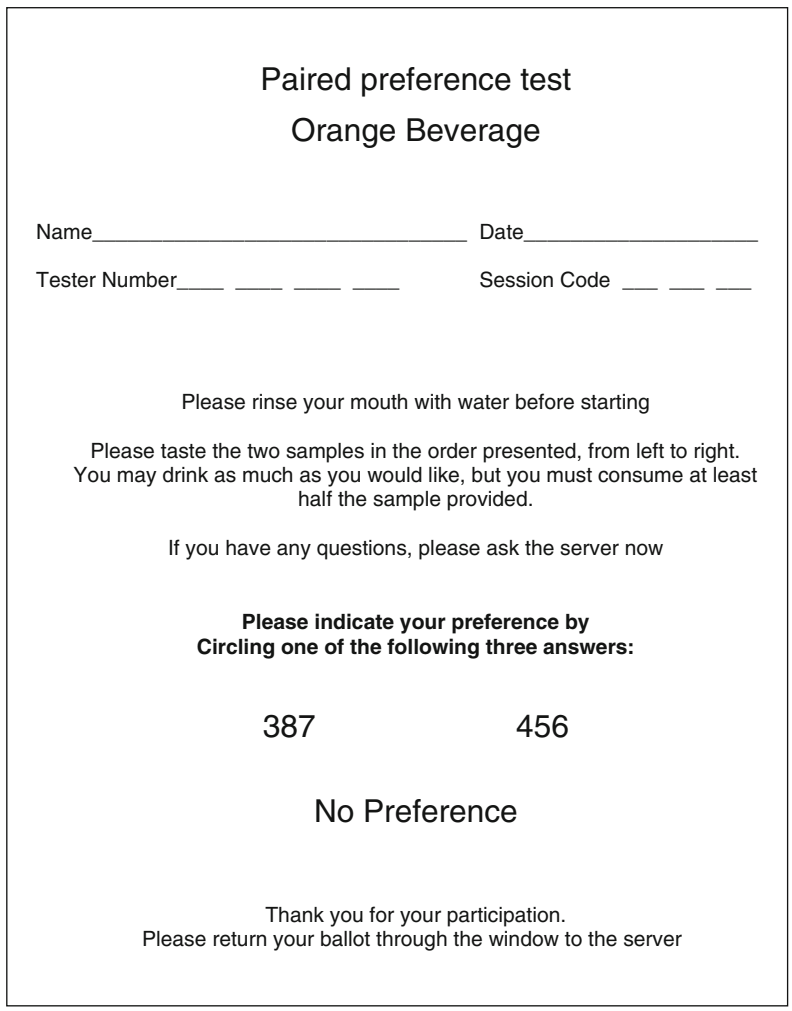
and our *p*-calculation proceeds as follows:

$$p = \Phi\left(\frac{|0.0092| - 0.05}{0.02031}\right) - \Phi\left(\frac{-|0.0092| + 0.05}{0.02031}\right) = .0204$$

and so the value of 0.0204 (p<0.05) is good evidence that the true population proportion lies within the interval of 0.5±0.05, based on the obtained proportion in our data. Further information on preference tests in claim substantiation is given in Chapter 19.

13.4 Non-forced Preference

A no-preference option is sometimes included in a paired preference test, in spite of the fact that it complicates the analysis considerably. Many practitioners question whether it is worth the effort. However, it may be required due to the legal regulations for claim substantiation (ASTM, 2008). The size of the no-preference response group could be useful information in its own right. Also, some have proposed that it could help decide if an equal preference split was due to indifferent response or whether there might in fact be stable groups of about equal size with strong preferences (Gridgeman, 1959). In other words, a 50/50 split in a preference test is no clear win for either product, but might be the result of two segments of consumers that each likes different versions of the product. Unfortunately this situation is not resolved by offering the no-preference option. A very robust finding is that people will avoid making this response, even with physically identical samples (Chapman and Lawless, 2005; Chapman et al., 2006; Marchisano et al., 2003). Although Gridgeman was correct in stating that the no-preference option offers additional information, the response option does not in fact resolve the question of stable segments. Nonetheless, the no-preference response may be required by someone requesting the test or be included for other legal considerations (ASTM, 2008). A ballot for a preference test with the no-preference option is shown in Fig. 13.2.


Paired preference test

Orange Beverage

Name_____ Date_____

Tester Number____ ____ ____ ____ Session Code ___ ___ ___

Please rinse your mouth with water before starting

Please taste the two samples in the order presented, from left to right. You may drink as much as you would like, but you must consume at least half the sample provided.

If you have any questions, please ask the server now

Please indicate your preference by Circling one of the following three answers:

387 456

No Preference

Thank you for your participation.
Please return your ballot through the window to the server

Fig. 13.2 Ballot for a paired preference test when a choice is not forced and a no-preference response is allowed.

There are four approaches to dealing with the no-preference responses in a paired preference test: (1) eliminate them, (2) apportion them, (3) use a confidence interval analysis, (4) use a signal detection model analysis. To some degree the choice of how to deal with the responses depends on what you assume the basis for the no-preference choices was. For example, if you assume that people responding "no preference" really do not care, then it would make sense to apportion them 50/50 to each of the other response options. Unfortunately there is rarely any good basis (i.e., some additional information or data) for making such assumptions.

In the first approach, they are simply discounted. The analysis proceeds using the simple two-tailed binomial statistics for a difference of proportions. This approach lowers the sample size and thus the power of the test to finding significant preferences. This approach seems reasonable if the actual number is fairly low, say less than 10% of the responses. If the proportion of no-preference responses is high, say above 20%, but there is still a significant result in the remaining sample, the test result may be qualified in the report as follows: "*Among those expressing a preference*, there was a significant preference for product X" (see ASTM, 2008). The researcher should also state the raw percentages including the size of the no-preference response group.

In the second approach, the no-preference responses are apportioned. One way to do this is simply split them 50/50 into the existing preference groups. This maintains the sample size but does dilute the test result somewhat since the 50/50 split is what one expects by chance. Another option is to divide the no-preference votes in proportion to those who did express a preference. That is, if there is a 60/40 split among those who did express a preference, the no-preference votes would be apportioned to those groups in the same 60/40 ratio. This approach is based on some findings from Odesky (1967), who found that the proportions of people expressing a preference when the no-preference option was available were similar to the obtained proportions when choice was forced. However, this finding has been questioned and it may not be a valid generalization (Angulo and O'Mahony, 2005). In some cases for advertising claims of product superiority, they must be apportioned to the competitor's product, providing a very strict hurdle for proving a significant preference (ASTM, 2008).

