Chapter 17

Quality Control and Shelf-Life (Stability) Testing

Abstract Two routine functions of a sensory department may be quality control testing and the measurement of product stability or shelf life. These activities may involve any of the three main kinds of sensory testing or modifications of them. However, there are unique constraints for these tests, different types of analyses, and specific models for these data. This chapter discusses different procedures for sensory quality control, presents a recommended procedure, and outlines the programmatic requirements for establishing and maintaining a sensory QC function. The second section of the chapter presents an introduction to shelf-life testing, its special considerations, and some of the models used for stability testing data.

Consumer researchers are well aware of the quality of products. The food industry constantly faces the demand to maintain both quality and profitability simultaneously. Quality, however, is an elusive concept and as such must be operationalized and measured in order for it to be maintained.
—H. R. Moskowitz (1995)

Contents

H.T. Lawless, H. Heymann, *Sensory Evaluation of Food*, Food Science Text Series,
DOI 10.1007/978-1-4419-6488-5_17, © Springer Science+Business Media, LLC 2010

17.1 Introduction: Objectives and Challenges

Product quality has been defined in a variety of different ways (Lawless, 1995). Most sensory researchers focus on issues of consumer satisfaction as a measure of quality (Cardello, 1995; Moskowitz, 1995) although there is an historic tradition of using expert judges, commodity graders, or government inspectors to be the arbiters of product quality (Bodyfelt et al., 1988; York, 1995). This tradition is tied to use of the senses for detection of well-known defects or expected problem areas. The approach was well suited to standard commodities where minimum levels of quality could be insured, but excellence was rarely the issue. Another strong tradition has been the emphasis on conformance to specifications (Muñoz et al., 1992). This approach is useful in the manufacturing of durable goods whose attributes and performance could be measured using instrumental or objective means. Another popular definition of quality has been fitness for use (Lawless, 1995). This definition recognizes that quality does not exist in a vacuum, but only in a context or frame of reference for the consumer. Finally, the reliability or consistency in sensory and performance experiences with a product has been recognized as an important feature of product quality. Consumer expectations arise out of experience, and maintaining the constancy of that experience does a lot to build consumer confidence.

There are a number of challenges and problems that face a sensory evaluation program when trying to provide sensory information for quality control (QC). Difficult situations occur in the manufacturing environment where sensory assessment is needed during the processing itself. Such online sensory quality testing is likely to be done under tight time constraints, for example, while the product is cooling and before a decision is made to bottle or pack a production run. Only a few qualified judges may be available on third shift in the middle of the night when these decisions have to be made. There is little luxury involved in terms of time, and a detailed descriptive evaluation and statistical analysis may not be possible due to time and resource constraints. At the same time, a flexible and comprehensive system may be desired, one that is also applicable to raw materials testing, finished products, packaging materials, and shelf-life tests (Reece, 1979). Such constraints and demands often entail compromises in sensory practices.

A basic requirement of any sensory QC system is the definition of standards or tolerance limits on a sensory basis for the product. This requires calibration studies. If the sensory QC program is new, management may be surprised to learn that some research needs to be done before the QC panel can be trained and begin to operate. Sometimes the identification of standard products and tolerance limits may incur more expense than the sensory panel operation itself, especially if consumers are used to define the limits of what is acceptable quality. Maintaining reference standards for a standard quality product may also present difficulties. Foods and consumer products may have short shelf lives, and even with optimal storage conditions the standards will need to be replaced sooner or later. It is difficult to prevent some drift in the product over time. Multiple references including both optimally stored and fresh products may be needed (Wolfe, 1979). Some products simply change with age and this is a desirable feature like the proteolysis in ham or in cheese ripening (Dethmers, 1979). Furthermore, the frame of reference of the panel can drift or change seasonally. This makes it difficult to insure that a sensory specification of a standard product is in fact the same as the last standard.

Other barriers to acceptance involve the different ways that sensory evaluation is performed as opposed to traditional quality control. Most sensory tests are designed to look at a few or limited number of products. Sometimes the products are even considered to be identical, as in a homogeneous product evaluated from the same batch, like a well-mixed tank-produced beverage. The major source of variability in this sensory test is in the measuring instruments, the panelists. Statistical tests are designed to look at mean scores against the background of variation among people. This is quite different from the usual operation of quality control, where many samples of the product are taken and measured only once or a very few times on an instrument. The variability measured by traditional QC and pictured in control charts and other plots is across products, not instruments. Sensory QC has to deal with both sources of variation. In the instrumental measures, one can sample hundreds of products and take one measurement on each. In sensory QC, there may be one sample of each product but multiple measurements across panelists.

17.2 A Quick Look at Traditional Quality Control

Traditional quality control involves three major requirements: the establishment of specifications, the establishment of tolerance limits, and a sampling plan appropriate to the product being manufactured or the system being monitored. By specification, we mean the characteristics of the ideal or average product. To set tolerance limits, the liabilities of Type I error, rejecting product that is acceptable and therefore incurring unnecessary cost, must be weighed against the potential for Type II error, letting bad product into the marketplace and offending loyal consumers. This is a management decision that will impact the nature of the tolerance limits that are set. These tolerance limits are the levels of variability and/or the ranges that are deemed acceptable (in specification) versus unacceptable (out of specification).

Historically, such analysis was born in the advent of statistical quality control. W.A. Shewhart, working with Bell Labs in the early 1900s, noticed that there was variability in the functioning of some signal transmission components and that since these were often buried, they were a problem to dig up and repair. Failures or severe problems needed to be differentiated from the normal expected variability in these systems. To address this, he coined the ideas of assignable cause versus chance cause variation. The idea was that some variation was to be expected, but when the observation was outside of some common range, it was likely that some other cause was at work, and the item would need to be replaced, repaired, or otherwise dealt with. Thus the approach was statistical and involved a number of charts or graphs depicting this variation and the limits at which an assignable cause might be suspected. His notions of statistical quality control were later adopted by W.E. Deming in support of the war effort and later in the reconstruction of Japanese industrial practices.

There are several common types of charts used in statistical quality control and the sensory evaluation specialist should be familiar with them as they are part of the common language used by traditional quality control departments. Three kinds of control charts are common: *X*-bar charts, *R* charts, and *I* charts. Various rules exist for warning levels and action levels using these charts. Warning levels generally mean that the process needs to be investigated but no change is necessary. If action levels are surpassed, then there is good evidence for an out-of-control situation or assignable cause, and the process must be changed.

The *X*-bar chart plots the mean scores for different test batches over some time period. Typically three to five products are pulled for evaluation (Muñoz et al., 1992). Upper and lower confidence limits (UCL and LCL) are generally set by ± 3 standard errors (sometimes referred to as 3-sigma) as shown in Fig. 17.1. Out-of-control conditions are spotted by a number of criteria, such as a point beyond the 3-sigma UCL or LCL, or some pattern of points such as "nine points in a row all on one side of the historical mean" or "six points in a row all increasing or decreasing" (Nelson, 1984). The *R*-chart measures the range of observations in any batch of product. Upper and lower limits are set as in the *X*-bar charts with warning and action levels at 2-sigma and 3-sigma (or 95 and 99% confidence levels). Sometimes the *X*-bar chart and the *R* chart may be combined to give a fuller picture. If only one unit is evaluated per batch, then the mean and range cannot be used, only the observed value itself. This is plotted on an I chart (Muñoz et al., 1992). As in the *X* and *R* charts, a mean and confidence limits can be set, as well as warning levels and action levels.

17.3 Methods for Sensory QC

17.3.1 Cuttings: A Bad Example

Muñoz and coauthors (1992) give both good and bad examples of applications of sensory QC procedures. Here is an example of a poor implementation of the in/out procedure:

> The panel consists of a small group of company employees (4 to 5) mainly from the management team. The panel evaluates a large amount of production samples (up to 20 to 40) per session without standardized and controlled protocols. Each product is discussed to determine if it is to be considered "in" or "out" of specifications. In this program, no defined specifications or guidelines for product evaluation exist, and no training or product orientation was held. As a result, each panelist makes decisions based on his or her individual experience and familiarity with production, or based on the highest ranking person on the panel. (p. 141)

This scenario highlights some of the pitfalls of a pass/fail procedure. It resembles a common daily

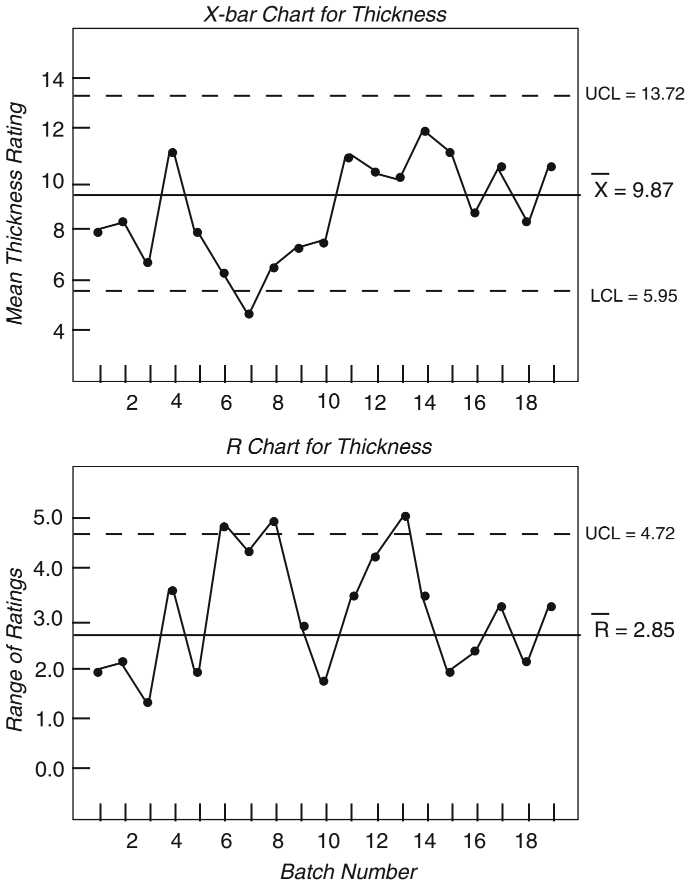


Fig. 17.1 *X*-bar and *R* charts for a hypothetical analysis of product thickness ratings over several batches. The *X*-bar chart shows the historic mean and upper and lower control limits, usually set at three standard deviations. The *X*-bar chart shows one batch with a mean value below the lower limits and the batches on either side show a range beyond the range limit as well. This should suggest action and/or investigation by process control personnel. Batch 13 was also above the range limit suggesting an out-of-control situation. Batches 5–10 also show the alarm pattern of six points on one side of the historic mean.

check on production that was often done by a convened committee of technical personnel and managers, called “cuttings.” Without the guidance of a sensory evaluation specialist, such a method can be put in place with a number of poor practices, such as having an open discussion to reach consensus and determining a final score.

17.3.2 In–Out (Pass/Fail) System

Muñoz et al. (1992) discuss four different approaches to sensory quality assessment. Their book, *Sensory Evaluation in Quality Control*, gives a detailed treatment of each. One of the methods is the in/out or pass–fail procedure. This method differentiates normal production from products that are considered different or outside specifications. It is a popular procedure at the plant level and is used in some binary decision-making scenarios such as Canadian fish inspection (York, 1995).

