Chapter 9

Context Effects and Biases in Sensory Judgment

Abstract Human judgments about a sensation or a product are strongly influenced by items that surround the item of interest, either in space or in time. This chapter shows how judgments can change as a function of the context within which a product is evaluated. Various contextual effects and biases are described and categorized. Some solutions and courses of action to minimize these biases are presented.

By such general principles of action as these everything looked at, felt, smelt or heard comes to be located in a more or less definite position relatively to other collateral things either actually presented or only imagined as possibly there.

— James (1913, p. 342)

Contents

9.1 Introduction: The Relative Nature of Human Judgment

This chapter will discuss context effects and common biases that can affect sensory judgments. Context effects are conditions in which the judgment about a product, usually a scaled rating, will shift depending upon factors such as the other products that are evaluated in the same tasting session. A mediocre product evaluated in the context of some poor-quality items may seem very good in comparison. Biases refer to tendencies in judgment in which the response is influenced in some way to be an inaccurate reflection of the actual sensory experience. In magnitude estimation ratings, for example, people have a tendency to use numbers that are multiples of 2, 5, and 10, even though they can use any number or fraction they wish. At the

H.T. Lawless, H. Heymann, *Sensory Evaluation of Food*, Food Science Text Series,
DOI 10.1007/978-1-4419-6488-5_9,

end of the chapter, some solutions to these problems are offered, although a sensory scientist should realize that we can never totally eliminate these factors. In fact, they are of interest and deserve study on their own for what they can tell us about human sensory and cognitive processes.

An axiom of perceptual psychology has it that humans are very poor absolute measuring instruments but are very good at comparing things. For example, we may have difficulty estimating the exact sweetness level of our coffee, but we have little trouble in telling whether more sugar has been added to make it sweeter. The question arises, if people are prone to making comparisons, how can they give ratings when no comparison is requested or specified? For example, when asked to rate the perceived firmness of a food sample, how do they judge what is firm versus what is soft? Obviously, they must either choose a frame of reference for the range of firmness to be judged or be trained with explicit reference standards to understand what is high and low on the response scale. In other words, they must relate this sensory judgment to other products they have tried. For many items encountered in everyday life, we have established frames of reference based on our experiences. We have no trouble forming an image of a "large mouse running up the trunk of a small elephant" because we have established frames of reference for what constitutes the average mouse and the average elephant. In this case the judgment of large and small is context dependent. Some people would argue that all judgments are relative.

This dependence upon a frame of reference in making sensory judgments demonstrates the influence of contextual factors in biasing or changing how products are evaluated. We are always prone to see things against a background or previous experience and evaluate them accordingly. A 40° (Fahrenheit) day in Ithaca, New York, in January seems quite mild against the background of the northeastern American winter. However, the same 40°C temperature will feel quite cool on an evening in August in the same location. This principle of frame of reference is the source of many visual illusions, where the same physical stimulus causes very different perceptual impressions, due to the context within which it is embedded. Examples are shown in Fig. 9.1.

A simple demonstration of context is the visual afterimage effect that gave rise to Emmert's law (Boring, 1942). In 1881, Emmert formalized a principle of size constancy based on the following effect: Stare for about 30 s at a brightly illuminated colored paper rectangle (it helps to have a small dot to aid in fixation in the center) about a meter away. Then shift your gaze to a white sheet on the table in front of you. You should see the rectangle afterimage in a complementary color and somewhat smaller in size as compared to the original colored rectangle. Next, shift your gaze to a white wall some distance off. The afterimage will now appear much larger, as the brain finds a fixed visual angle at greater distance to represent larger physical objects. Since the mind does not immediately recognize that the afterimage is just a creation of the visual sensory system, it projects it at the distance of the surface upon which it is "seen." The more distant frame of reference, then, demands a larger size perception.

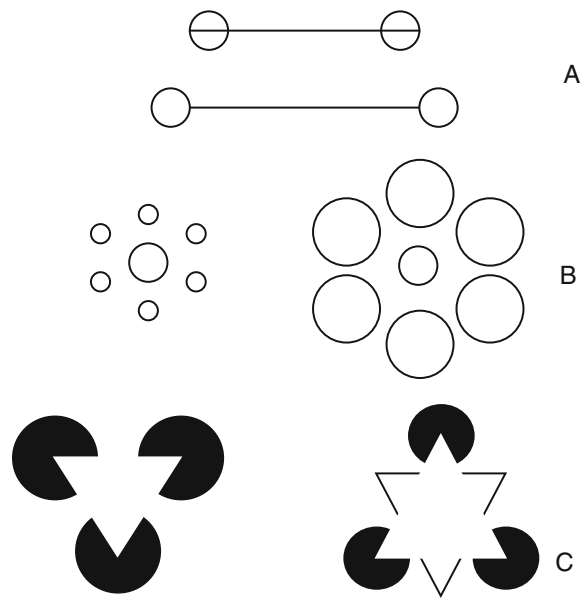

Fig. 9.1 Examples of contextual effects from simple visual illusions. (**a**) The dumbell version of the Muller–Lyer illusion. (**b**) The Ebbinghaus illusion. (**c**) Illusory contours. In this case the contexts induce the perceptions of shapes.

The close link between sensory judgments and context presents problems for anyone who wants to view ratings as absolute or comparable across different times, sessions, or settings. Even when the actual sensory impression of two items is the same, we can shift the frame of reference and change the overt behavior of the person to produce a different response. This problem (or principle of sensory function) was glossed

over by early psychophysical scientists. In psychological terms, they used a simple stimulus–response (S–R) model, in which response was considered a direct and unbiased representation of sensory experiences. Certain biases were observed, but it was felt that suitable experimental controls could minimize or eliminate them (Birnbaum, 1982; Poulton, 1989).

A more modern view is that there are two or three distinct processes contributing to ratings. The first is a psychophysical process by which stimulus energy is translated into physiological events that result in a subjective experience of some sensory intensity. The second, equally important process is the function by which the subjective experience is translated into the observed response, i.e., how the percept is translated onto the rating scale (see Fig. 9.2). Many psychophysical researchers now consider a "judgment function" to be an important part of the sequence from stimulus to response (Anderson, 1974, Birnbaum, 1982, McBride and Anderson, 1990; Schifferstein and Frijters, 1992; Ward, 1987). This process is also sometimes referred to as a response output function. A third intermediate step is the conversion of the raw sensory experience into some kind of encoded percept, one that is available to memory for a short time, before the judgment is made (Fig. 9.2c).