In the third approach, a confidence interval may be constructed around each proportion of those who did express a preference. If the confidence intervals do not overlap, one may conclude that there is a significant preference win for the product with the larger proportion. This approach is justified if the sample size is large ($N > 100$) and the number of no-preference choices is relatively low (less than 20%). The formula for the confidence interval is

$$\mathrm{CI} = \frac{\chi^2 + 2X \pm \sqrt{\chi^2\left[\frac{\chi^2 + 4X(N-X)}{N}\right]}}{2\left(N + \chi^2\right)} \tag{13.5}$$

where X is the number of panelists preferring one of the two products, N is the total number of panelists, and χ^2 is the chi-square value for 2 df or 5.99. A worked example is shown in Appendix 3 at the end of this chapter.

A fourth approach is based on a Thurstonian model, which states that the degree of difference in liking must exceed a person's criterion, called tau, below which they will choose a no-preference option. This approach is based on an extension of the paired comparison test for difference with an "equal" option (see Braun et al., 2004). The degree of preference/difference is called *d*-prime (or in some literature, the Greek letter delta). See Chapter 5 for more information on Thurstonian models. The size of the difference, *d*-prime (d'), and the tau criterion can be solved by the following equations:

$$z_1 = (-\mathrm{tau} - d')/\sqrt{2} \tag{13.6a}$$

$$z_2 = (+\mathrm{tau} - d')/\sqrt{2} \tag{13.6b}$$

where z_1 is the z-score for the proportion preferring product A and z_2 is the z-score for the sum of the proportion for A *and* the no-preference votes.

This provides two equations in two unknowns, which can be solved for tau and d'. Once d' is obtained, a z-test can be performed which will tell if the d' value is significantly different from zero. The standard deviation (S) is found from the value $\sqrt{B/N}$. Then $z = d'/S$, which must be greater than 1.96. This would be taken as evidence of a significant preference. *B*-values (see Bi, 2006) are found for different d' values in Table O.

13.5 Replicated Preference Tests

Although replication is not often done with preference tests, there are a few good reasons to consider replicating in the design of a sensory test. First, the effort and cost to recruit, screen, and get a consumer panelist on-site is far larger than the time and cost associated with the conduct of the actual test. So why not get some additional information while the consumers are present? Second, there is evidence that many people will change their minds from trial to trial. Koster et al. (2003) summarized some data with children and novel foods and found, on the average, less than 50% consistency from trial to trial in the most liked (chosen) food and somewhat higher levels with adults. This result is similar to findings of Chapman and Lawless (2005) who found about 45% switching in a test of milks that allowed the no-preference option, although the marginal proportions of the groups that preferred each milk remained stable. Finally, repeated testing may be the only way to answer the question of whether a 50/50 split in preference represents equal appeal of the two products (or lack of preference) or whether there are two stable segments of equal size preferring each version of the product. In other words, repeated testing could yield evidence for stable segments, in which case both versions of the product become candidates for further consideration, development, and marketing.

Analysis of these data can be simple or more complex. A simple approach is to consider the expected value one would obtain on a replicated test if people were behaving randomly (i.e., had no real stable preference). For example, on a test with two trials, one would expect 25% of the consumers to choose product A twice, 25% to choose product B twice, and 50% to choose one of each. Given these expected frequencies, a simple chi-square analysis can be performed to see if the results differ from chance expectations (Harker et al., 2008; Meyners, 2007). The approach can obviously be extended to tests with more than two trials. In the case of two trials the data can also be cast in a 2×2 contingency table (showing the preferences on trial 1 in the columns and trial 2 in the rows) which classifies each consumer into one of four cells. A chi-square analysis can be conducted on this table as well. Individual binomial tests on proportions can also be done to compare the group consistently preferring A to that consistently preferring product B (but note that this lowers the N by eliminating the people who switched choices).

A variety of other approaches have been suggested, and several are reviewed in the extensive paper by Cochrane et al. (2005). A beta-binomial approach is discussed by Bi (2006), similar to the beta-binomial analysis used with replicated difference tests. This approach is informative as it not only gives some overall statistical significance level but also calculates the gamma statistic which shows how random versus consistently grouped the panel appears to be behaving. If there are stable segments with consistent preferences, a significant (non-zero) gamma statistic could be obtained. Further information on gamma and the beta-binomial is given in Chapter 4. Also, this analysis formally recognizes that the data on the two trials are related, and that individual consistency or variation is part of the picture (i.e., "overdispersion"). In the next section we will see an approach that combines both replication and allowance for the no-preference option.

13.6 Replicated Non-forced Preference

Next we will look at perhaps the most complicated situation, involving both replication and the use of the no-preference response option. A simple yet elegant approach to the issue of stable preference segments is found in a 1958 paper by George E. Ferris (also discussed in Bi, 2006). An alternative analysis uses the Dirichlet-multinomial (DM) model, an extension of the beta-binomial approach to the multinomial situation. This is discussed in Gacula et al. (2009) and illustrated with a worked example. The DM model is attractive because it considers the heterogeneity of the consumer group, i.e., whether there are different pockets of individuals with consistent patterns or response across replications, much like the gamma statistic of the beta-binomial.

Ferris called his approach the k-visit method of consumer testing and the examples he gives are in sequential home use tests for paired preference, with a no-preference option (for our purposes, $k = 2$). In other words, there were at least two separate tests conducted using the same consumers, with a preference choice between products "A" and "B." There are several reasonable assumptions made, which are as follows: (1) persons with a consistent preference

would choose A on both trials or B on both trials; (2) sometimes consumers would choose to respond A or B even if they had no preference or could not discriminate the samples, in order to please the experimenter; (3) thus the double response for A (or B) could include some people who had no preference (or could not discriminate); (4) the amount of switching responses (i.e., inconsistency) that goes on was a clue to the proportion of false preference expressed by non-preferring consumers. Ferris was then able to get some good estimates of (1) the true proportion of consumers who consistently liked A, (2) the true proportion of consumers who consistently liked B, (3) and the true proportion of everyone left over (people with consistent non-preference, non-discriminators, and switchers). The calculations for this analysis and some simple tests for statistical significance are shown below. Note that if you wanted to just test for a difference between the two trials, you could do a Stuart–Maxwell test, but this would not answer the questions about a product win nor give any information about stable segments.