Panelists are trained to recognize the characteristics that defined “out-of-spec” products as well as the range of characteristics that are considered “in spec” (Nakayama and Wessman, 1979). This enhances the uniformity of criteria among the panelists. As in any yes/no procedure, the effects of bias and criterion setting can be as influential as the actual sensory experience (see Section 5.8). Different panelists may be more or less conservative in the degree of sensory difference they require in order to call something out of spec. In quality control, the liability of differences in criterion setting is high, since there are always pressures to pass poor products to maintain productivity. Obviously, the presentation of blind control samples is necessary to estimate a false alarm rate (false positives), and the introduction of purposely defective samples can be useful in estimating the false-negative (miss) rate.

Muñoz and coauthors stressed the need for standardized protocols for sample handling and evaluation and the need for independent judgments, rather than discussion and consensus. York (1995) described how government fish inspectors are involved in standards development workshops. Training includes definition of sensory characteristics that define wholesomeness, taint, and decomposition and how these characteristics at different levels contribute to the binary decision for acceptance or rejection.

The major advantages of the in/out procedure are its apparent simplicity and use as a decision-making tool. It is especially suited to simple products or those with a few variable attributes. The disadvantages include the criterion-setting problems described above. Also, the method does not necessarily provide diagnostic reasons for rejection or failure, so there is a lack of direction to be used in fixing problems. It may also be difficult to relate these data to other measures such as microbial or instrumental analyses of food quality. The data necessarily consist of mere frequency counts of the number of panelists judging the product out of spec. Finally, it may also be difficult for some panelists to be analytical and look for specific problems and defects, while at the same time providing an overall integrated judgment of product quality.

17.3.3 Difference from Control Ratings

A second major approach to sensory quality control is to use ratings for an overall degree of difference from a standard or control product. This works well if it is feasible to maintain a constant "gold standard" product for comparison (Muñoz et al., 1992). It is also well suited to products where there is a single sensory characteristic or just a few sensory characteristics that vary in production. The procedure uses a single scale as illustrated in the paper by Aust et al. (1985) such as the following:

extremely different from the standard	the same as the standard

Ratings on this scale may be transcribed from zero (rightmost point) to ten (leftmost point). For purposes of rapid analysis, a simple 10-point category scale can be used. Additional points along the scale are sometimes labeled with other verbal descriptions of different degrees of difference.

Training with a range of references and establishing the nature and conditions for reproducing the control sample are critical in this procedure. The panelists must be shown samples in training that represent points along the scale. These can be cross-referenced to consumer opinion or chosen by management tastings (Muñoz et al., 1992). Preferably there is some consumer input for calibration at an early stage of the program development. Muñoz and coauthors also presented a more descriptive version where differences from control on several individual attributes of a flaked breakfast cereal were evaluated. This more detailed procedure can provide more actionable information about the attributes responsible for any differences. If just a single scale is used, panelists may weight attributes differently in determining an overall degree of difference. Specific characteristics may be more or less influential for a given panelist.

Management should choose some level of the difference as a cutoff for action. The scale is useful in that it provides for a range of differences that are acceptable. At some point, regular users of the product will notice and object to differences, and this should be the benchmark for action standards. If at all possible, the panelists should *not* be informed of where the breakpoint in decision making occurs along the scale. If they know where management sets the cutoff, they may become too cautious and tend to give scores that may approach but not surpass the cutoff (Rutenbeck, 1985).

As with the pass/fail method, an important part of this procedure is the introduction of blind control samples. During every test session, a blind-labeled sample of the standard should be inserted into the test set, to be compared against the labeled version of itself. This can help establish the baseline of variation on the scale, since two products are rarely rated as identical. Another way to think about this is that it provides a false alarm rate or an estimate of a placebo effect (Muñoz et al., 1992). In the original paper by Aust et al. (1985), an additional control sample was a product from a different batch of the same production. Thus a test product's variability could be measured against the response bias or variation within the ratings of the standard against itself, as well as the batch-to-batch variation. This approach could also be useful in comparing products from different manufacturing sites. Aust et al. proposed an analysis of variance model for this design. If the control comparison is simply the standard against itself, a paired or dependent *t*-test can compare the mean scores for the test product against the mean score of the standard product rated against itself. This presumes that there are sufficient judges to warrant a statistical test. With small panels, more qualitative criteria for action have to be adopted, e.g., any three out of five panelists below some cutoff score. The false alarm level should also be considered in making decisions. The ratings for the blind standard against itself must be low relative to the test product scores in order to reject a batch.

The major disadvantage of this test, like the yes/no procedure, is that it does not necessarily provide any diagnostic information on the reasons for the difference if only the single scale is used. Of course, open-ended reasons for difference can be given, or additional questions, scales, or checklists can be provided for attributes that are common problems or show common variation.

17.3.4 Quality Ratings with Diagnostics

A third method, similar to the overall difference from control method, is to use quality ratings. This entails an even more complex judgment procedure on the part of the panelists, since it is not only the differences that matter but also how they are weighted in determining product quality. This idea of an integrated quality score is part of the tradition of food commodity judging, as discussed in Section 17.6.

There are three main abilities of the trained or expert judge that are necessary in order to use a quality judging system. The expert judge must maintain a mental standard of what the ideal product is in terms of sensory characteristics. Second, the judge must learn to anticipate and identify common defects that arise as a function of poor ingredients, poor handling or production practices, microbial problems, storage abuse, and so on. Finally, the judge needs to know the weight or influence of each defect at different levels of severity and how they detract from overall quality. This usually takes the form of a point deduction scheme. In the case of seafood, deterioration as a function of aging or mishandling will go through a sequence of flavor changes and sensory spoilage characteristics. These changes in sensory characteristics can be translated into a scale for fish quality (Regenstein, 1983).

The common characteristics of quality ratings are these (Muñoz et al., 1992): Scales directly represent the quality judgment, rather than just sensory difference, and can use words like poor to excellent. This wording itself can be a motivator, as it gives the impression to panelists that they are directly involved in decision making. Quality grading works best when there is management or industry consensus on what is good. In some cases specific product characteristics can be rated in addition to overall quality, for example, quality of texture, flavor, appearance. In some schemes like wine judging, quality scores for individual attributes are then summed to give the overall score (Amerine and Roessler, 1981). Unfortunately, the quality scoring approach is prone to abuse, where small numbers of poorly trained judges evaluate dozens of product "cuttings," use their own personal criteria, and use consensus (discussion) methods to make decisions. Muñoz and coauthors presented this example of good practice:

> The panel consists of 8 to 12 panelists who are trained in the procedures to assess the quality of a given product type. They learned the company's quality guidelines, which were established using the input from consumers and management. . . . These guidelines are shown to panelists by actual products representing various quality levels. The program was designed by a sensory professional using sound methodology and adequate testing controls for the evaluation process. In routine evaluations, panelists rate "overall" quality as well as the quality of selected attributes using a balanced scale (very poor to excellent). The data are treated like interval data and panel means are used to summarize the results of the evaluations. The results are provided to management, which makes decisions on the disposition of the production batches evaluated. (p. 109)

Although there are apparent time and cost advantages in this direct approach, there are also disadvantages. The ability to recognize all the defects and integrate them into a quality score may require a lengthy training process. There is a liability that individual subjectivity in likes and dislikes can creep into the judge's evaluations. The specialized vocabulary of technical defects may seem arcane to non-technical managers. Finally, with small panels, statistical difference tests are rarely applied to such data, so the method is primarily qualitative.

17.3.5 Descriptive Analysis

A fourth approach to sensory quality control is a descriptive analysis method, as described in Chapter 10. The goal is to provide intensity ratings for individual sensory attributes by a trained panel. The focus is on the perceived intensity of single attributes and not quality or overall difference. Intensity rating of single sensory characteristics demands an analytical frame of mind and focused attention on dissecting the sensory experience into its component parts. Muñoz and coauthors called this as a "comprehensive descriptive method," but they do allow for limitation of the scorecard to a small set of critical attributes. For QC purposes, attention to a few critical attributes may be appropriate.

As in the other techniques, calibration must be done. Specifications for the descriptive profile must be set via consumer testing and/or management input. This will consist of a range of allowable intensity scores for the key attributes. Table 17.1 shows an example of a descriptive evaluation of potato chip samples and the range of sensory specifications, as previously determined in a calibration study with consumers and/or management input (from Muñoz et al., 1992). This sample is below the acceptable specification limits in evenness of color and is too high in cardboard flavor, a characteristic of lipid oxidation problems.

Table 17.1 Evaluation of potato chip samples using descriptive specification

	Mean panel score	Acceptable range
Appearance		
Color intensity	4.7	3.5–6.0
Even color	4.8	6.0–12.0
Even size	4.1	4.0–8.5
Flavor		
Fried potato	3.6	3.0–5.0
Cardboard	5.0	0.0–1.5
Painty	0.0	0.0–1.0
Salty	12.3	8.0–12.5
Texture		
Hardness	7.5	6.0–9.5
Crisp/crunch	13.1	10.0–15.0
Denseness	7.4	7.4–10.0

From Muñoz et al. (1992)

Descriptive analysis requires extensive panel training. Panelists should be shown reference standards to learn the meaning of the key attributes. Next they must be shown intensity standards to anchor their quantitative ratings on the intensity scale. They do not have to be shown examples that are labeled as "in specification" or "out of specification," however, since that decision is based on the overall profile of the product and is done by the sensory panel leader or QC management. Defective samples can be used in training intensity ratings, but the actual cutoff points are better kept in confidence by managers making the decisions about product disposition (Muñoz et al., 1992). This will avoid the tendency for panelists to gravitate toward scores that are just within the acceptable range.

Advantages. The detail and quantitative nature of the descriptive specification lends itself well to correlation with other measures such as instrumental analysis. The second advantage is that it presents less of a cognitive burden on panelists, once they have adopted the analytical frame of mind. They are not required to integrate their various sensory experiences into an overall score, but merely report their intensity perceptions of the key attributes. Finally, the reasons for defects and corrective actions are easier to infer since specific characteristics are rated. These can be more closely associated with ingredients and process factors than an overall quality score.

Limitations. Because the method depends on good intensity anchoring, it tends to be more laborious in panel training than some of the other techniques. Due to the need for data handling and statistical analysis, as well as having a sufficient number of trained judges on hand for the method, it is better suited for quality evaluation of finished products. It may be difficult to arrange for descriptive evaluation for ongoing production, particularly on later work shifts if production is around the clock. The training regimen is difficult and time consuming to set up since examples must be found for the range of intensities for each sensory attribute in the evaluation. This can require a lot of technician time in sample preparation. Another liability is that problems may occur in some attributes that were not included on the scorecard and/or outside of the training set. Thus, the method lends itself to situations where the problem areas are well known and the production and ingredient variability can be easily reproduced to make up the training set.