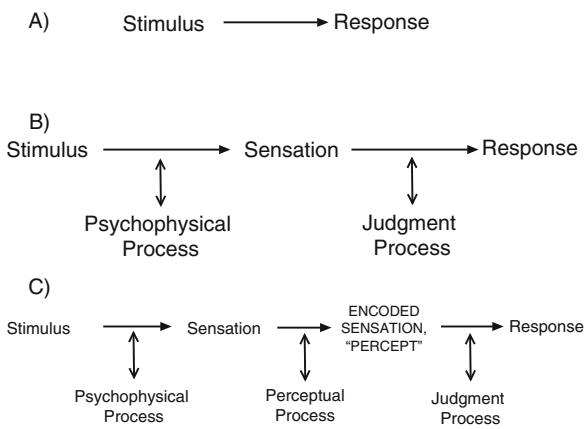

Fig. 9.2 Models for sensory-response processes. (**a**) The simple stimulus–response model of twentieth-century behavioral psychology. (**b**) Two processes are involved in sensation and response, a psychophysical process and then a response output or a judgment process in which the participant decides what response to give for that sensation. (**c**) A more complex model in which the sensation may be transformed before the response is generated. It may exist in short-term memory as an encoded percept, different from the sensation. Contextual effects of simultaneous or sequential stimuli can influence the stimulus–response sequence in several ways. Peripheral physiological effects such as adaptation or mixture inhibition may change the transduction process or other early stages of neural processing. Other stimuli may give rise to separate percepts that are integrated into the final response. Contextual factors may also influence the frame of reference that determines how the response output function will be applied. In some models, an additional step allows transformation of the percept into covert responses that are then translated as a separate step into the overt response R.

Given this framework, there are several points at which stimulus context may influence the sensory process. First, of course, the actual sensation itself may change. Many sensory processes involve interaction effects of simultaneous or sequential influences of multiple items. An item may be perceived differently due to the direct influence of one stimulus upon another that is nearby in time or space. Simultaneous color contrast is an example in color vision and some types of inhibitory mixture interactions and masking in taste and smell are similarly hard wired. Quinine with added salt is less bitter than quinine tasted alone, due to the ways that sodium ions inhibit bitterness transduction. Sensory adaptation weakens the perception of a stimulus because of what has preceded it. So the psychophysical process itself is altered by the milieu in which the stimulus is observed, sometimes because of physical effects (e.g., simple buffering of an acid) or physiological effects (e.g., neural inhibition causing mixture suppression) in the peripheral sensory mechanisms. A second point of influence is when the context shifts the frame of reference for the response output function. That is, two sensations may have the same subjective intensity under two conditions, but because of the way the observer places them along the response continuum (due to different contexts), they are rated differently. A number of studies have shown that contextual factors such as the distribution of stimuli along the physical continuum affect primarily (although not exclusively) the response output function (Mellers and Birnbaum, 1982, 1983). A third process is sometimes added in which the sensation itself is translated into an implicit response or encoded image that may also be affected by context (Fig. 9.2c). This would provide another opportunity to influence the process if contextual factors affect this encoding step.

Contextual change can be viewed as a form of bias. Bias, in this sense, is a process that causes a shift or a change in response to a constant sensation. If one situation is viewed as producing a true

and accurate response, then the contextual conditions that cause shifts away from this accurate response are "biased." However, bias need only have a negative connotation if there is a reason to presume that one condition of judgment is more accurate than all others. A broader view is to accept the idea that all judgments are a function of observing conditions and therefore all judgments are biased from one another in different ways. Fortunately, many of these biases and the conditions that cause them are predictable and well understood, and so they can be eliminated or minimized. At the very least, the sensory practitioner needs to understand how these influences operate so as to know when to expect changes in judgments and ratings. An important endpoint is the realization that few, if any, ratings have any absolute meaning. You cannot say that because a product received a hedonic rating of 7.0 today, it is better than the product that received a rating of 6.5 last week. The context may have changed.

9.2 Simple Contrast Effects

By far the most common effect of sensory context is simple contrast. Any stimulus will be judged as more intense in the presence of a weaker stimulus and as less intense in the presence of a stronger stimulus, all other conditions being equal. This effect is much easier to find and to demonstrate than its opposite, convergence or assimilation. For example, an early sensory worker at the Quartermaster Corp., Kamenetsky (1957), noticed that the acceptability ratings for foods seemed to depend upon what other foods were presented during an evaluation session. Poor foods seemed even worse when preceded by a good sample. Convergence is more difficult to demonstrate, although under some conditions a group of items may seem more similar to each other when they are in the presence of an item that is very different from that group (Zellner et al., 2006).

9.2.1 A Little Theory: Adaptation Level

As we noted above, a 40° day in January (in New York) seems a lot warmer than the same temperature in August. These kinds of effects are predicted Helson's theory of adaptation level. Helson (1964) proposed that we take as a frame of reference the average level of stimulation that has preceded the item to be evaluated. The mild temperature in the middle of a hot and humid summer seems a lot more cool and refreshing than is the mild temperature after a period of cold and icy weather. So we refer to our most recent experiences in evaluating the sensory properties of an item. Helson went on to elaborate the theory to include both immediate and distant predecessors. That is, he appreciated the fact that more recent items tend to have a stronger effect on the adaptation level. Of course, mere reference to the mean value of experience is not always sufficient to induce a contrast effect—it is more influential if the mean value comes to be centered near the middle of the response scale, an example of a centering bias, discussed below (Poulton, 1989).

The notion of adaptation, a decrease in responsiveness under conditions of constant stimulation, is a major theme in the literature on sensory processes. Physiological adaptation or an adjustment to the ambient level of stimulation is obvious in light/dark adaptation in vision. The thermal and tactile senses also show profound adaptation effects—we become easily adjusted to the ambient room temperature (as long as it is not too extreme) and we become unaware of tactile stimulation from our clothing. So this mean reference level often passes from consciousness or becomes a new baseline from which deviations in the environment become noticeable. Some workers have even suggested that this improves discrimination—that the difference threshold is smallest right around the adaptation level or physiological zero, in keeping with Weber's law (McBurney, 1966). Examples of adaptation effects are discussed in Chapter 2 for the senses of taste and smell. In the chemical, thermal, and tactile senses, adaptation is quite profound.

However, we need not invoke the concept of neural adaptation to a preceding item or a physiological effect to explain all contrast effects. It may be simply that more or less extreme stimuli change our frame of reference or the way in which the stimulus range and response scales are to be mapped onto one another. The general principle of context is that human observers act like measuring instruments that constantly recalibrate themselves to the experienced frame of reference. What we think of as a small horse may depend upon whether the frame of reference includes Clydesdales, Shetland ponies, or tiny prehistoric equine species. The

following examples show simple effects of context on intensity, sensory quality, and hedonics or acceptability. Most of these examples are cases of perceptual contrast or a shift in judgment *away from other stimuli* presented in the same session.

9.2.2 Intensity Shifts

Figure 9.3 shows a simple contrast effect of soups with varying salt levels presented in different contexts (Lawless, 1983). The central stimulus in the series was presented either with two lower or two higher concentrations of salt added to a low sodium soup. Ratings of saltiness intensity were made on a simple nine-point category scale. In the lower context, the central soup received a higher rating, and in the higher context, it received a lower rating, analogous to our perception of a mild day in winter (seemingly warmer) versus a mild day in summer (seemingly cooler). Note that the shift is quite dramatic, about two points on the nine-point scale or close to 25% of scale range.

A simple classroom demonstration can show a similar shift for the tactile roughness of sandpapers varying in grit size. In the context of a rougher sample, a medium sample will be rated lower than it is in the context of a smoother sample. The effects of simple contrast are not limited to taste and smell.