There are several areas for minor concern in using this analysis. First, it is possible that a person could have a real preference on the first trial, but change his or her mind on the second trial (Koster et al., 2003). For example, I might like product A on trial one, then get tired of the product, fatigued by the test, lose interest, etc., so that on trial two I really do not care any more and so I respond "no preference." Second, the opinion could actually change from trial to trial for good reason, e.g., perhaps with some very sweet foods which are liked at first, then become too cloying. Third, it is possible that the preference is not all-or-nothing (as assumed in this model) but rather that people have some percentage of the time they like A and another percentage they like B and perhaps even some of the time when the same person simply does not care. If any of these is the case, then this model is a bit too simple, but it can still be applied for decision-making purposes.

Here are the two experimental questions: Is there a significant preference? Is there a consistent segment even if there is not a "win"? The test design uses N consumers, who participate in each of two paired tests and must respond either "Prefer *A*," "Prefer *B*," or no preference on each questionnaire. The samples have different blind codes in each trial so the consumers are unaware that the test is repeated. The data are tabulated in a 3 × 3 table as shown below, with trial one answers in the columns and trial two answers in the rows. We have retained the original notation of Ferris in the calculations (see also Bi, 2006). N's in the table are the actual frequency counts (not proportions). The subscript a means the consumer chose product A, b for product B, and o for no preference. Thus N_{ao} refers to the number of consumers who choose product A on the first trial and said "no preference" on the second. N without a subscript is the total number of consumers in the test (Table 13.3).

Table 13.3 Notation for analysis of non-forced replicated preference

	Trial 1 response		
	Prefer A	No preference	Prefer B
Trial 2 response			
Prefer A	N_{aa}	N_{oa}	N_{ba}
No preference	N_{ao}	N_{oo}	N_{bo}
Prefer B	N_{ab}	N_{ob}	N_{bb}

Calculations:

(1) Tabulate some "inconsistent behavior" totals. These will help to shorten some of the further calculations below.

$$N_y = N_{ao} + N_{oa} + N_{bo} + N_{ob} \tag{13.7a}$$

(some people switching to or from no pref)

$$N_x = N_{ab} + N_{ba} \tag{13.7b}$$

(some people switching products, A to B or B to A)

$$M = N - N_{aa} - N_{bb} \tag{13.7c}$$

(all those showing no consistent preference)

(2) Calculate a consistency parameter, p.

$$p = \frac{M - \sqrt{\left[M^2 - (N_{oo} + N_y/2)(2N_x + N_y)\right]}}{2N_{oo} + N_y} \tag{13.8}$$

(3) Figure the maximum likelihood estimates, fitted proportions π_A and π_B and π_o. These are the true or population estimates we are trying to obtain from the analysis. Note that π_A does not equal N_{aa}/N. In other words we are going to adjust the p_A estimate based on the notion that some of the consistent N_{aa} responders may have included some non-preferring consumers who were just trying to

please the experimenter. This was a remarkable insight on the part of Ferris, given the rediscovery of false preferences in the tests with identical products, just recently (see Chapman et al., 2006; Marchisano et al., 2003). This is the reason for the consistency parameter, p, in Eq. (13.8). Put another way, we can get a good estimate of the amount of random responding or false consistency by looking at the amount of no-preference votes and the amount of switching that went on in the data set.

$$\pi_A = \frac{\left[N_{aa}(1-p^2)\right] - \left[(N-N_{bb})p^2\right]}{N(1-2p^2)} \quad (13.9a)$$

$$\pi_B = \frac{\left[N_{bb}(1-p^2)\right] - \left[(N-N_{aa})p^2\right]}{N(1-2p^2)} \quad (13.9b)$$

$$\pi_o = 1 - \pi_A - \pi_B \quad (13.9c)$$

Note that it is these adjusted proportions π_A and π_B that we really want to know giving us estimates for the sizes of the consistent segments, not simply N_{aa}/N or N_{bb}/N which are the proportions of supposedly "consistent" consumers in the original data. The original raw data are probably tainted by some random responders or no-preference responders trying to please the experimenter.

(4) These are point estimates, so in order to get confidence limits and do some statistical testing, we need some variability estimates, too. Next, calculate some variance and covariance estimates using the parameters calculated so far:

$$\mathrm{Var}(\pi_A) = \frac{\pi_A(1-\pi_A) + (3\pi_o p^2)/2}{N} \quad (13.10a)$$

$$\mathrm{Var}(\pi_B) = \frac{\pi_B(1-\pi_B) + (3\pi_o p^2)/2}{N} \quad (13.10b)$$

$$\mathrm{COV}(\pi_A, \pi_B) = \frac{(\pi_o p^2/2) - (\pi_A \pi_B)}{N} \quad (13.10c)$$

(5) Now we can see if there is a win for product A or for product B. We test for difference of π_A versus π_B using a Z-test:

$$Z = \frac{\pi_A - \pi_B}{\sqrt{\mathrm{Var}(\pi_A) + \mathrm{Var}(\pi_B) - 2\mathrm{Cov}\pi_A\pi_B}} \quad (13.11)$$

This assumes we have a good-sized consumer test with $N > 100$ and preferably $N > 200$.

(6) If needed, test for the size of a consistent segment versus some benchmark. For example, if we needed to see the proportion of true preference for product A greater than 45% to make some advertising claim or take some action in further product marketing we would also use a Z-test against that benchmark, e.g., $\pi_A = 0.45$ (a 45% segment size).

$$Z = \frac{\pi_A - 0.45}{\sqrt{\mathrm{Var}(\pi_A)}} \quad (13.12)$$

We can also test to see whether we have surpassed a certain size of the no-preference segment by a Z-test using π_o and its benchmark. For example, we might have an action standard which states that the no-preference segment must be at or below 20%. We might also have an action standard that included several of these tests. For example, if π_o was less than 20% and π_A and π_B both higher than 35%, we might explore marketing two versions of the product.