17.3.6 A Hybrid Approach: Quality Ratings with Diagnostics

Gillette and Beckley proposed a reasonable compromise between the quality rating method and a comprehensive descriptive approach at the 1992 meeting of the Institute of Food Technologists (described in Beckley and Kroll, 1996). The centerpiece of this procedure is a scale for overall quality. The quality scale is accompanied by a group of diagnostic scales for individual attributes. These attributes are key sensory components that are known to vary in production. Muñoz et al. (1992, pp. 138–139) describe a similar modification of the overall quality ratings method to include the collection of descriptive information on key attributes. In the method of Gillette and Beckley the main scale takes this form:

1	2	3	4	5	6	7	8	9	10
Reject		Unacceptable			Acceptable			Match	

On this scale, a product that is so clearly deficient to call for immediate disposal gets a score of 1 or 2. Products that are unacceptable to ship but might be reworked or blended get a score in the range of 3–5. If evaluation is online during processing, these batches would not be filled into retail containers or packaged but would be held for rework or blending. If the samples are different from the standard but in an acceptable range, they receive scores of 6–8, and samples that are a near match or considered identical to the standard receive a 9 or 10, respectively. According to Muñoz et al. (1992), the use of the terms "acceptable" and "unacceptable" here is unfortunate, for it gives the panelists an impression of the action standards for products passing or failing and the feeling that they are responsible for decisions about product disposition. This creates a tendency to use the middle to the upper end of the scale, to avoid grading products as unacceptable (Rutenbeck, 1985).

The advantages of this method are its outward simplicity in using an overall rating and the addition of attribute scales to supply reasons for product rejection. The method also recognizes that there are situations where products will not match the gold standard exactly, but still are acceptable to ship. As in the other procedures, the boundaries for out-of-spec product and the selection of a gold standard must be undertaken before training, preferably in a consumer study, but at least with management input. These defined samples must then be shown to subjects to establish concept boundaries. In other words, tolerance ranges must be shown to panelists (Nakayama and Wessman, 1979).

17.3.7 The Multiple Standards Difference Test

Amerine et al. (1965) mentioned a variation on simple difference tests that would include a non-uniform or variable standard. This has come to be known as the multiple standards difference test. Although there is scant literature on the procedure, it has apparently enjoyed some popularity. The idea is to give a forced-choice test in which participants pick which one of several alternative products is the most different from the rest of the set. The simplest approach is to have one test product and K alternative versions of the standard product. Rather than representing identical versions of a gold standard, the standards are now chosen to represent the acceptable range of production variability. The choice of standards to represent the range of acceptable variation is critical to the success of this approach. Historically, this method resembles Torgerson's "method of triads" of which the triangle test is a special case (Ennis et al., 1988). Pecore et al. (2006) and Young et al. (2008) used a similar approach, except that overall degree of difference rating was used (discussed below). The choice of products to be included in the set of acceptable standards is critical. If they do not reasonably bracket the range of acceptable variation, then the test will be too sensitive (if the range is small) or too insensitive to detecting bad samples (if their range is too large).

If there are a large number of testers ($N = 25$ or more), as in a discrimination procedure, the z-score approximation to the binomial distribution may be used for hypothesis testing. The approximation is

$$z = \frac{(P - 1/k) - (1/2N)}{\sqrt{(1/k)(1 - 1/k)/N}} \tag{17.1}$$

where k is the total number of alternatives (test product plus variable standards), P is the proportion selecting the test product as the outlier (the most different sample), and N is the number of judges. This is the same formula as in the triangle and other forced-choice procedures, except that k may be 4 or larger, depending on the number of references.

Although this method appears simple at first, there are a few concerns and potential pitfalls in its application. First, in many QC situations, it may not be feasible to conduct a difference test with sufficient numbers of judges to have a meaningful and sensitive application of the statistical significance test. Second, the failure to reject the null (failure to get a significant result) does not necessarily imply sensory equivalence. From a statistical perspective, it is difficult to have confidence in a "no-difference" result, unless the power of the test is very high. Statistical confidence in the equivalence decision can only be obtained after beta-risk is estimated against a suitable alternative hypothesis (see Appendix E on test power). One approach is to use the analysis for significant similarity, as outlined in Chapter 5. This will necessarily entail a larger number of judges ($N = 80$ or so). Finally, the tests for overall difference such as the triangle procedure are known to have high inherent variability, so the introduction of even more variability by multiple standards makes it very difficult to get a significant difference and reject product. This factor may contribute to a high level of beta-risk, i.e., the chance of missing a true difference.

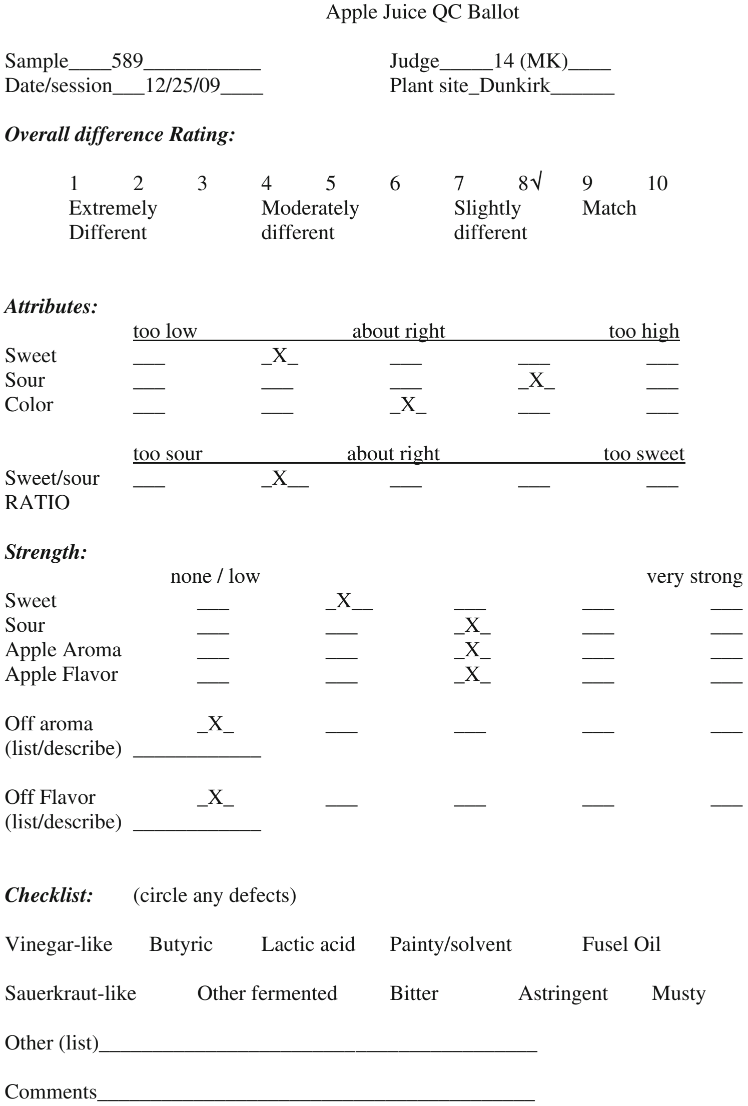
A similar approach to the multiple standards choice test was described by Pecore et al. (2006) and Young et al. (2008) but using degree of difference ratings, rather than a choice test. This approach was part of the original intent of the degree of difference test proposed by Aust et al. (1985). In this method, a test lot is compared to each of two control lots. The two control lots are also compared to each other. From these three pieces of data, three comparisons can be made: Each mean difference score of the test–control pair is compared to the mean difference score of the control–control pair. Also the *average* test–control rating is compared to the same baseline. Thus the within-control variation is taken into account in the comparisons, which must be significantly exceeded if the test lot is to be found different (and thus actionable). Of course, this kind of test requires a sufficient panel size to get a meaningful and statistically powerful test. Later, Young et al. (2008) extended the model and procedure to include two test lots as well as two control lots and used an incomplete block design to limit the six comparisons to three comparisons per panelist. The critical comparison in this case is between the average of the four means comparing tests to controls versus the average of the mean control–control score and the mean test–test lot score. So the baseline becomes the average difference within control lots and within test lots.

17.4 Recommended Procedure: Difference Scoring with Key Attribute Scales

This method is similar to the hybrid procedure of Gillette and Beckley except that it substitutes the overall difference scale for the quality scale. This avoids the problem that panelists may react to words like "reject" and avoid them. So the method is similar to that of Section 17.3.6. Some category or line version of the difference scale should be used such as the following:

1	2	3	4	5	6	7	8	9	10
Completely different		Very different			Somewhat different			Match	

The ballot should also include diagnostics on key attributes, those that will vary in production and are likely to cause consumer rejection. For attributes that can be too strong or too weak, using just-right scales is appropriate. Some defects may be a problem at higher levels, and intensity scales are useful for those attributes. Others may warrant product rejection at any level whatsoever, and a checklist can be provided for



Apple Juice QC Ballot

Sample____589___________ Judge_____14 (MK)____
Date/session___12/25/09____ Plant site_Dunkirk______

Overall difference Rating:

1	2	3	4	5	6	7	8√	9	10
Extremely Different			Moderately different			Slightly different		Match	

Attributes:

	too low		about right		too high
Sweet	___	_X_	___	___	___
Sour	___	___	___	_X_	___
Color	___	___	_X_	___	___

	too sour		about right		too sweet
Sweet/sour RATIO	___	_X__	___	___	___

Strength:

	none / low				very strong
Sweet	___	_X__	___	___	___
Sour	___	___	_X_	___	___
Apple Aroma	___	___	_X_	___	___
Apple Flavor	___	___	_X_	___	___
Off aroma (list/describe) ____________	_X_	___	___	___	___
Off Flavor (list/describe) ____________	_X_	___	___	___	___

Checklist: (circle any defects)

Vinegar-like Butyric Lactic acid Painty/solvent Fusel Oil

Sauerkraut-like Other fermented Bitter Astringent Musty

Other (list)____

Comments____

Fig. 17.2 A sample ballot for apple juice, using the recommended procedure of degree-of-difference scale plus diagnostic attribute ratings. Note that some attributes use the just-about-right scale while others are better suited to a simple intensity scale. For off-flavors or defects that are objectionable at any level, a simple checklist is useful. Note that this method must be used with a well-trained panel.

those more serious faults. A sample ballot for apple juice is shown in Fig. 17.2.

Screening panelists for sensory acuity and a good training regimen are key here, as in setting up other quality control panels. The screening procedure should use the types of products that people are going to eventually judge and insure that they can discriminate among common levels of ingredients like sugar or acid content and process variables like heating times or processing temperatures. A sample screening procedure for an apple juice panel is shown in Appendix 1 of this chapter. Screening should involve a number of attributes and, if possible, different tasks or tests (Bressan and Behling, 1977). The top performers can be invited for panel training and others who score well can be kept on file for future replacements as panelist attrition occurs (plan from day 1!). Ideally the screening pool of volunteers should be two to three times the desired panel size. Supervisory approval is a key to good attendance and participation.

After screening, training may take six to ten sessions depending on the complexity of the product. Gross differences are illustrated early in training and smaller differences shown as training progresses. The goal is to solidify the conceptual structure of the panelists so they know the category boundaries for the quality ratings and the expected levels of the sensory attributes. Panelists must also come to recognize how off-flavors, poor texture, or appearance problems factor into their overall score.