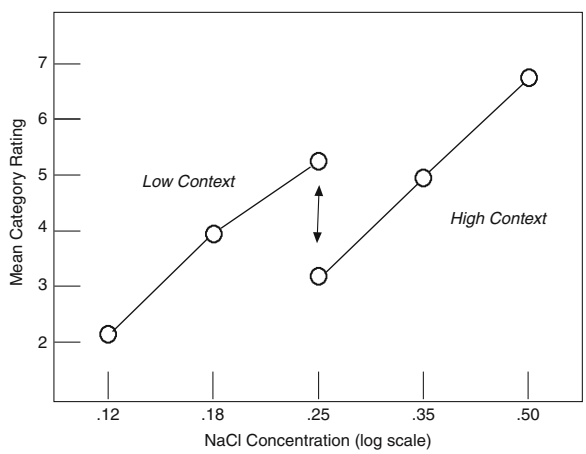

Fig. 9.3 Saltiness ratings of soups with added NaCl. The sample at 0.25 M was evaluated in two contexts, one with higher concentrations and one with lower concentrations. The shift is typical of a simple contrast effect of contrast. Replotted from Lawless (1983). Copyright ASTM, used with permission.

Contrast effects are not always observed. In some psychophysical work with long series of stimuli, some item-to-item correlations have been observed. The effects of immediately preceding versus remotely preceding stimuli have been measured and a positive correlation among adjacent responses in the series was found. This can be taken as evidence for a type of assimilation, or underestimation of differences (Ward, 1979, 1987; but see also Schifferstein and Frijters, 1992).

9.2.3 Quality Shifts

Visual examples such as color contrast were well known to early psychologists like William James: "Meanwhile it is an undoubted general fact that the psychical effect of incoming currents does depend on what other currents may be simultaneously pouring in. Not only the perceptibility of the object which the current brings before the mind, but the quality of it is changed by the other currents." (1913, p. 25). A gray line against a yellow background may appear somewhat bluish, and the same line against a blue background may seem more yellowish. Paintings of the renowned artist Josef Albers made excellent use of color contrast. Similar effects can be observed for the chemical senses. During a descriptive panel training period for fragrance evaluation, the terpene aroma compound dihydromyrcenol was presented among a set of woody or pine-like reference materials. The panelists complained that the aroma was too citrus-like to be included among the woody reference materials. However, when the same odor was placed in the context of citrus reference materials, the same panelists claimed that it was far too woody and pine-like to be included among the citrus examples. This contextual shift is shown in Fig. 9.4. In a citrus context, the item is rated as more woody in character than when placed in a woody context. Conversely, ratings for citrus intensity decrease in the citrus context and increase in the woody context. The effect is quite robust and is seen whether or not a rest period is included to undo the potential effects of sensory adaptation. It even occurs when the contextual odor follows the target item and judgments are made after both are experienced (Lawless et al., 1991)! This striking effect is discussed further in Section 9.2.5.

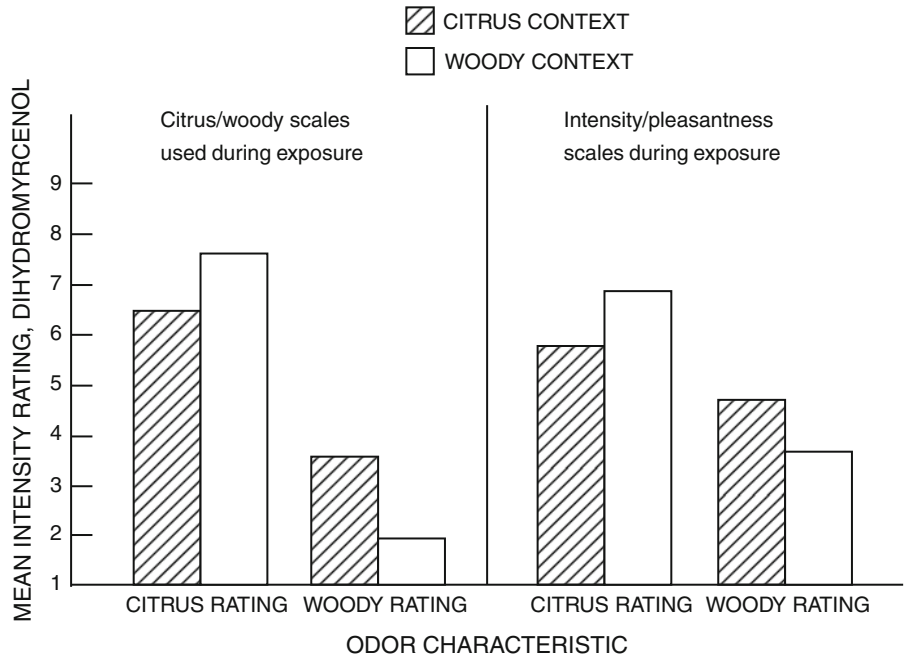

Fig. 9.4 Odor quality contrast noted for the ambiguous terpene aroma compound dihydromyrcenol. In a citrus context, woody ratings increase and citrus character decreases. In a woody context, the woody ratings decrease. The group using different scales did not rate citrus and woody character during the contextual exposure phase, only overall intensity and pleasantness were rated. From Lawless et al. (1991) by permission of Oxford University Press.

Context effects can also alter how items are identified and characterized. When people categorize speech sounds, repeated exposure to one type of simple phoneme changes the category boundary for other speech sounds. Repeated exposure to the sound of the phoneme "bah," which has an early voice onset time, can shift the phoneme boundary so that speech sounds near the boundary are more likely classified as "pah" sounds (a later voice onset) (Eimas and Corbit, 1973). Boundary-level examples are shifted across the boundary and into the next category. This shift resembles a kind of contrast effect.

9.2.4 Hedonic Shifts

Changes in the preference or acceptance of foods can be seen as a function of context. Hedonic contrast was a well-known effect to early workers in food acceptance testing (Hanson et al., 1955; Kamenetzky, 1959). An item seems more appealing if it followed an item of poor quality and less appealing if it followed something of better quality. The effect was known to Beebe-Center (1932), who also attributed it to Fechner in 1898. This kind of contrast has been observed for tastes (Riskey et al., 1979; Schifferstein, 1995), odors (Sandusky and Parducci, 1965), and art (Dolese et al., 2005). Another effect observed in these kinds of experiments is that a contrasting item causes other, generally lower rated stimuli to become more similar or less discriminable, an effect termed condensation (Parker et al., 2002; Zellner et al., 2006;). In the study by Zellner et al. (2006), pre-exposure to a good-tasting juice reduced the magnitude of preference ratings among less appealing juices. Mediocre items were both worse and more similar.