A worked example of this analysis for the Ferris k-visit method is shown in Appendix 1 at the end of this chapter.

13.7 Other Related Methods

13.7.1 Ranking

In these tests the consumers are asked to rank several products in either descending or ascending order of preference or liking. The participants are usually not allowed to have ties in the ranking, thus the method is usually a forced choice. The paired preference test is a special case of a preference ranking, when the participant is asked to rank only two samples. Note that rankings do not give a direct estimate of the size of any difference in preference, although it is possible to derive some Thurstonian scale values from the proportions. Preference ranking is intuitively simple for the consumer, it can be done quickly and with relatively little effort. The ranks are based on a frame of reference that is internal to the specific set of products and

thus the consumer does not have to rely on memory. A disadvantage of preference ranking is that the data from different sets of overlapping products cannot be compared since the rankings are based on this internal frame of reference. Visual and tactile preference rankings are relatively simple but the multiple comparisons involved in ranking samples by flavor or taste can be very fatiguing. A sample score sheet is shown in Fig. 13.3. An example of the use of ranking in sensory consumer testing is found in the study by Tepper et al. (1994).

13.7.2 Analysis of Ranked Data

The data are ordinal and are treated as nonparametric. Preference ranking data may be analyzed either by using the so-called Basker tables (Basker 1988a, b), those by Newell and MacFarlane (1987) (see Table J) or the Friedman test (Gacula and Singh, 1984). The tables require that the panelists were forced to make a choice and that there are no tied ranks. The Friedman test is tolerant of a small number of tied opinions. Examples follow.

Preference test - Ranking

Fruit Yogurt

Name_____ Date_____

Tester Number____ _____ _____ ____ Session Code___ ___ ___

Please rinse your mouth with water before starting.
You may rinse again at any time during the test if you need to.

Please taste the five samples in the order presented, from left to right.
You may re-taste the samples once you have tried all of them.

Rank the samples from most preferred to least preferred using the following numbers:
1 = most preferred, 5 = least preferred

(If you have any questions, please ask the server now)

Sample	Rank (1 to 5) (ties are NOT allowed)
387	_____
589	_____
233	_____
694	_____
521	_____

Thank you for your participation,
Please return your ballot through the window to the server.

Fig. 13.3 Sample ballot for a ranking test.

To use Table J, assign numerical values to each of the n products (1–n) starting with one for the most preferred sample. Then sum the values across the group of panelists to obtain a rank sum for each sample. Next, consult a table for rank totals (Table J). In this example, six consumers ranked seven products using a rank scale with 1 = preferred most and 7 = preferred least. Clearly we would never, in real life, do a consumer ranking study with only six panelists, but this is only an example to illustrate the calculations associated with the tables. In this example rank totals for the products A through G are as follows:

Product	A	B	C	D	E	F	G
Rank total	18	28	20	10	26	32	34
Significance group	ab	ab	ab	a	ab	ab	b

The table indicates that the critical difference value for six consumers and seven products is 22. Products with the same letter below their rank sum are not significantly different by this test. Product D is thus significantly more preferred to Product G with no other preferences observed in this comparison.

The Friedman test is the nonparametric equivalent to the two-way analysis of variance without interaction. The Friedman test equation is based on the χ^2 distribution:

$$\chi^2 = \left\{ \frac{12}{[K(J)(J+1)]} \left[\sum_{j=1}^{J} T_j^2 \right] \right\} - 3K(J+1) \tag{13.13}$$

where

J = number of samples
K = number of panelists
T_j = rank total and
degrees of freedom for χ^2 = (J–1)

Once we determine that the χ^2 test is significant then a comparison of rank total separation is done to determine which samples differ in preference from one another. Informally we have called the value that determines the significant difference in preference ranking the "least significant ranked difference" or LSRD, in an analogy to the LSD test used to test differences of means after analysis of variance.

$$\text{LSRD} = t\sqrt{\frac{JK(J+1)}{6}} \tag{13.14}$$

where

J = number of samples
K = number of panelists and
t is the critical t-value at a = 5% and degrees of freedom = 1

To return to the example previously used to explain the use of the Newell and MacFarlane or Basker tables, we will now use those data in a Friedman test. First, according to the overall test for differences, χ^2 = 15.43. The critical χ^2-value for α at 5% and six degrees of freedom is 12.59. Therefore, the preference ranks for this data set differ significantly at $p < 5\%$. We now need to determine which products were ranked significantly higher in preference from one another. The least significant rank difference (LSRD) for the Friedman test is calculated from Eq. (13.14). In the example above, the LSRD = 14.7, giving the following pattern of statistical differences in preference:

Product	A	B	C	D	E	F	G
Rank total	18	28	20	10	26	32	34
Significance group	ab	bc	abc	a	bc	bc	c

Products sharing the same significance group letter show no difference in ranked preference. This pattern can be summarized as follows: Product D is significantly more preferred than products B, E, F, and G. Product G is significantly less preferred than products A and D. Note that the results from the Basker table were more conservative than the results from the Friedman test. Slight differences will sometimes occur in these two approaches.

13.7.3 Best–Worst Scaling

Best–worst scaling (also called maximum difference or max-diff) is a technique in which more than two products are given and the person chooses the one he or she likes best and the one he or she likes the least.

Although the data can result in scaled values for the overall appeal of each product, it is really a choice method and therefore falls into the same class as paired preference choice and ranking. This method has been popular in other areas like marketing research but has recently received some attention for food testing (Hein et al., 2008; Jaeger and Cardello, 2009; Jaeger et al., 2008). The psychometric models for the data from this method suggest that the data can yield scores on an interval and sometimes ratio scale (Finn and Louviere 1992). The method would seem to have some psychological benefit in terms of ease of use, under the notion that it is easier for people to differentiate products at the end of a continuum, as opposed to what is in the middle. However, it may not work well with very fatiguing products like wines (see Mueller et al., 2009).