With small panels, there is no statistical analysis, but rules of thumb must be established for taking actions. It is difficult to apply mean ratings to less than about eight panelists. Since there are differences in individual sensory ability, poor ratings by just a few individuals may be indicative of potential problems. Thus action criteria should take into account negative minority opinions and weight them more heavily than high outliers or a few panelists who thought the product matched the standard (but who may have missed some important difference). For example, if two panelists rate the sample on scale point two, but the rest give it six, seven, or eight, the mean score could be in the acceptable range in spite of the two panelists who spotted potentially important problems. The panel leader should take note of the two low values and at least call for retesting of this questionable sample. Of course, consistent patterns of disagreement between panelists are a hint that some retraining may be needed.

17.5 The Importance of Good Practice

In all small-panel sensory assessments, the general principles of good testing become especially important since there are shortcuts often taken in these procedures that are not part of standard sensory evaluation practices. Most notably, quality assessment may not entail any statistical analysis, due to the small numbers of panelists. Statistical methods provide some insurance against false alarms due to random variation or errors of missing important differences. Without the aid of statistical analysis, other safeguards for insuring the quality of the information take on an even higher level of importance.

It is worth considering an example concerning pork inspection since it illustrates many of the pitfalls involved in small-panel experiments. This was a study of boar taint or sex aroma from androstenone, a problem odor in the fatty tissues of adult male swine. The goal was to correlate sensory panel scores for this taint with instrumental measures of androstenone content (Thompson and Pearson, 1977). Two sensory analyses were done. In the first, three to five panelists sampled boar taint aroma in the packinghouse using a hot-iron technique to elicit the aroma and came to a consensus judgment using a 6-point scale for odor intensity. A second evaluation was done after the samples were sent to the laboratory for instrumental analysis. In this case, three panelists were screened for sensitivity to androstenone, and means were calculated from a 9-point scale for odor intensity. Evaluations were performed in a laboratory exhaust hood and preparation procedures were standardized. The correlation with instrumental measures was +0.27 for the first evaluation (not significantly different from zero correlation) and +0.40 (statistically significant) in the second. The increased correlation could be due to a number of methodological factors that were improvements in the second evaluation. These include (1) better evaluation location (fume hood versus packing house), (2) screening of judges, (3) constancy of panel members instead of people dropping in and out, (4) averaging scores versus a consensus procedure, and (5) a more standardized sample preparation method. Each of the shortcuts might have been introduced on some practical grounds, but their combined effect was to increase the error level in the data. This made a difference in the statistical significance and conclusions of the study regarding instrumental-sensory correlations.

Table 17.2 gives a number of guidelines for good sensory practice in quality evaluations and Table 17.3 gives guidelines for judges (adapted from Nelson and Trout, 1964). As in any other sensory test, product samples should be blind coded and presented in different random orders to each panelist. If production personnel are used in the panel who know the identities of some of the products pulled for evaluation, another technical person must blind code them and insert blind controls into the test set. The person who pulls the samples must not evaluate the samples. It is not reasonable to expect that person to be objective and discount any knowledge of the product identity. Serving temperature, volume, and any other details concerning product preparation and the tasting method should be standardized and controlled. Facilities should be odor free and distraction free. Evaluations should be made in a clean sensory testing environment with booths or separators, not on the benchtop of an analytical instrument lab or on the manufacturing floor (Nakayama and Wessman, 1979). Warm-up samples are useful. Blind replicates can be introduced to check judge consistency. Judges should taste a representative

Table 17.2 Ten guidelines for sensory quality testing

1. Establish standards for optimum quality (“gold standard”) target plus ranges of acceptable and unacceptable products
2. Standards should be calibrated by consumer testing if possible. Alternatively, experienced personnel may set standards but these should be checked against consumer opinion (users of product)
3. Judges must be trained, i.e., familiarized with standards and limits of acceptable variation
4. Unacceptable product standards should include all types of defects and deviations likely to occur from materials, processing, or packaging
5. Judges may be trained to give diagnostic information on defects, if standards are available typifying these problems. Scaled responses for intensity or checklists may be used
6. Data should always be gathered from at least several panelists. Ideally, statistically meaningful data should be gathered (ten or more observations per sample
7. Test procedures should follow rules of good sensory practice—blind testing, proper environment, test controls, random orders
8. Blind presentation of standards within each test should be used to check for judge’s accuracy. It is important to include a (blind) gold standard for reference purposes as well
9. Judge reliability may be tested by blind duplicates
10. Panel agreement is necessary. If unacceptable variation or disagreement occurs, re-training is warranted

Table 17.3 Guidelines for participation in sensory assessments

1. Be in correct physical and mental condition
2. Know the score card
3. Know the defects and the range of probable intensities
4. For some foods and beverages, it is useful to observe aroma immediately after opening the sample container
5. Taste a sufficient volume (Be professional—not timid!)
6. Pay attention to the sequence of flavors
7. Rinse, occasionally, as the situation and product type warrant
8. Concentrate. Think about your sensations and block out all other distractions
9. Do not be too critical. Also, do not gravitate to the middle of the scale
10. Do not change your mind. Often the first impression is valuable, especially for aromas
11. Check your scoring after the evaluation. Get feedback on how you are doing
12. Be honest with yourself. In the face of other opinions, “stick to your guns”
13. Practice. Experience and expertise come slowly. Be patient
14. Be professional. Avoid informal lab banter and ego trips Insist on proper experimental controls—watch out for benchtop “experiments”
15. Do not smoke, drink, or eat for at least 30 min before participation
16. Do not wear perfume, cologne, aftershave, etc. Avoid fragranced soaps and hand lotions

Modified from Nelson and Trout (1964)

portion (not end of batch or other anomalous parts of production).

Other rules of thumb apply to the judges or panelists. Panelists should be screened, qualified, and motivated with suitable incentives. They must not be overtaxed or asked to test too many samples in one day. Rotation of the panel at regular intervals can improve motivation and relieve boredom. Judges should be in good physical condition, i.e., free from ailments like colds or allergies that would detract from their performance. They should not be mentally harried from other problems on the job when arriving for testing, but should be relaxed and able to concentrate on the task at hand. They must be trained to recognize the attributes, scoring levels and, of course, know the scorecard. Judgments should be independent without conferring, jury style. Discussion or feedback can be given later to provide for ongoing calibration. A special liability arises when manufacturing personnel are used who have a lot of pride in the product accompanied by false confidence in the infallibility of the manufacturing process. Such panelists may be unwilling to “rock the boat” and call attention to problem areas. Testing with blind out-of-spec samples, known defects, and other such “catch trials” accompanied by feedback when they pass defective samples can help counteract this overly positive attitude.

The data should consist of interval scale measurements where possible. If large panels are used (ten or more judges), statistical analysis is appropriate and data can be summarized by means and standard errors. If very small panels are used, the data should be treated qualitatively. Frequency counts of individual scores should be reported and considered in action standards. Deletion of outliers can be considered, but a few low scores (i.e., a minority opinion) may be indicative of an important problem, as noted above. Re-tasting may be warranted in situations where there is strong disagreement or high panel variability.

17.6 Historical Footnote: Expert Judges and Quality Scoring

17.6.1 Standardized Commodities

The food industry benefits from standardization of grades for foods that are minimally processed from raw ingredients, from a single source without multiple components, and closely associated with a single agricultural commodity. Such "food commodities" include many dairy products such as milk, cheese, and butter; fruits such as olives; some kinds of meat; and wine. Various industries and governments have established quality grades for food commodities or systems for scoring them based on two main factors: similarity to a product ideal and lack of defects. The value in such quality grading, scoring, or monitoring is the assurance to the consumer that the product will have the sensory properties that they have come to expect.

Sometimes these systems are defined by international organizations in order to provide standards of identity for the food commodity. An example is the International Olive Oil Council (COI). The COI provides written standards for sensory evaluation including definitions for the vocabulary of sensory properties and defects, a standardized scorecard, a point system for assigning grades or classifications, methods for panel training, certification of laboratories evaluating olive oil, and even specifications of the tasting glasses that are to be used in the evaluations. Their website provides all of this information in the various languages of olive oil producing countries (International Olive Oil Council, 2007).

Two further examples of commodity judging systems by trained or expert panels are shown below. The sensory specialist should search for such professional organizations and specifications if they are assigned to develop methods for such a food commodity. The methods are poorly suited to processed engineered foods that do not fall into the category of a standardized commodity, but they can provide a useful starting place for development of a quality monitoring system for a closely related product. The sensory specialist should be careful, however, not to force-fit a standardized grading scheme to a product that is substantially different. For example, the quality evaluation scheme for grading vanilla ice cream would be only poorly suited to sensory testing on a frozen yogurt product made from goat's milk.

17.6.2 Example 1: Dairy Product Judging

A longstanding tradition in the field of dairy products has been the quality grading schemes for assessing product defects and assigning overall quality scores. The American Dairy Science Association continues to hold a decades-old student judging competition, in which students and teams of students attempt to duplicate the quality scores of established experts. Various defective products are supplied, and students must be able to recognize the defect, subtract the appropriate penalty given the type and severity of the problem, and arrive at an overall score (Bodyfelt et al., 1988). The support for quality judging in dairy products is not universal, however. Some countries like New Zealand have replaced the overall quality judging method with ratings on specific key attributes for dairy product analysis.

However, these methods do persist and find some utility in small plant quality control and in government inspections (Bodyfelt et al., 1988; York, 1995). An example of the quality judging scheme for cottage cheese is shown in Table 17.4, listing defects and their point deduction values for slight, definite, and pronounced levels of sensory intensity. An extensive discussion of quality judging for dairy products can be found in "Sensory Evaluation of Dairy Products" by Bodyfelt et al. (1988).

Table 17.4 A point deduction scheme for cottage cheese quality grading. Cottage cheese scoring guide

	Slight	Distinct	Pronounced
Appearance (5 points maximum):			
Lacks cream	4	3	2
Shattered curd	4	3	2
Free cream	4	2	1
Free whey	4	2	1
Texture (5 points maximum):			
Weak/soft	4	3	2
Firm/rubbery	4	2	1
Mealy/grainy	4	2	1
Pasty	3	2	1
Gelatinous	3	2	1
Flavor (10 points maximum):			
High acid	9	7	5
High salt	9	8	7
Flat	9	8	7
Bitter	7	4	1
Diacetyl/coarse	9	7	6
Feed	9	7	5
Acetaldehyde/green	9	7	5
Lacks freshness	8	5	1
Malty	6	3	1
Oxidized	5	3	1
Fruity	5	3	1
Musty	5	3	1
Yeasty	4	2	1
Rancid	4	2	1

Rate the presence of each defect as slight, distinct, or pronounced. Give scores for appearance, texture, and flavor based on the table.