An example of hedonic shifting was found in a study on the optimization of the saltiness of tomato juice and also the sweetness of a fruit beverage using the method of adjustment (Mattes and Lawless, 1985). When trying to optimize the level of sweetness or saltiness in this study, subjects worked in two directions. In an ascending series, they would concentrate a dilute solution by mixing the beverage with a more concentrated

version having the same color, aroma, and other flavor materials (i.e., only sweetness or saltiness was different). In a second descending series, they would be given a very intense sample as the starting point and then dilute down to their preferred level. This effect is shown in Fig. 9.5. The adjustment stops too soon and the discrepancy is remarkable, nearly a concentration range of 2:1. The effect was also robust—it could not be attributed to sensory adaptation or lack of discrimination and persisted even when subjects were financially motivated to try to achieve the same endpoints in both trials. This is a case of affective contrast. When compared to a very sweet or salty starting point, a somewhat lower item seems just about right, but when starting with a relatively sour fruit beverage or bland tomato juice, just a little bit of sugar or salt helps quite a bit. The stopping point contrasts with the starting material and seems to be better than it would be perceived in isolation. In an ascending or descending sequence of products, a change in responses that happens too soon is called an "error of anticipation."

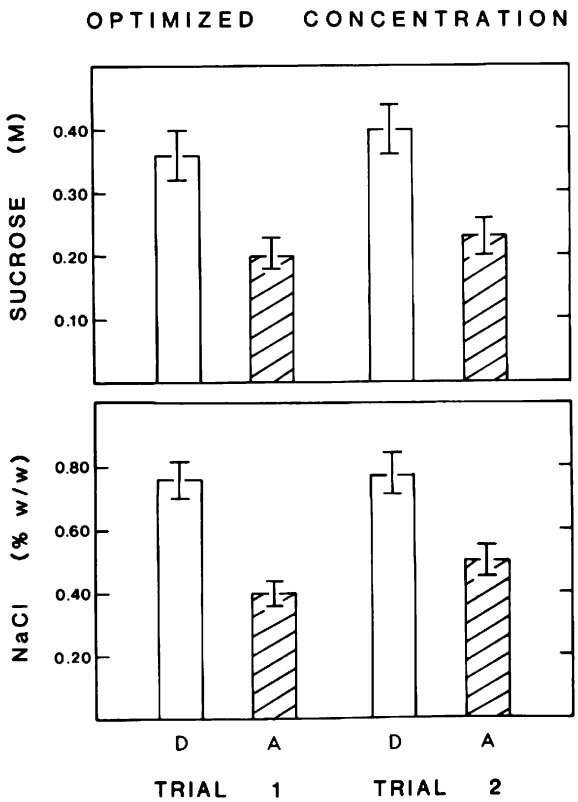

Fig. 9.5 Optimized concentrations of salt in tomato juice and sucrose in a fruit beverage. In the trials labeled D, the concentration was diluted from a concentrated version of the test sample. In the trials marked A, the concentration was increased from a dilute version of the test sample. Concentrations of other ingredients were held constant. The contextual shift is consistent with reaching the apparent optimum too soon as if the apparent optimum was shifted in contrast to the starting point. From Mattes and Lawless (1985) with permission.

9.2.5 Explanations for Contrast

At first glance, one is tempted to seek a physiological explanation for contrast effects, rather than a psychological or a judgmental one. Certainly sensory adaptation to a series of intense stimuli would cause any subsequent test item to be rated much lower. The importance of sensory adaptation in the chemical senses of taste and smell lends some credence to this explanation. However, a number of studies have shown that precautions against sensory adaptation may be taken, such as sufficient rinsing or time delays between stimuli and yet the context effects persist (Lawless et al., 1991; Mattes and Lawless, 1985; Riskey, 1982). Furthermore, it is difficult to see how sensory adaptation to low-intensity stimuli would cause an increase in the ratings for a stronger item, as adaptation necessarily causes a decrement in physiological responsiveness compared to a no-stimulation baseline.

Perhaps the best evidence against a simple adaptation explanation for contrast effects is from the *reversed-pair* experiments in which the contextual item follows the to-be-rated target item and therefore can have no physiologically adapting effect on it. This paradigm calls for a judgment of the target item from memory after the presentation of the contextual item, in what has been termed a reversed-pair procedure (Diehl et al., 1978). Due to the reversed order, the context effects cannot be blamed on physiological adaptation of receptors, since the contextual item follows rather than precedes the item to be rated. Reversed-pair effects are seen for shifts in odor quality of aroma compounds like dihydromyrcenol and are only slightly smaller in magnitude than the contextual shift caused when the contextual item comes first (Lawless et al., 1991). The reversed-pair situation is also quite capable of causing simple contrast effects in sensory intensity. A sweetness shift was observed when a higher or a lower sweetness item was interpolated between the tasting and rating (from memory)

of a normal-strength fruit beverage (Lawless, 1994). Looking back at Fig. 9.2, it seems more likely that the effect changes the response function. However, not all workers in the field agree. In particular, Marks (1994) has argued that the contextual shifts are much like an adaptation process and that for auditory stimuli this is a peripheral event. It is possible that what changes is not the sensation/experience, but to some encoded version of the sensation, or to some kind of implicit response, not yet verbalized. If a person, when rating, is evaluating some memory trace of the experience, it is possible that this memory, for example, could be altered.

9.3 Range and Frequency Effects

Two of the most common factors that can affect ratings are the sensory range of the products to be evaluated and the frequency with which people use the available response options. These factors were nicely integrated into a theory that helped to explain shifts in category ratings. They are also general tendencies that can affect just about any ratings or responses.

9.3.1 A Little More Theory: Parducci's Range and Frequency Principles

Parducci (1965, 1974) sought to go beyond Helson's (1964) simple idea that people respond to the mean or the average of their sensory experiences in determining the frame of reference for judgment. Instead, they asserted that the entire distribution of items in a psychophysical experiment would influence the judgments of a particular stimulus. If this distribution was denser (bunched up) at the low ends and a lot of weak items were presented, product ratings would shift up. Parducci (1965, 1974) proposed that behavior in a rating task was a compromise between two principles. The first was the range principle. Subjects use the categories to sub-divide the available scale range and will tend to divide the scale into equal perceptual segments. The second was the frequency principle. Over many judgments, people like to use the categories an equal number of times (Parducci, 1974). Thus it is not only the average level that is important but also how stimuli may be grouped or spaced along the continuum that would determine how the response scale was used. Category scaling behavior could be predicted as a compromise between the effects of the range and frequency principles (Parducci and Perrett, 1971).

9.3.2 Range Effects

The range effect has been known for some time, both in category ratings and other judgments including ratio scaling (Engen and Levy, 1958; Teghtsoonian and Teghtsoonian, 1978). When expanding or shrinking the overall range of products, subjects will map their experiences onto the available categories (Poulton, 1989). Thus short ranges produce steep psychophysical functions and wide ranges produce flatter functions. An example of this can be seen in two published experiments on rating scales (Lawless and Malone, 1986a, b). In these studies, four types of response scales and a number of visual, tactile, and olfactory continua were used to compare the abilities of consumers to use the different scales to differentiate products. In the first study, the consumers had no trouble in differentiating the products and so in the second study, the stimuli were spaced more closely on the physical continua so that the task would be more challenging. However, when the experimenters closed the stimuli in, the range principle took over, and participants used more of the rating scale than expected. In Fig. 9.6, ratings for perceived thickness of (stirred) silicone samples are shown in the wide and narrow stimulus ranges. Note the steepening of the response function. For the same one log unit change in physical viscosity, the range of responses actually doubled from the wide range to the narrower range.