The method works as follows: the set of products is partitioned into a series of three or more products, with each product representing an equal number of times. In one example, there are four products presented in triads, with four possible combinations of three products at a time (Jaeger et al., 2008). The consumer sees all four triads in random order and with each triad indicates which product is liked best and which is liked the least. There are then two ways of handling the data. The simplest is to sum up the number of times a product was liked best, and then subtract from that the number of times it was liked the least. This differencing procedure generates a score from every panelist which can then be submitted to analysis of variance or any other parametric statistical test. An alternative analysis is to fit the data by a multinomial logistic analysis, which will also yield scores and variance estimates, as well as test for differences among products and any other variables in the test. It is claimed that the simple difference scores have interval scale properties and that the multinomial logistic analysis has true ratio properties (one of the only scales for which this has any reasonable substantiation). See Finn and Louviere (1992) for the psychometric model upon which these claims are based.

Given the "natural" user-friendly nature of the task and the potential to get detailed interval or ratio scale data from the method, it has some appeal for food preference testing. The only drawback is that it requires much more testing to do all the possible combinations than a more straightforward acceptance tests, although the number of trials is somewhat more efficient than giving all possible pairs in a set of multiple preference tests. Furthermore the number of combinations will decrease as more products are included in each trial (there is no law against using more than three), although the task may become more complex for the consumer as they have more choices to consider. Recent data in real food tests show the task to be easy to do, and the resulting data are as good or better than acceptance scales (9-point or labeled affective magnitude (LAM) scales, see Chapter 14) in differentiating among products (Jaeger and Cardello, 2009; Jaeger et al., 2008). This was confirmed by a comparison to the 9-point, LAM, unstructured line scale and ranking test by Hein et al. (2008). However, it should be noted that when the acceptance scales were replicated in that study, in order to better equate the total number of product exposures and judgments, the scaling data improved in terms of product differentiation, with the 9-point scale improving markedly on the second trial. Thus it seems that the better differentiation with best–worst may simply be a function of having more product comparisons.

The future for this choice method presents some opportunities. First, it may be well suited to food preference surveys, in which no foods are tasted (only names of foods presented) and thus the fatigue factor is not present. Second, incomplete block designs should be applicable to the situation where many products are tested, reducing the burden of a large number of triads that would be needed in a complete design. In their original study, Jaeger et al. (2008) also found that the best–worst scaling provided more useful data for preference mapping than did the line scaling for acceptance. That is the data were better fit and more information was forthcoming from the preference maps (Jaeger et al., 2008). There were also better fits to vectors from descriptive analysis attributes that were projected into the preference maps (see Chapter 18, under external preference mapping).

13.7.4 Rated Degree of Preference and Other Options

A few additional options are available for preference and ranking tests which we will briefly describe here, mostly because they appear in the literature. One option is to provide a rating scale for the degree of preference (Filipello, 1957). The simple preference test

Preference Test

Funistrada

Name_____Date_____

Tester number: ___ ___ ___ ___ Session Code____

Please rinse your mouth with water before starting

Please taste the two samples of Funistrada in the order presented, from left to right. You may taste as much as you would like but you must consume at least one-third of the sample.

If you have any questions, please ask the server now.

Check one answer that best describes your preference:

____ Strongly prefer 387 over 589

____ Prefer 387 over 589

____ Slightly prefer 387 over 589

____ No preference

____ Slightly prefer 589 over 387

____ Prefer 589 over 387

____ Strongly prefer 589 over 387

Thank you for your participation.
Please return your ballot through the window to the server.

Fig. 13.4 Sample ballot for a preference test with rated degree-of-preference and a no-preference option is allowed.

does not indicate the strength of the consumer's opinion. Their preferences might be a small matter or they might have a strong favorite among the alternatives. To get this information a rating scale could be used as shown in Fig. 13.4. If the sample size is large, as it is usually in a consumer test, the choices can be transcribed as –3 to +3 and the resulting distribution tested against a population value of zero by a simple *t*-test. As in the case of the just-right scale discussed in the next chapter, it is important to look at the frequency counts in each category and not just the mean score, in case there are unexpected patterns of response (perhaps two groups that strongly prefer different products). Scheffe' (1952) proposed an analysis of variance model for these kinds of data.

Another option that is sometimes used is "dislike both equally" and "like both equally" as a variation of the no-preference option. This obviously provides

some additional information. The data should be examined and then both of these combined and treated as if they were the usual no-preference option. An additional category is "do not care" which would provide a response for those individuals who have no preference and no liking or disliking anyway.

13.8 Conclusions

Preference testing is primarily based on a simple choice procedure. A consumer must choose from a pair of products which one is liked best. The analysis is straightforward: A two-tailed test is conducted against a binomial expectation of a 50/50 split under the null hypothesis. If one product has a significantly higher percentage than the expected 50%, a win is declared and the product may move forward in the development scenario.

Two complications arise, however, in common practice. The first is the use of a "no-preference" option. This non-forced preference task is desired for some advertising claim substantiation scenarios due to regulatory agency requirements. However, it complicates the analysis and under most product development situations, the option is not needed. A choice is forced under the assumption that if there is no clear preference (or people just do not care) they will split their votes according to the null expectations. Recent research demonstrates that people avoid the no-preference option, even with physically identical products, so its utility is questionable.

The second complication arises from replication. Replication is less troublesome than the non-forced or no-preference option. It offers a chance to examine the stability of preference choices. In the case where there is nearly an even split among the two products, replication can offer some evidence as to whether there are stable segments with strong loyalty to one of the two versions of the product or whether a substantial amount of shifting might occur. Various analyses of replicated preference are available, including a beta-binomial analysis similar to the one used for discrimination tests. The sensory scientist should weigh the advantages and disadvantages of these procedures carefully before choosing one of them over the simpler and more straightforward paired preference test.

Appendix 1: Worked Example of the Ferris *k*-Visit Repeated Preference Test Including the No-Preference Option

A consumer test with 900 respondents is completed with the following results.

Questions: (1) Is there a significantly higher preference for product A or product B? (2) Is the preference for the winning product higher than 45%? (Example from Ferris (1958) and Bi (2006), pp. 72–76).