Other problems may include discoloration, matted curd, slimy texture, foreign flavors, unclean flavors (describe), and fermented flavors

Modified from Bodyfelt et al. (1988)

Such methods are poorly suited to food research where the processed or engineered food is not a standard commodity and/or when the sensory changes are not likely to be a set of predictable defects. In new food product development, it is not necessarily clear what consumers or segments of consumers may like, so assignment of quality scores based on some arcane or traditional knowledge of experts is not useful. The dairy judging methods have been repeatedly criticized for lack of applicability to research problems, violations of sensory evaluation principles, and problems in scaling and statistical analysis (Hammond et al., 1986; McBride and Hall, 1979; O'Mahony, 1981; Pangborn and Dunkley, 1964; Sidel et al., 1981). Furthermore, the opinions of expert judges and standard point deduction schemes may not correspond to consumer opinion as shown in Fig. 17.3. The oxidized defects in milk were viewed less critically on the average by consumers than the suggested ADSA scores would dictate. Of course, having a point deduction scheme that is more severe that the average consumer opinion provides a kind of safety net and insures that the most sensitive consumers will not be offended by poor products. The liability in a stringent "safety net" is that acceptable product batches will be rejected.

17.6.3 Example 2: Wine Scoring

Beyond manufacturing control and government inspection there are other situations in which the consuming public desires information on product quality. Rather than deducting points from some widely accepted standard, there are also products

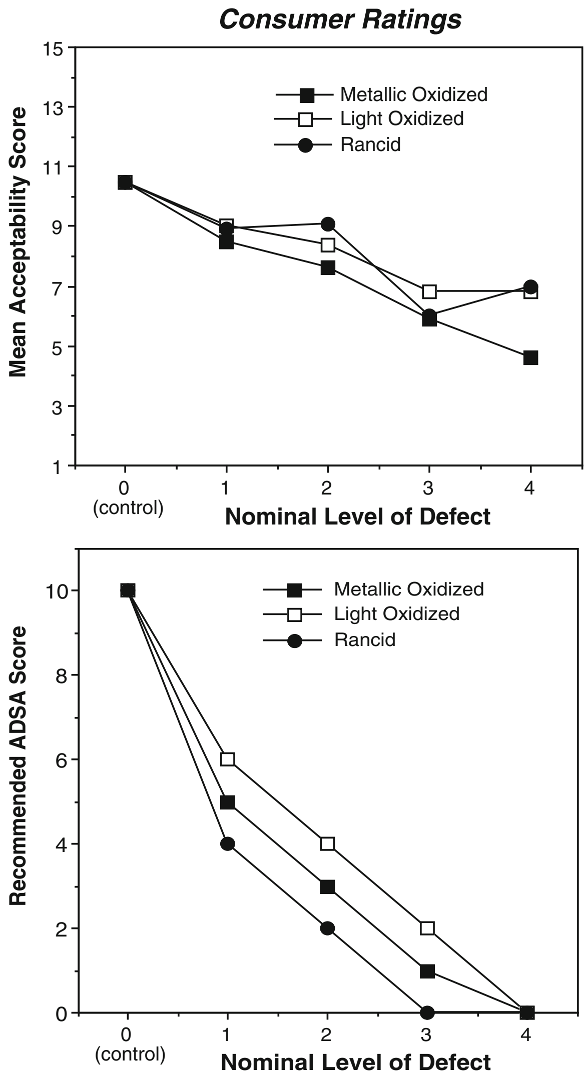


Fig. 17.3 Consumer ratings of abused milk samples compared to the recommended ADSA scores for those products based on the recipes and rating systems shown in Bodyfelt et al. (1988). From Lawless and Claassen (1993) with permission.

where excellence is recognized beyond the merely acceptable. Garvin (1987) remarked that considerations of quality should include the ability to please consumers, not just protect them from annoyances. This idea is further developed in the Kano model in Chapter 19. Some products show a wide range of better-than-acceptable variation. Wines are a good example. What kind of quality measurement system can go beyond point deductions for defects to provide degrees of difference on the positive end of the quality continuum?

An early method for wine quality assessment was the 20-point rating system developed at the University of California at Davis (Amerine and Roessler, 1981; Ough and Baker, 1961). This was an additive scheme for giving overall quality scores. It was based on the analysis of quality for sensory categories such as appearance, body, flavor, and aftertaste, as well as some specific attributes like sweetness, bitterness, and acidity. As shown in Table 17.5, different points are given for different categories, i.e., there is uneven weighting, presumably due to the different contributions of each category to the overall quality. Note that this does not produce a scale value in the psychophysical sense, but a score.

Like other quality grading schemes, this method can be criticized on a number of grounds. First, the weighting system is somewhat arbitrary—different versions can be found—and it was based on the expert opinion of the method's originators, rather than any consumer opinion. Second, whether wine quality can actually be captured by an additive scheme is questionable. Some defects (e.g., bitterness) are simply too serious to provide any good score at all, even though all other attributes might add up to some positive number. Some versions of the technique try to allow for this by providing a few overall quality points to add into the total, a kind of global fudge factor. Also, it is an anecdotal observation that some judges who gain experience with the technique score an overall quality level of wines first and do not bother assigning individual category points to start. Rather, they first decide on an overall score and then allot points into the individual categories, using the method backward.

A simplified alternative procedure is based on hedonic scoring by experienced fine wine drinkers (Goldwyn and Lawless, 1991). The assumption was that a small panel of experienced tasters can provide recognition of good to superior products based on their personal likes and dislikes. The method works to the extent that fine wine drinkers form a cultural and linguistic community (Solomon, 1990) with known and consensual standards of taste at least within a given geographical area. The method used a balanced 14-point hedonic scale (like extremely to dislike extremely) with no neutral center point. Wines were tasted twice and the second final score was recorded. The procedure followed principles of good practice such as randomized orders, blind coding, independent judgments (no conferring), standardized presentation,

Table 17.5 Example of 20-point wine scoring scheme

Characteristic	Scoring guide	Maximum points
Appearance	Cloudy 0, clear 1, brilliant 2	2
Color	Distinctly off 0, slightly off 1, correct 2	2
Aroma and bouquet	Vinous 1, distinct but not varietal 2, varietal 3 subtract 2 for off-odors and 1 for bottle bouquet	4
Vinegary	Obvious 0, slight 1, none 2	2
Total acidity	Distinctly high or low 0, slightly high or low 1, normal 2	2
Sweetness	Too high or low 0, normal 1	1
Body	Too high or low 0, normal 1	1
Flavor	Distinctly abnormal 0, slightly abnormal 1, normal 2	2
Bitterness	Distinctly high 0, slightly high 1, normal 2	2
General quality	Lacking 0, slight 1, impressive 2	2
Total score		20

Modified from Amerine and Roessler (1981)

palate cleansers, and a reasonable pace of tasting. Flights consisted of seven wines and the pace was limited to at least 30 min per flight, allowing time for palate recovery. Blind duplicates were periodically introduced to check on judge reliability.

These methods represent improvements over the types of informal consensus tastings done by juries to award medals at state fair competitions. Such evaluations have almost no scientific merit, i.e., they are of about the same value as a movie critic's review. Analysis of 3 years of data from blind duplicate samples in the California state wine judging has shown that among non-defective wines, any grade or medal may be assigned to any wine in different competitions, and about 90% of judges were unable to reproduce their scores (Hodgson, 2008).

17.7 Program Requirements and Program Development

17.7.1 Desired Features of a Sensory QC System

Rutenbeck (1985) and Mastrian (1985) outlined the program development of a sensory QC system in terms of specific tasks. These included research into availability and expertise of panelists, availability or access to reference materials, and time constraints. Panelist selection, screening, and training on objective terms (such as "high saltiness," as opposed to vague terms like "poor quality") must be undertaken. Sampling schemes must be developed and agreed upon as well as standard procedures for sample handling and storage. Data handling, report format, historical archiving, and tracking and panelist monitoring are all important tasks. It is extremely important that a sensory evaluation coordinator with a strong technical background in sensory methods should be assigned to carry out these tasks (Mastrian, 1985). Aside from these practical operational concerns, the system should also have certain features that maintain the quality of the evaluation procedures themselves. For example, a method for measuring the overall effectiveness of the system should be identified (Rutenbeck, 1985). External auditing at periodic intervals may be useful (Bauman and Taubert, 1984).

Gillette and Beckley (1992) listed requirements for a good in-plant sensory QC program and ten other desirable features. These concerns are taken from the perspective of ingredient suppliers to a major food manufacturer but can be modified to fit other manufacturing situations. A sensory QC program must involve human evaluation of the products. It must be acceptable to both suppliers and customers. Results must be easily communicated so that reasons for rejection and actions to be taken are both made clear. It should take into account an acceptable range of deviation, recognizing that some products will not match the gold standard but will still be acceptable to consumers. Of course, the program must be able to detect unacceptable production samples.

Additional desirable features include the following: Potential transfer over time to an instrumental measure is a good goal if the evaluations are very repetitive, as instruments do not become fatigued or bored with the testing regimen. This presumes that tight sensory–instrumental correlations can be established. Ideally, a sensory QC program should provide rapid detection for online corrections. Information should be quantitative and interface with other QC methods. As quality control and shelf-life tests are often similar, the methods will be more useful if they are transferable to shelf-life monitoring. Many of the changes over time in a stored product are also quality problems, such as deterioration in texture, browning, oxidation, syneresis, oiling-out, staling, and off-flavor development (Dethmers, 1979). Sensory evaluations may include raw ingredient testing as well as in-process and finished products. A good sensory QC program will produce a track record of actually flagging bad products to prevent further problems down the line or in consumer opinion in the marketplace.

Considering these desired features, there are some traditional test methods from the mainstream of sensory evaluation that simply do not apply well to in-plant quality control work. Problems arise if the tests cannot handle a sufficient volume of production samples. Any test procedure that has a slow turnaround in analysis and reporting of results will not be suitable for online corrective action in the manufacturing environment. For example, it is difficult to implement a descriptive analysis panel for QC work if decisions have to be made in the middle of the night on third shift and the data cannot be statistically analyzed through an automated system. At first glance, finding defective products would seem to suggest that a simple difference test from a standard product would be a good approach. However, most sensory difference tests take the form of forced-choice tasks, like the triangle procedure. The triangle test is useful for detecting any difference at all, but is not suitable when there is a range of acceptable variation. Just because a product is found to be different from the standard does not mean that it is unacceptable.

17.7.2 Program Development and Management Issues

Management may need to be educated as to the cost and practical issues that are involved in sensory QC. Rutenbeck (1985) described the "selling" of a sensory QC program and suggested calculations of measurable results, such as reductions in consumer complaints, cost savings in avoiding rework or scrapping materials, and potential impact on sales volume. Manufacturing executives unfamiliar with sensory testing can easily underestimate the complexity of sensory tests, the need for technician time to setup, the costs of panel startup and panelist screening, and training of technicians and panel leaders as well as panelist incentive programs (Stouffer, 1985). If employees are used as panelists, another stumbling block can be the personnel time away from the person's main job to come to sensory testing (and any associated costs). However, panel participation can be a welcome break for workers, can enhance their sense of participation in corporate quality programs, can expand their job skills and their view of manufacturing, and does not necessarily result in a loss of productivity. There are considerable advantages in using panelists from the processing operation, notably in accessibility and interest (Mastrian, 1985). Arranging for a sensory testing space may also involve some startup costs. An important issue concerns what will be done to insure continuity in the program. Management must be made to see that the sensory instrument will need maintenance, calibration, and eventual replacement. Concerns include panelist attrition and retraining, refreshment, or replacement of reference standards (Wolfe, 1979).