Another kind of stimulus range effect occurs with anchor stimuli. Sarris (1967) previously showed a strong effect of anchor stimuli on the use of rating scales, unless the anchors were very extreme, at which point their influence would tend to diminish, as if they had become irrelevant to the judgmental frame of reference. Sarris and Parducci (1978) found similar effects of both single and multiple end anchors that generally take the form of a contrast effect. For example, a low anchor stimulus, whether rated or unrated, will cause stronger stimuli to receive higher ratings than

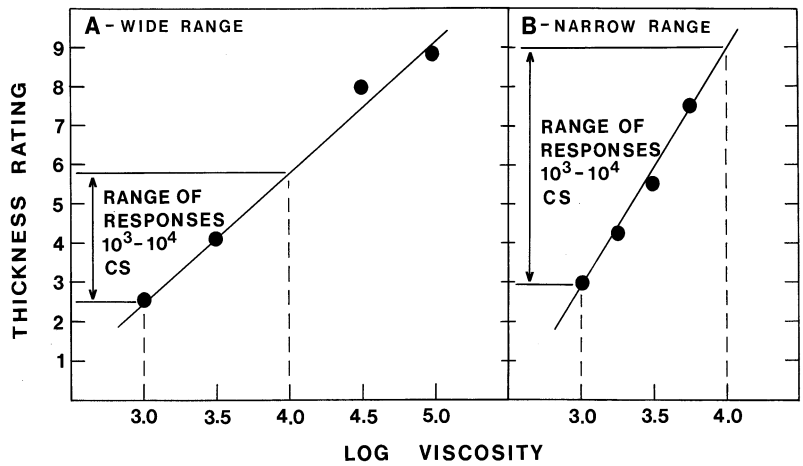

Fig. 9.6 A simple range effect. When products are presented over a wide range, a shallow psychophysical function is found. Over a narrow range, a steeper psychophysical function will be observed. This is in part due to the tendency of subjects to map the products (once known) onto the available scale range. From Lawless and Malone (1986b), with permission.

they would if no anchor were presented, unless the "anchor" is so extreme as to seem irrelevant. Sarris and Parducci (1978) provided the following analogy: A salesman will consider his commissions to be larger if coworkers receive less than he does. However, he is unlikely to extend or shift his scale of judgment by hearing about others who are in an entirely different bracket of income (1978, p. 39). Whether an outlying product is similar enough to have an influence (or whether it is a "horse of a different color") should be of concern in product comparisons of diverse items.

9.3.3 Frequency Effects

The frequency effect is the tendency of people to try to use the available response options about the same number of times across a series of products or stimuli to be rated. The frequency effect can cause shifts that look like simple contrast and also a local steepening of the psychophysical function around points where stimuli were closely spaced or very numerous (compared with less "dense" portions of the stimulus range). The frequency principle dictates that when judging many samples, products that are numerous or bunched at the low or high ends of the distributions tend to be spread out into neighboring categories. This is illustrated in the two panels of Fig. 9.7. The upper panel shows four hypothetical experiments and how products might be bunched in different parts of the range. In the upper left panel, we see how a normal replicated psychophysical experiment would be conducted with equal presentations of each stimulus level. The common outcome of such a study using category ratings would be a simple linear function of the log of stimulus intensity. However, if the stimulus presentations were more frequent at the high end of the distribution, i.e., negative skew, the upper categories would be overused, and subjects would begin to distribute their judgments into lower categories. If the samples were bunched at the lower end, the lower response categories would be overused and subjects would begin to move into higher categories. If the stimuli were bunched in the midrange, the adjacent categories would be used to take on some of the middle stimuli, pushing extreme stimuli into the ends of the response range, as shown in the panel for a quasi-normal distribution.

Such behavior is relevant to applied testing situations. For example, in rating the presence of off-flavors or taints, there may be very few examples of items with high values on the scale and lots of weak (or zero) sensations. The frequency effect may explain why the low end of the scale is used less often than anticipated, and higher mean values are obtained than one would deem appropriate. Another example is screening a number of flavor or fragrance candidates for a new product. A large number of good candidates are sent for testing by suppliers or a flavor development group. Presumably these have been pre-tested or at least have received a favorable opinion from a flavorist or a perfumer. Why do they then get only mediocre ratings from the test panel? The high end of the distribution is overrepresented (justifiably so and perhaps on purpose), so the tendency for the panel is to drop into lower categories. This may partly explain why in-house testing

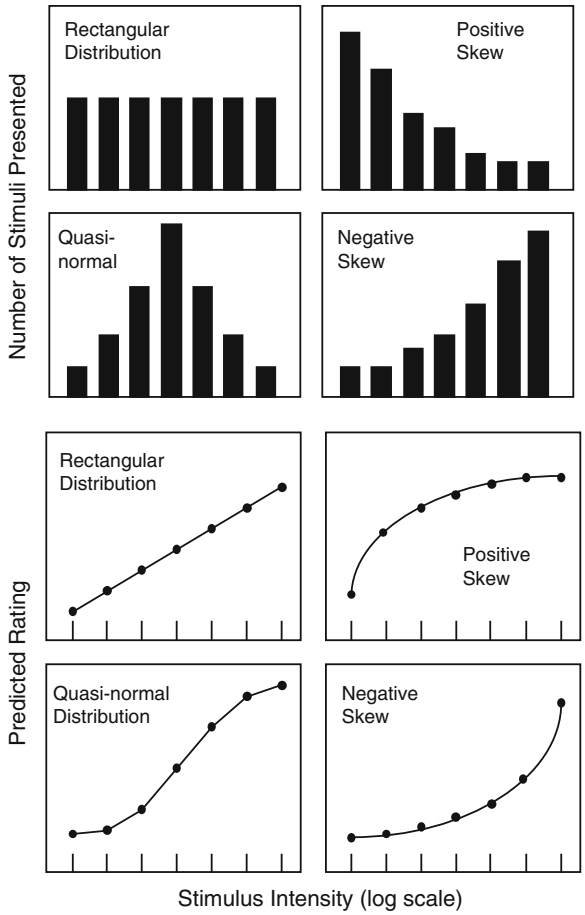

Fig. 9.7 Predictions from the Parducci range–frequency theory. Distributions of stimuli that are concentrated at one part of the perceptual range (*upper quartet*) will show local steepening of the psychophysical functions (*lower quartet*). This is due to subjects' tendencies to use categories with equal frequency, the resulting shifting into adjacent categories from those that are overused.

panels are sometimes more critical or negative than consumers when evaluating the same items.

Although the great majority of experiments on the range and frequency effects have been performed with simple visual stimuli, there are also examples from taste evaluation (Lee et al., 2001; Riskey et al., 1979; Riskey, 1982; Schifferstein and Frijters, 1992; Vollmecke, 1987). Schifferstein and Frijters found similar effects of skewed distributions with line-marking responses as seen in previous studies with category ratings. Perhaps line marking is not a response scale with infinite divisions, but panelists sub-divide the line into discrete sub-ranges as if they were using a limited number of zones or categories. The effect of grouping or spacing products also intensifies as the exposure to the distributions increases. Lawless (1983) showed that the shift that occurred with a negative skew (bunching at the upper end) into lower response categories would intensify as the exposure to the skewed distribution went from none to a single exposure to three exposures. Thus the contextual effects do not suddenly appear but will take hold of the subjects' behavior as they gain experience with the sample set.