	Prefer A	No preference	Prefer B
Prefer A	$N_{aa} = 457$	$N_{oa} = 12$	$N_{ba} = 14$
No preference	$N_{ao} = 14$	$N_{oo} = 24$	$N_{bo} = 17$
Prefer B	$N_{ab} = 8$	$N_{ob} = 11$	$N_{bb} = 343$

($N = 900$)

Here are the basic equations we need:

$$N_y = N_{ao} + N_{oa} + N_{bo} + N_{ob}$$
$$N_x = N_{ab} + N_{ba}$$
$$M = N - N_{aa} - N_{bb}$$

$$p = \frac{M - \sqrt{\left[M^2 - (N_{oo} + N_y/2)(2N_x + N_y)\right]}}{2N_{oo} + N_y}$$

So for this data set:

$M = 100$ (all those not showing AA or BB behavior, i.e., a consistent choice for one of the two products)

$$N_x = 8 + 14 = 22$$
$$N_y = 14 + 12 + 17 + 11 = 54$$
$$p = 0.257$$

Now we need the equations for the best estimates of each segment/proportion:

$$\pi_A = \frac{\left[N_{aa}(1 - p^2)\right] - \left[(N - N_{bb})p^2\right]}{N(1 - 2p^2)}$$
$$\pi_B = \frac{\left[N_{bb}(1 - p^2)\right] - \left[(N - N_{aa})p^2\right]}{N(1 - 2p^2)}$$

and

$$\pi_o = 1 - \pi_A - \pi_B$$

Now we can get our segment size estimates:

$\pi_A = 0.497$ or 49.7% true preference for product A.
$\pi_B = 0.370$ or 37% true preference for product B.
$\pi_o = 0.133$ or 13.3% no real preference.

Next, we need the variability and covariance estimates for the *Z*-tests:

$$\mathrm{Var}(\pi_A) = \frac{\pi_A(1-\pi_A) + (3\pi_o p^2)/2}{N}$$

$$\mathrm{Var}(\pi_B) = \frac{\pi_B(1-\pi_B) + (3\pi_o p^2)/2}{N}$$

$$\mathrm{COV}(\pi_A, \pi_B) = \frac{(\pi_o p^2/2) - (\pi_A \pi_B)}{N}$$

$\mathrm{Var}(\pi_A) = 0.000296$ (so π_A is 47.9% ± 1.7%, $0.017 = \sqrt{0.000296}$)
$\mathrm{Var}(\pi_B) = 0.000297$
$\mathrm{COV}(\pi_A, \pi_B) = -0.000198$

Now for the hypothesis tests:

$$Z = \frac{\pi_A - \pi_B}{\sqrt{\mathrm{Var}(\pi_A) + \mathrm{Var}(\pi_B) - 2\mathrm{Cov}\pi_A\pi_B}}$$

Note that this is a little different from the simple binomial test for paired preference. In the simple case we test the larger of the two proportions against a null value of 0.5. In this case we actually test for a difference of the two proportions, since we do not expect a 50/50 split any more with the no-preference option.

So the *Z* for test of A versus B gives $Z = 4.067$, an obvious win for product A.

Finally, a test against a minimum required proportion or benchmark:

$$Z = \frac{\pi_A - 0.45}{\sqrt{\mathrm{Var}(\pi_A)}}$$

Z-test for A versus benchmark of 0.45 (45%).

Appendix 2: The "Placebo" Preference Test

In this method a pair of identical samples are given on one of two preference test trials (Alfaro-Rodriguez et al., 2007; Kim et al., 2008). These physically identical samples are not expected to differ, hence the parallel to a placebo, or a sham medical intervention with no expected therapeutic value. In theory, this could provide a baseline or control condition, against which performance in the preference test (with a no-preference option) could be measured. However, the amount of information gained from this design is relatively small and once again the analysis becomes more complicated. For these reasons, the sensory professional should consider the potential cost, additional analysis, and interpretations that are necessary. A recommended analysis is given at the end of this section.

Issues and complications. The use of a no-preference option was proposed to be a solution to the problem of a 50/50 preference test result, which could result from two stable segments of consumers who have a (perhaps strong) preference for each of the versions, respectively. Hence the idea was to offer a no-preference option with the reasoning that if there were no preferences (rather than stable segments) respondents should opt for the no-preference response. However, persons given identical samples will avoid the no-preference option 70–80% of the time, as discussed earlier in this chapter. So an answer concerning the question of stable segments cannot be obtained by this approach. Evidence for stable segments could be found by replicated testing or by converging evidence from different kinds of tests and/or questions.

Possible analyses. It might be tempting to just eliminate those persons expressing a preference for one of the identical pair members, on the grounds that such respondents are biased. However, this could eliminate 70–80% of the consumers. It is generally not advisable to pre-select consumers on any other grounds than their product consumption, and this approach eliminates individual who are in fact a portion of the representative population we are trying to generalize the results to. Such persons may not be "biased" in any dysfunctional way. Even identical samples may seem different from moment to moment. Furthermore, individuals are

clearly responding to the demands of the task (you are expecting them to state a preference in a preference test!).

If historical data are available concerning the frequencies of response to identical pairs, a chi-square test can be performed as shown below. Note that this analysis cannot be done for a test in which the same subjects provide the "placebo" judgments because the chi-square test assumes independent samples.

Placebo analysis #1. Using historical data for expected frequencies.

Cells in the top row (A1, NP1, and B1) are expected frequencies (expected proportion × *N* judges). Cells in row 2 (A2, NP2, and B2) are the obtained data, frequencies of response in the preference comparison for the actual (different) test samples.

	Prefer A	No preference	Prefer B
Historical data for identical samples	A1	NP1	B1
Test samples	A2	NP2	B2

Placebo analysis #2. The same consumers participate in the placebo trial and the test pair.

If the same people are used for the "placebo" trial and the normal preference comparison trial, an option is to recast the data into another 2 × 3 table, with the rows now representing whether the individual expressed a preference (or not) on the placebo trial. The columns remain Prefer A, no preference, Prefer B. A chi-square test will now tell you whether the proportions of preference changed comparing those who elected the no-preference option for the placebo pair versus those who expressed a "false preference" for the identical pair. This does not provide evidence for the existence of any stable segments nor will it tell you if there is a significant preference from this analysis alone.

If there is no significant chi square, you can feel justified in combining the two rows. If not, you may then analyze each row separately. The correct analyses are as stated in the no-preference Section 13.4: eliminating no-preference judgments, apportioning them, doing a test of *d*-prime values, or the confidence interval approach if the assumptions are met.