An early issue in program development concerns the definition of standards and cutoffs or specification limits (Stevenson et al., 1984). Management or preferably experienced technical personnel can do the evaluation and set the limits. This approach is fast and simple, but risky, since there is no consumer input (McNutt, 1988). The safest but slowest and most expensive approach is to give a range of products with representative production variation to consumers for evaluation. This calibration set should include known defects that are likely to occur and all ranges of processing and ingredient variables. As a small number of consumers will always be insensitive to any sensory differences, a conservative estimate of problem areas should be set based on rejection or failing scores from a minority of participants.

A third issue concerns the level of thoroughness in sampling that is needed for management comfort versus the cost of overtesting. Ideal quality control programs would sample materials along all stages of production, in every batch and every shift (Stouffer, 1985). This is rarely practical for sensory testing. Sampling multiple products from a batch or production run or performing replicate measurements with a sensory panel will give insurance against missing out-of-spec products, but will increase time and costs of testing.

Additional challenges arise from reporting structures, multiple test sites, and the temptation to substitute instruments for sensory panels. A built-in conflict of interest occurs when a QC department reports directly to manufacturing, since manufacturing is usually rewarded for productivity. A separate reporting structure may be desirable for quality control, so that executives committed to a corporate quality program can insulate the QC department from pressures to pass bad products. Across multiple plants, there is a need to standardize sensory QC procedures and coordinate activities (Carlton, 1985; Stouffer, 1985). This includes maintenance of consistent production samples and reference materials that can be sent to all other plants for comparison. Setting up similar sensory QC systems in different countries and cultures may present difficult challenges to the sensory program coordinator. Considerations in panel setup in other cultures are given in Carlton (1985). Finally, instruments cannot replace sensory evaluation for many important product characteristics (Nakayama and Wessman, 1979). Odor analysis is a good example. In other cases, the instrumental–sensory relationship may be nonlinear or indicate changes that are not perceivable at all (Rutenbeck, 1985; Trant et al., 1981).

17.7.3 The Problem of Low Incidence

A special problem with QC testing as well as shelf-life studies is that the majority of the evaluations result in positive results (favorable decisions for manufacturing). This is in the very nature of the test scenario. Good products are much more often tested than defective ones, and there is a much lower incidence of negative test results than those found in research support testing. This can present a special challenge to the credibility of the sensory testing program.

Table 17.6 shows an incidence diagram for a fairly high rate of problem products, in this case 10% (the credibility problem gets even worse if there is a lower rate of defective products). In over 1,000 tests, then, 100 are objectively defective, while 900 are objectively trouble free. If the tests are properly done, and there is statistical protection against Type I and Type II errors, there will still be some occasions where errors do occur. For the sake of easy calculation, let the long-term alpha- and beta-risks for the testing program be 10%. This means that 10% of the time when a defective product is sent for testing, it will go undetected by the evaluation, and 10% of the time a product which has no defects will be flagged, due to random error. This will lead to 810 correct “pass” decisions and 90 correct detections of sensory problems. Unfortunately, due to the high incidence of good products being tested, the 10% false alarm rate leads to 90 products also being flagged where there is no true sensory problem. Note that this assumes that there is good sensory testing and proper statistical treatment of the results! The problem arises when the sensory QC leader picks up the phone and calls the manufacturing manager and “rings the alarm bell.” Given this incidence, the probability of being right about the problem is only 50%, in other words, no better than a coin toss! Even if alpha is reduced to the usual 5%, there is still a one-in-three chance of false alarms.

How can this occur if good sensory testing is done and proper statistics are applied? The answer is that our normal inferential statistics are used to view the outcome chart in Table 17.6 across rows and not down columns. The problem in sensory QC is that given a low incidence of problems, there is simply a high rate of false alarms relative to correct detections. This can hurt the credibility of the program if manufacturing managers develop a feeling that the sensory department is prone to “cry wolf.” Thus it is wise to build in a system for additional or repeated testing of product failures to insure that marginal products are in fact defective before action is taken.

17.8 Shelf-Life Testing

17.8.1 Basic Considerations

Shelf-life or stability testing is an important part of quality maintenance for many foods. It is an inherent part of packaging research because one of the primary functions of food packaging is to preserve the integrity of a food in its structural, chemical, microbiological

Table 17.6 Bayesian incidence chart

	Outcome of evaluation		
	Problem reported	No problem reported	Incidence (=total across row)
Problem exists (description)	90 ("hit rate")	10 (Type II error)	100
No problem exists	90 ("false alarm")	810 (correct acceptance)	900
(Total)	180	820	1000

Let alpha = 0.10 and beta = 0.10

Assume 1,000 tests are conducted, with a 10% rate of faulty products

Numbers in cells show estimated numbers of problems reported or not, based on alpha and beta rates of 10%

Given that a problem was reported, you stand a 50/50 chance of having made the wrong decision (90/180)!

and sensory properties. A good review of shelf-life testing can be found in the packaging text by Robertson (2006) and the reader is referred there for further information on modeling and accelerated storage tests. For many foods, the microbiological integrity of the food will determine its shelf life, and this can be estimated using standard laboratory practices; no sensory data are required. The sensory aspects of a food are the determining factor for the shelf life of foods that do not tend to suffer from microbiological changes such as baked goods. Sensory tests on foods are almost always destructive tests, so sufficient samples must be stored and available, especially during the period in which the product is expected to deteriorate (Gacula, 1975).

Shelf-life testing may employ any of the three major kinds of sensory tests, discrimination, descriptive, or affective, depending on the goals of the program (Kilcast, 2000). Thus one can view shelf-life tests as no special category of sensory testing, but simply a program of repeated testing using accepted methods. The objectives of the study may dictate what method is most suitable to answer the research questions (Dethmers, 1979). For a designed study to evaluate the effects of a new packaging film, a simple discrimination test might be appropriate to test for changes versus the existing packaging. For purposes of establishing an open dating system, consumer acceptability tests would be appropriate to establish the time that the product is likely to become unacceptable. If the product is new, a descriptive analysis profile is needed to establish the full sensory specification of what a fresh product tastes like. If a product has failed a consumer evaluation, it is often appropriate to submit the samples to descriptive testing to try and understand the reasons for failure and which aspects have deteriorated (Dethmers, 1979). If the purpose is to establish a suitable degree of stability, i.e., that the failure time exceeds the typical distribution and use time by consumers, a combination of two tests may be appropriate. It is cost efficient in this case to perform discrimination tests against a fresh control or standard product and then perform consumer acceptance tests if any difference is detected.

According to Peryam (1964) and Dethmers (1979) a shelf-life program will involve the following steps: (1) formulating objectives, (2) obtaining representative samples, (3) determining the physical and chemical composition of the test products, (4) setting up a test design, (5) choosing the appropriate sensory method, (6) choosing the storage conditions, (7) establishing the control product or products to which the stored product will be compared, (8) conduct the periodic testing, and (9) determine the shelf life based on the results.

Important strategic choices include the nature of the control product and the storage conditions. Storage conditions should mimic the conditions found in distribution and those in stores, unless some accelerated storage conditions are required. Ideal conditions are generally a poor choice. The control product presents special problems. If a fresh product is available at the different time intervals, how does one know that subsequent batches have not drifted or changed since the initial product was manufactured? If a fresh product is stored from the initial batch under ideal conditions, how does one know it has not changed? There is no perfect solution to this problem. Sometimes a study will involve more than one standard. Reference standards should be clearly identified by date, lot number, production location, etc. A separate program may be instituted to insure the integrity and constancy of reference standards. Descriptive evaluation may be beneficial for this purpose. Options for references include the following: current plant product that has passed QC, current pilot-plant prototype, historical product, optimally stored product, a written descriptive profile, and a mental reference (Wolfe, 1979).

Two main choices are used for criteria for product failure. These include a cutoff point from a critical descriptive attribute (or set of them) and consumer data when the product is rejected as unacceptable. Statistical modeling with equations such as a hazard function or survival analysis is discussed below. Note that product failure is an all-or-none phenomenon, and decreases in sensory measures such as falling acceptability or increasing percents of consumer rejection are more continuous in nature. This opens the opportunity for other kinds of models, such as logistic regression against percent rejecting (Giminez et al., 2007).

17.8.2 Cutoff Point

The choice of a cutoff point has two implications. The first occurs when the cutoff point itself is used as an action standard. When the product gets to this point, we consider it to have reached the end of its useful life. It is no longer salable. The time estimate may be used for some purpose like open dating or "use-by" dates printed on the product package. The second implication is that when a product in a designed study reaches this cutoff point, it defines "failure" and will be used as a data point in some kind of statistical modeling such as survival analysis.

Determination of a cutoff point requires careful consideration. Several options are available including (1) a significant difference in a discrimination test, (2) some degree of difference from control product on a scaled attribute or overall degree of difference scale, and (3) consumer reaction. Consumer data may involve a significant difference in acceptability ratings from control, a cutpoint on an acceptance score, or some percent of consumer rejection (e.g., 50% or 25%). Giminez et al. (2007) found the first significant difference to be too conservative an estimate in the sense that acceptance scores were still above 6 on the 9-point scale. This makes sense because two products may differ but still be acceptable (Kilcast, 2000). Another option is to use any value less than 6 on the 9-point hedonic scale (6 = like slightly, i.e., just above neutral) (Muñoz et al., 1992). Another option is to use consumer rejection ("I would not buy/eat this product") (Hough et al., 2003). These two measures are not necessarily equivalent. Giminez et al. (2008) found that for certain baked products, consumers might not like the product, but they would answer "yes" when asked if they would consume it at home (having already purchased it). This finding suggests that consumer rejection may not be sufficiently conservative, i.e., that a product may become disliked and even generate consumer complaints before it reaches the point of rejection. Giminez et al. (2007) found that acceptability scores could be related to percent of rejection by logistic regression analysis. The logistic equations for two different countries (Spain and Uruguay) for a baked product were different, a warning about cultural and/or national differences. Logistic regression is a useful general approach to data in the form of proportions that accumulate in an S-shaped curve. The general form is

$$\ln \frac{p}{1-p} = b_o + b_1 X \tag{17.2}$$

where p is the proportion rejected and X is the variable that is the predictor, such as time, or in this case acceptability scores (b are constants).

17.8.3 Test Designs

Several options are available for shelf-life tests regarding how samples are stored and test times. The simplest method is to make one large batch of product, store it under normal conditions, and test it at various intervals. However, this is not very efficient in terms of test time and risks the panel drifting its criteria. Another option is to stagger the production times, so all the products of different ages are tested on the same day. A variation on this is to store the product under conditions that essentially stop all aging processes, for example, at very low temperature. This is obviously not possible with all products (you cannot freeze lettuce). Then products are pulled from the optimal storage conditions at different times and allowed to age at normal temperatures. Another variation of this procedure is to allow products to age for different times and then place them into the optimal storage conditions, pulling everything out of storage at the test date.

17.8.4 Survival Analysis and Hazard Functions

The literature on survival analysis is very large, because a number of different fields use these kinds of statistical models, such as actuarial science for the insurance industry. Some of the models are similar to those used in chemical kinetics. These functions are useful when the product has a single process or a group of processes that are occurring at about the same time. However, some products will show a "bathtub" function with two phases of product failure (Robertson, 2006). In the early stages, some product failures occur due to faulty packaging or improper processing (see Fig. 17.4). Then the remaining products from that batch enter a period of product stability. After some time, *X*, the products begin to fail again, due to deterioration. Gacula and Kubala (1975) suggest that the shelf-life modeling should only consider those failures after time X_2.