9.4 Biases

9.4.1 *Idiosyncratic Scale Usage and Number Bias*

People appear to have preferred ranges or numbers on the response scale that they feel comfortable using. Giovanni and Pangborn (1983) noted that people using magnitude estimation very often used numbers that were multiples of 2 and 5 (or obviously 10), an effect that is well known in the psychophysical literature (Baird and Noma, 1978). With magnitude estimation, the idiosyncratic usage of a favorite range of numbers causes a correlation of the power function exponents across different sensory continua for an individual (Jones and Marcus, 1961; Jones and Woskow, 1966). This correlation can be explained if people are more or less expansive (versus restrictive) in their response output functions, i.e., in how they apply numbers to their sensations in magnitude estimation studies. Another version of such personal idiosyncrasy is the common observation in time–intensity scaling that people show a kind of personal "signature" or a characteristic curve shape (Dijksterhuis, 1993; McGowan and Lee, 2006).

Another version of self-induced response restriction can be seen when people use only selected portions of the scale in a line-marking rating task. On a line scale with verbal labels, people may choose to make markings only near the verbal labels, rather than distributing them across the response scale. This was first observed by Eng (1948) with a simple hedonic line scale labeled Like Very Highly at one end, Dislike Very Highly at the other, and Neither Like nor Dislike at the center. In a group of 40 consumers, 24 used only the three labeled parts of the scale, and Eng deleted them

from the data analysis! This kind of behavior was also noted with the labeled affective magnitude scale (LAM scale) by Cardello et al. (2008) with both Army laboratory and student groups. Lawless et al. (2010a) found a very high frequency (sometimes above 80%) of people making marks within ±2 mm of a phrase mark on the LAM scale in a multi-city consumer central location test. Lawless et al. (2010b) found that instructions did not seem to change this behavior much but that expanding the physical size of the scale on the ballot (from about 120 to 200 mm) decreased the "categorical" behavior somewhat. Categorical rating behavior can also be seen as a step function in time–intensity records (rather than a smooth continuous curve).

Finding product differences against the background of these individual response tendencies can be facilitated by within-subject experimental designs. Each participant is used as his or her own baseline in comparisons of products, as in dependent *t*-tests or repeated measure analysis of variance in complete block designs. Another approach is to compute a difference score for a comparison of products in each individual's data, rather than merely averaging across people and looking at differences between mean values.

9.4.2 Poulton's Classifications

Poulton (1989) published extensively on biases in ratings and classified them. Biases in Poulton's system go beyond Parducci's theory but are documented in the psychophysical literature. These include centering biases, contraction biases, logarithmic response bias with numerical ratings, and a general transfer bias that is seen when subjects carry the context from a previous session or study into a new experiment. The centering bias is especially relevant to just-right scales and is discussed in a later section. The response range bias is also a special case and follows this section.

The contraction biases are all forms of assimilation, the opposite of contrast. According to Poulton, people may rate a stimulus relative to a reference or a mean value that they hold in memory for similar types of sensory events. They tend to judge new items as being close (perhaps too close) to this reference value, causing underestimation of high values and overestimation of low values. There may also be overestimation of an item when it follows a stronger standard stimulus or underestimation when it follows a weaker standard stimulus, a sort of local contraction effect. Poulton also classifies the tendency to gravitate toward the middle of the response range as a type of contraction effect, called a response contraction bias. While all of these effects undoubtedly occur, the question arises as to whether contrast or assimilation is a more common and potent process in human sensory judgment. While some evidence for response assimilation has been found in psychophysical experiments through sequential analysis of response correlations (Ward, 1979), contrast seems much more to be the rule with taste stimuli (Schifferstein and Frijters, 1992) and foods (Kamenetzky, 1959). In our experience, assimilation effects are not as prevalent as contrast effects, although assimilation has certainly been observed in experiments on consumer expectation (e.g., Cardello and Sawyer, 1992). In that case, the assimilation is not toward other actual stimuli but toward expected levels.

The logarithmic response bias can be observed with open-ended response scales that use numbers, such as magnitude estimation. There are several ways to view this type of bias. Suppose that a series of stimuli have been arranged in increasing magnitude and they are spaced in subjectively equal steps. As the intensity increases, subjects change their strategy as they cross into ranges of numerical responses where there are more digits. For example, they might be rating the series using numbers like 2, 4, 6, and 8, but then when they get to 10, they will continue by larger steps, perhaps 20, 40, 60, 80. In Poulton's view they proceed through the larger numerical responses "too rapidly." A converse way of looking at this problem is that the perceived magnitude of the higher numbers is in smaller arithmetic steps as numbers get larger. For example, the difference between one and two seems much larger compared to the difference between 91 and 92. Poulton also points out that in addition to contraction of stimulus magnitude at very high levels, the converse is also operating and that people seem to illogically expand their subjective number range when using responses smaller than the number 3. One obvious way to avoid the problems in number bias is to avoid numbers altogether or to substitute line scaling or cross-modality matching to line length as a response instead of numerical rating techniques like magnitude estimation (Poulton, 1989).

Transfer bias refers to the general tendency to use previous experimental situations and remembered judgments to calibrate oneself for later tasks. It may involve any of the biases in Poulton's or Parducci's theories. The situation is common when subjects are used in multiple experiments or when sensory panelists are used repeatedly in evaluations (Ward, 1987). People have memories and a desire to be internally consistent. Thus the ratings given to a product on one occasion may be influenced by ratings given to similar products on previous occasions. There are two ways to view this tendency. One is that the judgments may not shift appropriately when the panelists' sensory experience, perception, or opinion of the product has in fact changed. On the other hand, one of the primary functions of panelist training and calibration in descriptive analysis is to build up exactly those sorts of memory references that may stabilize sensory judgments. So there is a positive light to this tendency as well. An open question for sensory evaluation is whether exposure to one continuum of sensory intensities or one type of product will transfer contextual effects to another sensory attribute or a related set of products (Murphy, 1982; Parducci et al., 1976; Rankin and Marks, 1991). And if so, how far does the transfer extend?

9.4.3 Response Range Effects

One of Poulton's biases was called the "response range equalizing bias" in which the stimulus range is held constant but the response range changes and so do the ratings. Ratings expand or contract so that the entire range is used (minus any end-category avoidance). This is consistent with the "mapping" idea mentioned for stimulus range effects (stimuli are mapped onto the available response range). Range stabilization is implicit in the way some scaling studies have been set up and in the instructions given to subjects. This is similar to the use of physical reference standards in some descriptive analysis training (Muñoz and Civille, 1998) and is related to Sarris's work on anchor stimuli (Sarris and Parducci, 1978). In Anderson's work with 20-point category scales and line marking, high and low examples or end anchors are given to subjects to show them the likely range of the stimuli to be encountered. The range of responses is known since it is visible upon the page of the response sheet or has been pre-familiarized in a practice session (Anderson, 1974). Thus it is not surprising that subjects distribute their responses across the range in a nicely graded fashion, giving the appearance that there is a reasonably linear use of the scale. Anderson noted that there are end effects that work against the use of the entire range (i.e., people tend to avoid using the endpoints) but that these can be avoided by indenting the response marks for the stimulus end anchors, for example, at points 4 and 16 on the 20-point category scale. This will provide psychological insulation against the end effects by providing a comfort zone for unexpected or extreme stimuli at the ends of the scale while leaving sufficient gradations and room to move within the interior points. The "comfort zone" idea is one reason why early workers in descriptive analysis used line scales with indented vertical marks under the anchor phrases.