Each judge is classified into one of the six cells, A1, A2, B1, B2, NP1, or NP2. The first row contains the data from people who reported no preference on the placebo pair. The second row contains the data from people who expressed a preference with the placebo pair. A chi-square test will now show whether the two rows have different proportions. If not, the rows may be combined. If they are different, separate analyses may be performed on each row, using the methods of analysis of the no-preference option discussed earlier in the chapter.

		Response—Trial 2:	
	Prefer A	No preference	Prefer B
Response, Trial 1:			
No preference on Trial 1	A1	NP1	B1
Preference on Trial 1	A2	NP2	B2

Appendix 3: Worked Example of Multinomial Approach to Analyzing Data with the No-Preference Option

This approach yields multinomial distribution confidence intervals for "no-preference" option, Data should be from a large test, $N > 100$, and the no-preference option was used rarely (<20%) (Quesenberry and Hurst, 1964, p. 193, Eq. (2.9)).

Upper and lower confidence interval boundaries are given by

$$\mathrm{CI} = \frac{\chi^2 + 2X \pm \sqrt{\chi^2\left[\chi^2 + \dfrac{4X(N-X)}{N}\right]}}{2\left(N + \chi^2\right)}$$

where

$\chi^2_{\text{critical}} = 5.99$ for α at 5% and 2 df,
X = number of observed preference votes for one sample,
N = sample size.

Example: For: $N = 162$, $X_1 = 83$, $X_2 = 65$, and no preference = 14.

First find the confidence interval for product X_1 (choices = 83/162)

$$\text{CI} = \frac{5.99 + 2(83) \pm \sqrt{5.99\left[\frac{5.99+4(83)(162-83)}{162}\right]}}{2\,(162 + 5.99)} = \frac{171.99 \pm 31.15}{335.98}$$

which gives an interval from 0.42 *to* 0.60 *for product* X_1.

Then find the confidence interval for product X_2 (choices = 65/162)

$$\text{CI} = \frac{5.99 + 2(65) \pm \sqrt{5.99\left[\frac{5.99+4(65)(162-65)}{162}\right]}}{2\,(162 + 5.99)} = \frac{135.99 \pm 30.39}{335.98}$$

which gives an interval from 0.31 *to* 0.50 *for product* X_2. The lower bound of the higher proportion (0.42) overlaps with the upper bound of the lower proportion (0.50). We therefore conclude that there is not enough evidence for any difference in the preference proportions.

References

Amerine, M. A. and Roessler, E. B. 1983. Wines: Their Sensory Evaluation. Freeman, San Francisco.

Angulo, O. and O'Mahony, M. 2005. The paired preference test and the no preference option: Was Odesky correct? Food Quality and Preference, 16, 425–434.

ASTM International. 2008. Standard Guide for Sensory Claim Substantiation. Designation E 1958–07. Vol. 15.08 Annual Book of ASTM Standards. ASTM International, Conshohocken, PA, pp. 186–212.

Basker, D. 1988a. Critical values of differences among rank sums for multiple comparisons. Food Technology, February 1988, 79–84.

Basker, D. 1988b. Critical values of differences among rank sums for multiple comparisons. Food Technology, July 1988, 88–89.

Bech, A. C., Engelund, E., Juhl, H. J., Kristensen, K. and Poulsen, C. S. 1994. Qfood: Optimal design of food products. MAPP Working Paper 19, Aarhus School of Business, Aarhus, Denmark.

Beckman, K. J., Chambers, E. IV and Gragi, M. M. 1984. Color codes for paired preference and hedonic testing. Journal of Food Science, 49, 115–116.

Berglund, B., Berglund, U. and Lindvall, T. 1975. Scaling of annoyance in epidemiological studies. Proceedings, Recent Advances in the Assessments of the Health Effects of Environmental Pollution. Commission of the European Communities, Luxembourg, Vol. 1, pp. 119–137.

Bi, J. 2006. Sensory Discrimination Tests and Measurements. Blackwell, Ames, IA.

Braun, V., Rogeaux, M., Schneid, N., O'Mahony, M. and Rousseau, B. 2004. Corroborating the 2-AFC and 2-AC Thurstonian models using both a model system and sparkling water. Food Quality and Preference, 15, 501–507.

Cardello, A. V. and Schutz, H. G. 2006. Sensory science: Measuring consumer acceptance. In: Handbook of Food Science, Technology and Engineering, CRC Press, Boca Raton, FL. Vol. 2, Ch. 56.

Caul, J. 1957. The profile method of flavor analysis. Advances in Food Research, 7, 1–40.

Chapman, K. W., Grace-Martin, K. and Lawless, H. T. 2006. Expectations and stability of preference choice. Journal of Sensory Studies 21, 441–455.

Chapman, K. W. and Lawless, H. T. 2005. Sources of error and the no-preference option in dairy product testing. Journal of Sensory Studies 20, 454–468.

Cochrane, C.-Y. C., Dubnicka, S. and Loughin, T. 2005. Comparison of methods for analyzing replicated preference tests. Journal of Sensory Studies, 20, 484–502.

Coetzee, H. 1996. The successful use of adapted paired preference, rating and hedonic methods for the evaluation of acceptability of maize meal produced in Malawi. Abstract, 3rd Sensometrics Meeting, June 19–21, 1996, Nantes, France, pp. 35.1–35.3

Coetzee, H. and Taylor, J. R. N. 1996. The use and adaptation of the paired-comparison method in the sensory evaluation of Hamburger-type patties by illiterate/semi-literate consumers. Food Quality and Preference, 7, 81–85.

Ennis, D. M. 2008. Tables for parity testing. Journal of Sensory Studies, 32, 80–91.

Ennis, D. M., and Ennis, J. M. 2009. Equivalence hypothesis testing. Food Quality and Preference, doi:10.1016/j.foodqual.2009.06.005.

Engen, T. 1974. Method and theory in the study of odor preferences. In: A. Turk, J. W. Johnson, Jr. and D. G. Moulton (Eds.), Human Responses to Environmental Odors. Academic, New York, pp. 121–141.

Ferris, G. E. 1958. The k-visit method of consumer testing. Biometrics, 14, 39–49.

Finn, A. and Louviere, J. J. 1992. Determining the appropriate response to evidence of public concern: The case of food safety. Journal of Public Policy and Marketing, 11, 12–25.