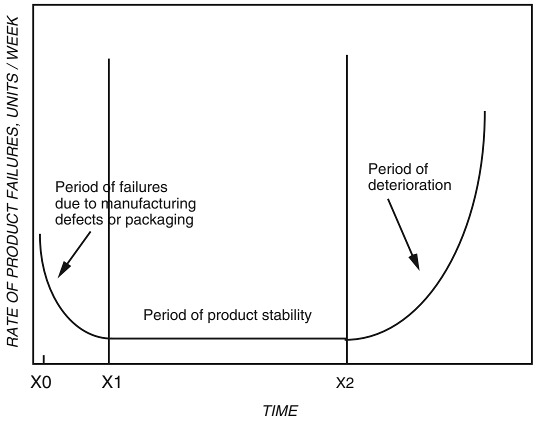


Fig. 17.4 The "bathtub" function showing a common pattern of failure rates changing over the sampling time in a shelf-life study. From time X_0 to X_1, some products will fail due to improper processing or faulty packaging. This is followed by a period of fairly low failure rates when products are stable or within specification limits. At time X_2, failures start to increase markedly. Researchers fitting hazard functions or doing survival analysis for estimation of shelf life should consider using only those times after X_2 in curve fitting as earlier failures are due to causes other than the time-related deterioration.

Survival analysis as applied to sensory data has two main tasks, the fitting of a hazard function to the data and the interpolation of some point used as the criterion for shelf life (such as 25 or 50% consumer rejection). Various functions have been used to fit the function of failure data (or percent of failures) over time. Percent survival (one minus percent of failures) often takes the form of a decaying exponential function.

An important choice of model includes the distribution used to fit the percent of failures. Many distributions have been tried (Gacula and Kubala, 1975) but two useful models are a log–normal distribution (surviving is a positively skewed distribution, few people live to 100) and a Weibull distribution. Weibull functions are useful distributions that can be used to fit a variety of data sets. They include a shape parameter and a scale parameter. When the shape parameter takes a value greater than 2, the distribution is approximately bell shaped and symmetric. These equations for failure take the following forms:

$$F(t) = \Phi\left(\frac{\ln(t) - \mu}{\sigma}\right) \text{(log −normal)} \qquad (17.3)$$

where Φ is the cumulative normal distribution function, *t* is time, $F(t)$ is the failure proportion at time *t*, μ is the mean failure time, and σ is the standard deviation.

$$F(t) = 1 - \exp\left[-\exp\left(\frac{\ln(t) - \mu}{\sigma}\right)\right] \quad \text{(Weibull)} \qquad (17.4)$$

where $\exp(x)$ is the notation for e^x.

If we make two substitutions and determine the mean (μ) and standard deviation (σ) of our failure times, a simple model can help us find the time for a given percentage of failures from the fitted Weibull equation. Let $\rho = \exp(-\mu)$. Then the following relationship holds for any proportion ($F(t)$):

$$t = \frac{-\ln(1 - F(t))^{\sigma}}{\rho} \qquad (17.5)$$

Using the log–normal model for $F(t)$, a simple graphic method for finding the interpolated 50% failure level is to do the following: For *N* samples of foods sampled over time that have known failure times T_i, rank all the batches, *i* as to the time of failure ($i = 1$ to *N*). Calculate the median ranks, MR values. The median rank can be found in some statistical tables or estimated as

$$MR = (i - 0.3)/(N + 0.4)$$

Plot the median rank on log probability paper versus T_i and interpolate at the 50% point. If a straight line fits the data, the 50% point can be estimated from the linear equation and standard deviations estimated from the probability paper. This is essentially a fit of log MRs to *z*-scores. Other percentages may be interpolated, of course, because 50% may be too high a failure percentage for many products in setting a useful limit. An equivalent mathematical solution is given in Appendix 2 of this chapter.

Hough et al. (2003) point out that the usual sensory experiment produces censored data. That is, for any batch that has failed, we only know that the time of failure was in some interval between the last test and the current test. Similarly, for a batch that has not failed at the final interval, we only know that its failure time is sometime after that final test. So the data are censored and the survival function can be estimated using maximum likelihood techniques.

17.8.5 Accelerated Storage

Sooner or later, product developers may figure out that sensory specialists do not own a time machine and cannot deliver shelf-life estimates without a long study. So they may request some accelerated storage tests to shorten the time. Such tests are based on the idea that at higher temperatures, many chemical reactions will proceed in a predictable manner, according to simple kinetic models, and thus the shelf life at long time intervals can be simulated by shorter intervals at higher temperatures (Mizrahi, 2000). Kinetic models are often based on the Arrhenius equation, and rate constants can be found from experiments conducted at different temperatures. Some of the models are shown in Appendix 3 of this chapter.

Problems in accelerated testing occur when changes in the product due to temperature are not the same as those due to storage time at normal temperatures (Robertson, 2006). Obviously, trying to measure the shelf life of a frozen food at higher temperatures makes little sense. Other foods may not follow simple predictive models as multiple processes occur with different rates. For example, at higher temperatures phase change may occur, solid to liquid. Carbohydrates in the amorphous state may crystallize. The water activity of dry foods may increase with temperature causing an increased reaction rate and overprediction of true shelf life at the normal temperature. If two reactions with different kinetic constants change at different rates with different temperatures, the one with the higher value may come to predominate. The sensory specialist should be familiar with the logic of this testing and the modeling that is commonly done, as well as the pitfalls.

17.9 Summary and Conclusions

Insuring the quality of products on a sensory basis is an important corporate goal in a competitive environment. Consumers have fixed expectations and will become disloyal to a brand if they experience substandard products. However, in spite of the need for sensory quality control, setting up and maintaining an in-plant QC program is difficult and costly. Commitment to the program is the corporate equivalent of diet and exercise. Everyone admits that it is a good idea. However, maintaining program integrity, avoiding shortcut procedures, and dealing with dwindling panel size can be challenging. The success of any program demands strong management commitment. Without management support, especially from manufacturing, a sensory QC program is bound to fail. In a typical case the program will amount to nothing more than "rubber stamping" of supervisory opinion, thus supporting a management policy that maximizes productivity at the expense of producing unacceptable products. Such programs will blow hot and cold, usually receiving some emergency attention when a truly bad product hits the retail shelves and consumer complaints filter back (Rutenbeck, 1985). After a period of improved production, some complacency may set in and loss of interest in sensory QC efforts (until the next disaster).

Implementation of a sensory QC program will involve four technical tasks. First, a range of products must be prepared for establishing quality specifications and limits. Specifications must be set in a research study with consumers or by management or expert opinion after sampling the various products. This range of products can also be used in panel training. The second step in the program is recruitment, screening, and training, in other words, panel setup. Next, standard protocols for product sampling, handling, storage, serving, blind coding, and maintenance of reference standards must be established. The fourth step is in systematizing the paperflow (Mastrian, 1985). This includes establishing standard reporting formats and processes for data handling, recommendations, and action criteria. This activity should also include mechanisms for archiving results and tracking both products and panelist performance across time.

Appendix 1: Sample Screening Tests for Sensory Quality Judges

Part 1. Paired comparison of sweetness levels

Adjust samples to three levels, e.g., 10, 11, and 12% sucrose wt/vol.

Give four pairs in counterbalanced orders, e.g.,

10 versus 11%, 11 versus 10% (hard discrimination)
10 versus 12 %, 12 versus 10% (easier discrimination)

Use a different order for different panelists. Blind code with random 3-digit labels.

Good performance: All four correct.
Acceptable performance: One error if other test sections perfect.

Part 2. Multiple choice odor identification. Done with blotters in capped jars

Circle correct answers on sheet. Make up multiple forms with different orders.

Use random 3-digit codes on bottles. Odors represent common notes in the product.

Use four alternatives, e.g., fruity, smoky, vinegar, onion

(a) Dilute ethyl hexanoate (or similar ester)
(b) Dilute ethyl 2-methyl butyrate
(c) Dilute vinegar
(d) Dilute phenylethanol
(e) Trans-2-hexenol(dilute until green or leafy smell is obtained)

Good performance: 4/5 correct
Acceptable performance:3/5 correct

Part 3. Odor discrimination test

Run triangle tests with base juice and base juice + 1% vinegar.
Run triangle test with base juice + 0.1% butyric acid.
Subjects should sniff first, then taste.
Provide palate cleansers (water, crackers).
Run duplicates of each test.

Good performance: 3/4 correct.
Acceptable performance: 2/4 correct.

Part 4. Acidity test

Adjust pH to about 0.5 versus 1% titratable acidity.

Use four paired comparisons, as in sweetness test above.

Good performance: 3/4 correct.
Acceptable performance: 2/4 correct.

After scoring, rank order candidates from highest to lowest.
Invite the top 50% from each shift for training.
Send thank you notes to all the people who try out.
Keep the rest “on file” for possible replacements if scores are acceptable.

Appendix 2: Survival/Failure Estimates from a Series of Batches with Known Failure Times

This procedure follows the graphic method given in Section 17.8.3 but allows a more exact fit by least squares regression.

1. For N samples of foods sampled over time that have known failure times T_i, rank all the batches, i as to the time of failure ($i = 1$ to N).
2. Calculate the median ranks, MR values. The median rank can be found in some statistical tables or estimated as

$$MR = (i - 0.3)/(N + 0.4)$$

3. Convert each T_i to $\ln(T_i)$, called Y_i. This will permit a fit of MR to the log–normal model.
4. Calculate the z-score for each MR at each T_i. Call this X_i.
5. Regress Y against X using least squares to get the linear equation $Y = a + bX$.

This is equivalent to finding the straight line fit to the log probability plot described in Section 17.8.

6. Then solve for $Y = 0$ (z-score for 50%) which is $X = –a/b$, to get the 50th percentile.
7. Convert back to the original units by exponentiating Time at 50% failure = $e^X = e^{-a/b}$.

Appendix 3: Arrhenius Equation and Q_{10} Modeling

The reaction time for product failures may be linear (zero order) or a decaying exponential (first order). Both allow determination of a rate constant, K. Let us consider a cutoff point rating on some scale, R, as the event to be modeled as function of time, t. R could also be any event that signals product failure.

The zero-order equation is

$$R = R_o - kt \tag{17.6}$$

And the first-order relationship is

$$R = R_o e^{-kt} \tag{17.7}$$

where R_o is the rating or failure at t=0
and

$$\ln \frac{R}{R_o} = -kt \tag{17.8}$$

Reaction rates are also dependent on temperature, so in accelerated storage studies, the Arrhenius equation provides a starting point or a generally useful approximation:

$$k = k_o e^{\left(\frac{-E_A}{RT}\right)} \tag{17.9}$$

and

$$\ln k = \ln k_o - \frac{E_A}{R}\left(\frac{1}{T}\right) \tag{17.10}$$

where k is the rate constant to be estimated, k_o is a constant independent of temperature (also known as the Arrhenius, pre-exponential, collision, or frequency factor), E_A is the activation energy (J/mol), R is the ideal gas constant, T is temperature (absolute, K).