An exception to the response range mapping rule is seen when anchor phrases or words on a scale are noted and taken seriously by participants. An example is in Green's work on the labeled magnitude scale, which showed a smaller response range when it was anchored to "greatest imaginable sensation" that included all oral sensations including pain, as opposed to a wider range when the greatest imaginable referred only to taste (Green et al., 1996). This also looks like an example of contrast in which the high-end anchor can evoke a kind of stimulus context, at least in the participant's mind. If the image evoked by the high-end phrase is very extreme, it acts like a kind of stimulus that compresses ratings into a smaller range of the scale. A similar kind of response compression was seen with the LAM scale when it was anchored to greatest imaginable liking for "sensations of any kind" as opposed to a more delimited frame such as "foods and beverages" (Cardello et al., 2008). A sensory scientist should consider how the high anchor phrase is interpreted, especially if he or she wants to avoid any compression of ratings along the response range. As Muñoz and Civille (1998) pointed out, the use of a descriptive analysis scale also depends a lot on the conceptualization of the high extreme. Does "extremely strong" refer to the strongest possible taste among all sensations and products, the strongest sensation in this product type, or just how strong this particular attribute can become in this particular product? The strongest sweetness in this product might be more intense than the strongest saltiness. The definition needs to be a

deliberate choice of the panel leader and an explicit instruction to the panelist to give them a uniform frame of reference.

9.4.4 The Centering Bias

The centering bias arises when subjects become aware of the general level of stimulus intensity they are likely to encounter in an experiment and tend to match the center or midpoint of the stimulus range with the midpoint of the response scale. Poulton (1989) distinguished a stimulus centering bias from a response centering bias, but this distinction is primarily a function of how experiments are set up. In both cases, people tend to map the middle of the stimulus range onto the middle of the response range and otherwise ignore the anchoring implications of the verbal labels on the response scale. Note that the centering bias works against the notion that respondents can use unbalanced scales with any effectiveness. For example, the "Excellent–very good–good–fair–poor" scale commonly used in marketing research with consumers is unbalanced. The problem with unbalanced scales is that over many trials, the respondents will come to center their responses on the middle category, regardless of its verbal label.

The centering bias is an important problem when there is a need to interpolate some value on a psychophysical function or to find an optimal product in just-right scaling. Poulton gives the example of McBride's method for considering bias in the just-about-right (JAR) scale (McBride, 1982; see also Johnson and Vickers, 1987). In any series of products to be tested, say for just-right level of sweetness, there is a tendency to center the series so that the middle product will come out closest to the just-right point. The function shifts depending upon the range that is tested. One way to find the true just-right point would be to actually have the experimental series centered on that value, but then of course you would not need to do the experiment. McBride gives a method for interpolation across several experiments with different ranges. The point at which the just-right function and the median of the stimulus series will cross shows the unbiased or true just-right level. This method of interpolation is shown in Fig. 9.8. In this method, you present several ranges of the products in separate sessions and plot how the judgments of the JAR point shift up and down. You can then interpolate to find the range in which the just-right point would have been from the center product in the series. This obviously takes more work to do the test a couple of times, but it could avoid a mistaken estimate of the JAR level.

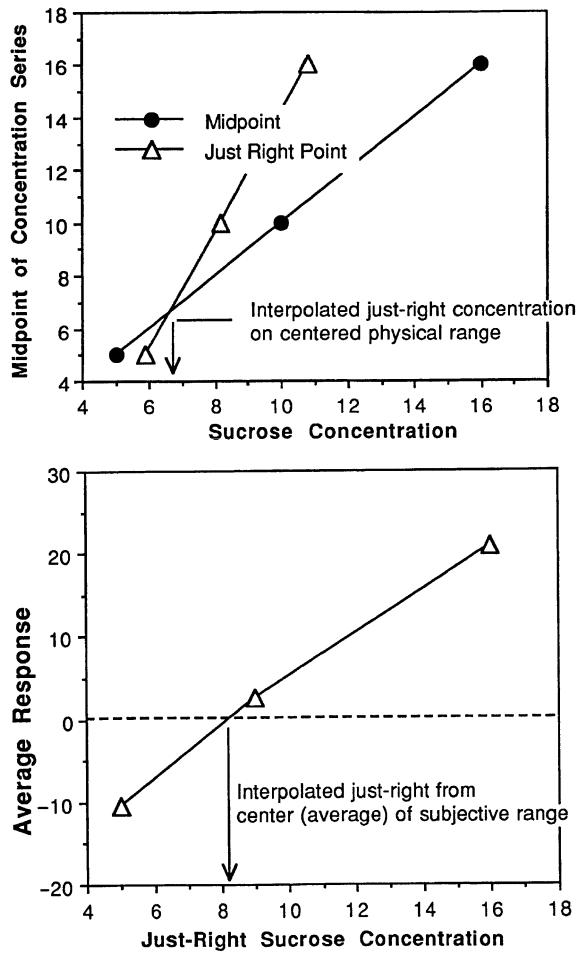

Fig. 9.8 Adjusting for the centering bias in just-right ratings. Three series of sucrose concentrations in lemonade were tested, a low series (2–8%), a middle series (6–14%), and a high range (10–22%). In the *upper panel*, the method of Poulton is used to interpolate the unbiased just-right point from the series where the midpoint concentration would correspond to the just-right item. In the *lower panel*, the method of McBride is used to interpolate the just-right point from a series in which the average response would correspond to the just-right point. When the average response would be just right (zero on this scale), the hypothetical stimulus range would have been centered on the just-right level. Replotted from Johnson and Vickers (1987), with permission.

9.5 Response Correlation and Response Restriction

Early experimental psychologists like Thorndike (1920) noted that one very positive attribute of a person could influence judgments on other, seemingly unrelated characteristics of that individual. In personnel evaluations of military officers, Thorndike noted a moderate positive correlation among the individual rated factors. People evaluate others like this in real life. If achievement in sports is influential in our assessment of a person, we might suppose a gifted athlete to also be kind to children, generous to charities, etc., even though there is no logical relationship between these characteristics. People like to have cognitive structures that form consistent wholes and are without conflicts or contradictions (called cognitive dissonance) that can make us uncomfortable. The halo effect has also been described as a carry-over from one positive product to another (Amerine et al., 1965), but its common usage is in reference to a positive correlation of unrelated attributes (Clark and Lawless, 1994). Of course, there can also be negative or horns effects, in which one salient negative attribute causes other, unrelated attributes to be viewed or rated negatively. If a product makes a mess in the microwave, it might be rated negatively for flavor, appearance, and texture as well.