Filipello, F. 1957. Organoleptic wine-quality evaluation. 1. Standards of quality and scoring vs. rating scales. Food Technology, 11, 47–51.

Gacula, M. C. and Singh, J. 1984. Statistical Methods in Food and Consumer Research. Academic, Orlando, FL.

Gacula, M., Singh, J., Bi, J. and Altan, S. 2009. Statistical Methods in Food and Consumer Research. Elsevier/Academic, Amsterdam.

Garnatz, G. 1952. Consumer acceptance testing at the Kroger food foundation. In: Proceeding of the Research Conference of the American Meat Institute, Chicago, IL, pp. 67–72.

Gridgeman, N. T. 1959. Pair comparison, with and without ties. Biometrics, 15, 382–388.

Harker, F. R., Amos, R. L., White, A., Petley, M. B. and Wohlers, M. 2008. Flavor differences in heterogeneous foods can be detected using repeated measures of consumer preferences. Journal of Sensory Studies, 23, 52–64.

Hein, K. A., Jaeger, S. R., Carr, B. T. and Delahunty, C. M. 2008. Comparison of five common acceptance and preference methods. Food Quality and Preference, 19, 651–661.

Jaeger, S. R. and Cardello, A. V. 2009. Direct and indirect hedonic scaling methods: A comparison of the labeled affective magnitude (LAM) scale and best-worst scaling. Food Quality and Preference, 20, 249–258.

Jaeger, S. R., Jørgensen, A. S., AAslying, M. D. and Bredie, W. L. P. 2008. Best-worst scaling: An introduction and initial comparison with monadic rating for preference elicitation with food products. Food Quality and Preference, 19, 579–588.

Jellinek, G. 1964. Introduction to and critical review of modern methods of sensory analysis (odour, taste and flavour evaluation) with special emphasis on descriptive sensory analysis (flavour profile method). Journal of Nutrition and Dietetics, 1, 219–260.

Kim, H. S., Lee, H. S., O'Mahony, M. and Kim, K. O. 2008. Paired preference tests using placebo pairs and different response options for chips, orange juices and cookies. Journal of Sensory Studies, 23, 417–438.

Kimmel, S. A., Sigman-Grant, M. and Guinard, J.-X. 1994. Sensory testing with young children. Food Technology, 48(3), 92–94, 96–99.

Koster, E. P., Couronne, T. Leon, F., Levy, C. and Marcelino, A. S. (2003) Repeatability in hedonic sensory measurement: A conceptual exploration. Food Quality and Preference, 14, 165–176.

Lucas, F. and Bellisle, F. 1987. The measurement of food preferences in humans: Do taste and spit tests predict consumption? Physiology and Behavior, 39, 739–743.

Marchisano, C., Lim, J., Cho, H. S., Suh, D. S., Jeon, S. Y., Kim, K. O. and O'Mahony, M. 2003. Consumers report preference when they should not: A cross-cultural study. Journal of Sensory Studies, 18, 487–516.

Meyners, M. 2007. Easy and powerful analysis of replicated paired preference tests using the c2 test. Food Quality and Preference, 18, 938–948.

Moskowitz, H. R. 1983. Product Testing and Sensory Evaluation of Foods. Marketing and R&D Approaches. Food and Nutrition, Westport, CT.

Mueller, S., Francis, I. L. and Lockshin, L. 2009. Comparison of best-worst and hedonic scaling for the measurement of consumer wine preferences. Australian Journal of Grape and Wine Research, 15, 1–11.

Newell, G. J. and MacFarlane, J. D. 1987. Expanded tables for multiple comparison procedures in the analysis of ranked data. Journal of Food Science, 52, 1721–1725.

Odesky, S. H. 1967. Handling the neutral vote in paired comparison product testing. Journal of Marketing Research, 4, 199–201.

Prescott, J., Norris, L., Kunst, M. and Kim, S. 2005. Estimating a "consumer rejection threshold" for cork taint in white wine. Food Quality and Preference, 18, 345–349.

Roessler, E. B., Pangborn, R. M., Sidel, J. L. and Stone, H. 1978. Expanded statistical tables for estimating significance in paired-preference, paired difference, duo-trio and triangle tests. Journal of Food Science, 43, 940–941.

Quesenberry, C. P. and Hurst, D. C. 1964. Large sample simultaneous confidence intervals for multinomial proportions. Technometrics, 6, 191–195.

Saliba, A. J., Bullock, J. and Hardie, W. J. 2009. Consumer rejection threshold for 1,8 cineole (eucalyptol) in Australian red wine. Food Quality and Preference, 20, 500–504.

Scheffe' H. 1952. On analysis of variance for paired comparisons. Journal of the American Statistical Association, 47, 381–400.

Schmidt, H. J. and Beauchamp, G. K. 1988. Adult-like odor preference and aversions in three-year-old children. Child Development, 59, 1136–1143.

Schraidt, M. F. 1991. Testing with children: Getting reliable information from kids. ASTM Standardization News, March 1991, 42–45.

Schutz, H. G. 1971. Sources of invalidity in the sensory evaluation of foods. Food Technology, 25, 53–57.

Sidel, J. L., Stone, H. and Bloomquist, J. 1981. Use and misuse of sensory evaluation in research and quality control. Journal of Dairy Science, 64, 2296–2302.

Sidel, J., Stone, H., Woolsey, A. and Mecredy, J. M. 1972. Correlation between hedonic ratings and consumption of beer. Journal of Food Science, 37, 335.

Stone, H. and Sidel, J. L. 1978. Computing exact probabilities in discrimination tests. Journal of Food Science, 43, 1028–1029.

Tepper, B. J., Shaffer, S. E. and Shearer, C. M. 1994. Sensory perception of fat in common foods using two scaling methods. Food Quality and Preference, 5, 245–252.

Tuorila, H., Hyvonen, L. and Vainio, L. 1994. Pleasantness of cookies, juice and their combinations rated in brief taste tests and following ad libitum consumption. Journal of Sensory Studies, 9, 205–216.

Van Trijp, H. C. M. and Schifferstein, H. N. J. 1995. Sensory analysis in marketing practice: Comparison and integration, Journal of Sensory Studies 10, 127–147.

Vickers, Z. and Mullan, L. 1997. Liking and consumption of fat free and full fat cheese. Food Quality and Preference, 8, 91–95.