So a plot of ln(k) versus 1/T can be used to find the activation energy, E_A.

This sometimes takes its derivative form:

$$\frac{d(\ln k)}{dT} = \frac{E_A}{RT^2} \tag{17.11}$$

The activation energy is fictitious in a way, because there is not a single chemical reaction going on during the aging of a food product, but a large number of simultaneous processes. Nonetheless, we can think of this as useful for two reasons. First, it gives an indication of the fragility of the food (lower activation energy would mean faster deterioration). Second, the E_A value becomes useful in predicting what happens at different temperatures. Of specific interest is predicting what will happen at "normal" temperature given that an accelerated storage study has been conducted at higher temperatures.

Experiments are often performed at varying temperatures, 10°C apart, to generate what is known as the Q_{10} factor.

$$Q_{10} = \frac{k_{T+10}}{k_T} = \frac{S_T}{S_{T+10}} \tag{17.12}$$

where k_{T+10} and k_T are the rate constants at temperature T and $T + 10$, and S_T and S_{T+10} are the corresponding shelf-life estimates. Note that the ratio of rate constants is the inverse of the ratio of the shelf-life times.

This produces some useful relationships, for example, to use in estimating the activation energy, E_A

$$\ln Q_{10} = \frac{10E_A}{RT^2} \tag{17.13}$$

Again, E_A can give some idea of the susceptibility of the product to deterioration. Once the Q_{10} factor has been determined for a product, the time–temperature relationship can be predicted for the accelerated tests. A useful factor to determine is an acceleration factor, AF. This will help us convert from the accelerated temperature back to a usage temperature or normal storage condition temperature like 20°C. Hough (2010) gives the following example, based on an off-flavor rating (OF) as a function of time and temperature:

$$\text{OF}_{T,\text{temp}} = \text{OF}_o + (\text{AF})k_u(T_{\text{temp}}) \tag{17.14}$$

where OF_o is the off-flavor at time zero, k_u is the rate constant at usage temperature, and T_{temp} is the accelerated test temperature. Knowing E_A, we can also estimate the AF from

$$\mathrm{AF} = \exp\left[\frac{E_A}{R}\left(\frac{1}{T_u} - \frac{1}{T_{\text{test}}}\right)\right] \tag{17.15}$$

where $\exp(X)$ is e^X, T_u is the usage temperature, and T_{test} is the accelerated test temperature.

For example, if we determine that the E_A is 6,500 cal/mol, then we can calculate an acceleration factor based on a test at 40°C and a usage temperature at 20°C:

$$\mathrm{AF} = \exp\left[6500\left(\frac{1}{293} - \frac{1}{313}\right)\right] = 4.13$$

Suppose we determine that at accelerated temperature of 40°C, we have a failure time of 35 days. Then we can find the failure time, FT_u, at usage temperature T_u, we merely multiply by the acceleration factor

$$\mathrm{FT}_u = \mathrm{FT}_{\text{test}}(\mathrm{AF}) = 35(4.13) = 145$$

Thus our accelerated test predicts a (mean) failure time at 145 days for the product stored at room temperature or about 20°C.

References

Amerine, M. R. and Roessler, E. B. 1981. Wines, Their Sensory Evaluation, Second Edition. W.H. Freeman, San Francisco, CA.

Amerine, M. R., Pangborn, R. M. and Roessler, E. B. 1965. Principles of Sensory Evaluation of Foods. Academic, New York, NY.

Aust, L. B., Gacula, M. C., Beard, S. A. and Washam, R. W. 1985. Degree of difference test method in sensory evaluation of heterogeneous product types. Journal of Food Science, 50, 511–513.

Bauman, H. E. and Taubert, C. 1984. Why quality assurance is necessary and important to plant management. Food Technology, 38(4), 101–102.

Beckley, J. P. and Kroll, D. R. 1996. Searching for sensory research excellence. Food Technology, 50(2), 61–63.

Bodyfelt, F. W., Tobias, J. and Trout, G. M. 1988. Sensory Evaluation of Dairy Products. Van Nostrand/AVI, New York, NY.

Bressan, L. P. and Behling, R. W. 1977. The selection and training of judges for discrimination testing. Food Technology, 31, 62–67.

Cardello, A. V. 1995. Food quality: Relativity, context and consumer expectations. Food Quality and Preference, 6, 163–170.

Carlton, D. K. 1985. Plant sensory evaluation within a multiplant international organization. Food Technology, 39(11), 130–133, 142.

Dethmers, A. E. 1979. Utilizing sensory evaluation to determine product shelf life. Food Technology, 33(9), 40–43.

Ennis, D. M., Mullen, K. and Frijters, J. E. R. 1988. Variations of the method of triads: Unidimensional Thurstonian models. British Journal of Mathematical and Statistical Psychology, 41, 25–36.

Gacula, M. C. 1975. The design of experiments for shelf life study. Journal of Food Science, 40, 399–403.

Gacula, M. C. and Kubala, J. J. 1975. Statistical models for shelf life failures. Journal of Food Science, 40, 404–409.

Garvin, D. A. 1987. Competing on the eight dimensions of quality. Harvard Business Review, 65(6), 101–109.

Gillette, M. H. and Beckley, J. H. 1992. In-Plant Sensory Quality Assurance. Paper presented at the Annual Meeting, Institute of Food Technologists, New Orleans, LA, June, 1992.

Giminez, A., Ares, G. and Gambaro, A. 2008. Survival analysis to estimate sensory shelf life using acceptability scores. Journal of Sensory Studies, 23, 571–582.

Giminez, A., Varela, P., Salvador, A., Ares, G., Fiszman, S. and Garitta, L. 2007. Shelf life estimation of brown pan bread: A consumer approach. Food Quality and Preference, 18, 196–204.

Goldwyn, C. and Lawless, H. 1991. How to taste wine. ASTM Standardization News, 19(3), 32–27.

Hammond, E., Dunkley, W., Bodyfelt, F., Larmond, E, and Lindsay, R. 1986. Report of the committee on sensory data to the journal management committee. Journal of Dairy Science, 69, 298.

Hodgson, R. T. 2008. An examination of judge reliability at a major U.S. wine competition. Journal of Wine Economics, 3, 105–113.

Hough, G. 2010. Sensory Shelf Life Estimation of Food Products. CRC Press, Boca Raton, FL.

Hough, G., Langohr, K., Gomez, G. and Curia, A. 2003. Survival analysis applied to sensory shelf life of foods. Journal of Food Science, 68, 359–362.

International Olive Oil Council. 2007. Sensory analysis of olive oil. Method for the organoleptic assessment of virgin olive oil. http://www.internationaloliveoil.org/.

Kilcast, D. 2000. Sensory evaluation methods for shelf-life assessment. In: D. Kilcast and P. Subramaniam (eds.), The Stability and Shelf-Life of Food. CRC/Woodhead, Boca Raton, FL, pp. 79–105.

Lawless, H. T. 1995. Dimensions of quality: A critique. Food Quality and Preference, 6, 191–196.

Lawless, H. T. and Claassen, M. R. 1993. Validity of descriptive and defect-oriented terminology systems for sensory analysis of fluid milk. Journal of Food Science, 58, 108–112, 119.

Mastrian, L. K. 1985. The sensory evaluation program within a small processing operation. Food Technology, 39(11), 127–129.

McBride, R. L. and Hall, C. 1979. Cheese grading versus consumer acceptability: An inevitable discrepancy. Australian Journal of Dairy Technology, June, 66–68.

McNutt, K. 1988. Consumer attitudes and the quality control function. Food Technology, 42(12), 97, 98, 108.

Mizrahi, S. 2000. Accelerated shelf-life tests. In: D. Kilcast and P. Subramaniam (eds.), The Stability and Shelf-life of Foods. CRC /Woodhead, Boca Raton, FL, pp. 107–142.

Moskowitz, H. R. 1995.Food Quality: conceptual and sensory aspects. Food Quality and Preference, 6, 157–162.

Muñoz, A. M., Civille, G. V. and Carr, B. T. 1992. Sensory Evaluation in Quality Control. Van Nostrand Reinhold, New York, NY.

Nakayama, M. and Wessman, C. 1979. Application of sensory evaluation to the routine maintenance of product quality. Food Technology, 33(9), 38, 39 ,44.

Nelson, L. 1984. The Shewart control chart-tests for special causes. Journal of Quality Technology, 16, 237–239.

Nelson, J. and Trout, G. M. 1964. Judging Dairy Products. AVI, Westport, CT.

O'Mahony, M. 1981. Our-industry today—psychophysical aspects of sensory analysis of dairy products: A critique. Journal of Dairy Science, 62, 1954–1962.

Ough, C. S. and Baker, G. A. 1961. Small panel sensory evaluations of wines by scoring. Hilgardia, 30, 587–619.

Pangborn, R. M. and Dunkley, W. L. 1964. Laboratory procedures for evaluating the sensory properties of milk. Dairy Science Abstracts, 26, 55–62.

Pecore, S., Stoer, N., Hooge, S., Holschuh, N., Hulting, F. and Case, F. 2006. Degree of difference testing: A new approach incorporating control lot variability. Food Quality and Preference, 17, 552–555.

Peryam, D. R. 1964. Consumer preference evaluation of the storage stability of foods. Food Technology, 18, 214.

Reece, R. N. 1979. A quality assurance perspective on sensory evaluation. Food Technology, 33(9), 37.

Robertson, G. L. 2006. Food Packaging, Principles and Practice, Second Edition. CRC/Taylor and Francis, Boca Raton, FL.

Rutenbeck, S. K. 1985. Initiating an in-plant quality control/sensory evaluation program. Food Technology, 39(11), 124–126.

Regenstein, J. M. 1983. What is fish quality? Infofish, June, 23–28.

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, 2292–2302.

Solomon, G. E. A. 1990. The psychology of novice and expert wine talk. American Journal of Psychology, 103, 495–517.

Stevenson, S. G.,Vaisey-Genser, M. and Eskin, N. A. M. 1984. Quality control in the use of deep frying oils. Journal of the American Oil Chemist's Society, 61, 1102–1108.

Stouffer, J. C. 1985. Coordinating sensory evaluation in a multiplant operation. Food Technology, 39(11), 134–135.

Thompson, R. H. and Pearson, A. M. 1977. Quantitative determination of 5 Androst-16-en-3-one by gas chromatography-mass spectrometry and its relationship to sex odor intensity of pork. Journal of Agricultural and Food Chemistry, 25, 1241–1245.

Trant, A. S., Pangborn, R. M. and Little, A. C. 1981. Potential fallacy of correlating hedonic responses with physical and chemical measurements. Journal of Food Science, 46, 583–588.

Wolfe, K. A. 1979. Use of reference standards for sensory evaluation of product quality. Food Technology, 33(9), 43–44.

York, R. K. 1995. Quality assessment in a regulatory environment. Food Quality and Preference, 6, 137–141.

Young, T. A., Pecore, S., Stoer, N., Hulting, F., Holschuh, N. and Case, F. 2008. Incorporating test and control product variability in degree of difference tests. Food Quality and Preference, 19, 734–736.