9.5.1 Response Correlation

A simple example of a halo effect is shown in Fig. 9.9. In this case, a small amount of vanilla extract was added to low-fat milk, near the threshold of perceptibility. Ratings were then collected from 19 milk consumers for sweetness, thickness, creaminess and liking for the spiked sample, and for a control milk. In spite of the lack of relationship between vanilla aroma and sweet taste and between vanilla and texture characteristics, the introduction of this one positive aspect was sufficient to cause apparent enhancement in sweetness, creaminess, and thickness ratings.

Apparent enhancement of sweetness is an effect long known for ethyl maltol, a caramelization product that has an odor similar to heated sugar (Bingham et al., 1990). When maltol is added to various products, sweetness ratings may rise compared to products lacking this flavor. However, the effect seems to be a case of the misattribution of olfactory stimulation to the taste sense. Murphy and Cain (1980) showed that citral (a lemon odor) could enhance taste ratings, but only when the nostrils were open, which allows diffusion of the odor into the nose and stimulation of the olfactory receptors (i.e., retronasal smell). When the nostrils are pinched shut, the diffusion is effectively eliminated and the enhancement disappears. Murphy and Cain interpreted this as convincing evidence that there was no true enhancement of taste intensity by citral, but only olfactory referral, a kind of confusion between taste and smell. Studies with other odors have also shown that the taste enhancement effect from volatile flavors can be eliminated by nose pinching (Frank and Byram, 1988) even for maltol (Bingham et al., 1990). The maltol effect is also minimized by training subjects who then learn to more effectively separate or localize their odor experiences from taste (Bingham et al., 1990). The sweetness enhancement may arise as a function of conditioning or experience with the pairing of sweet tastes with some odors in foods (Stevenson et al., 1995).

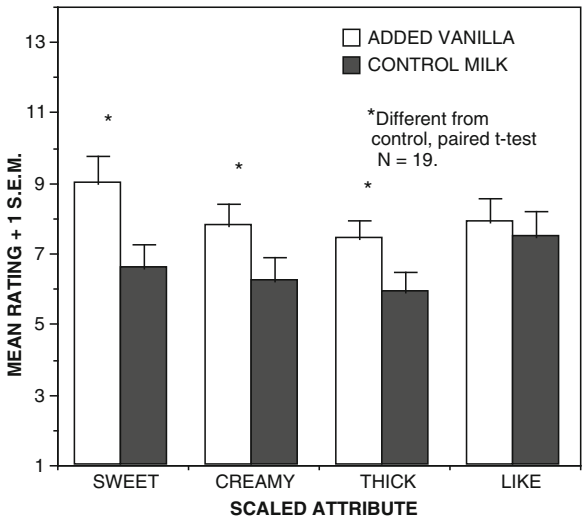

Fig. 9.9 Adding a just perceivable level of vanilla extract to low-fat milk causes increases in rated sweetness, thickness, creaminess, and liking, an example of the Halo effect. From Lawless and Clark (1994), with permission.

Several lessons can be learned from the vanilla halo effect shown in Fig. 9.9. First, untrained consumers

cannot be trusted to provide accurate sensory specifications of product characteristics. While it may be common practice to collect some diagnostic attribute ratings from consumers in central location or home use tests, such information must be viewed with caution. There are well-known correlations among attributes (easily shown by principal components analysis, see Chapter 18), halo effects, and taste–smell confusions, all of which can bias the obtained ratings. Second, consumers see products as Gestalten, as whole patterns. They do not act analytically in short product tests. They do not learn to separate their sensations and attend effectively to individual product attributes. Third, if consumers do not have a chance to comment on a salient product characteristic, they may find some other place on the questionnaire to voice that feeling, perhaps in an inappropriate place. This last tendency was taken advantage of in our milk example—note that no scale for vanilla flavor was provided. The effect of misusing response scales in this way on a questionnaire is called response restriction or simply the "dumping effect."

9.5.2 "Dumping" Effects: Inflation Due to Response Restriction in Profiling

It is part of the folklore of consumer testing that if there is one very negative and salient attribute of a product, it will influence other attributes in a negative direction, an example of a horns effect. The effect is even worse when the salient negative attribute is omitted from the questionnaire. Omission could be due to some oversight or failure to anticipate the outcome in a consumer test or simply that it was not observed in the laboratory conditions of preliminary phases of testing. In this case, consumers will find a way to dump their frustration from not being able to report their dissatisfaction by giving negative ratings on other scales or reporting negative opinions of other even unrelated attributes. In other words, restricting responses or failure to ask a relevant question may change ratings on a number of other scales.

A common version of this restriction effect can be seen in sweetness enhancement. Frank et al. (1993) found that the enhancement of sweet ratings in the presence of a fruity odor was stronger when ratings were restricted to sweetness only. When both sweetness and fruitiness ratings were allowed, no enhancement of sweetness was observed. Exactly the same effect was seen for sweetness and fruitiness ratings and for sweetness and vanilla ratings (Clark and Lawless, 1994). So allowing the appropriate number of attributes can address the problem of illusory enhancement. Schifferstein (1996) gave the example of hexenol, a fresh green aroma, which when added to a strawberry flavor mixture caused mean ratings in several other scales to increase. The enhancement of the other ratings occurred only when the "green" attribute was omitted from the ballot. When the "green" attribute was included in the ballot, the response was correctly assigned to that scale, and there was no apparent enhancement in the other attributes in the aroma profile.

There is good news and bad news in these observations. From a marketing perspective, ratings can be easily obtained from consumers that will show apparent sweetness enhancements if the questionnaires cleverly omit the opportunity to report on sensations other than sweetness. However, the nose pinch conditions and the use of complete sets of attributes show us that these volatile odorants such as maltol are not sweet taste enhancers but they are sweet *rating* enhancers. That is, they are not affecting the actual perception of sweet taste intensity but are changing the response output function or perhaps broadening the concept of sweetness to go beyond taste and include pleasant aromas as well. It would not be wise to try to use maltol to sweeten your coffee.

Are there other situations in sensory testing where the dumping effect can show up? One area in which responses are usually restricted to one attribute at a time is in time–intensity scaling (Chapter 8). In a common version of this technique, the subject moves a pointer, a mouse, or some other response device to provide a continuous record of sensory intensity for a specified attribute. Usually, just one attribute at a time is rated since it is very difficult to attend continuously or even by rapid shifting of attention to more than one attribute. This would seem to be a perfect opportunity for the dumping tendency to produce illusory enhancements (e.g., Bonnans and Noble, 1993). This idea was tested in experiments with repeated category ratings across time, a time–intensity procedure that allows for ratings of multiple attributes. These studies showed sweetness enhancement in sweet–fruity mixtures when sweetness alone was rated, but little or

no enhancement when both sweetness and fruit intensity were rated over time (Clark and Lawless, 1994). This is exactly parallel to the sweetness enhancement results seen by Frank and colleagues. Workers using single-attribute time–intensity ratings should be wary of apparent enhancements due to response restriction.

9.5.3 Over-Partitioning

In the data of van der Klaauw and Frank (1996), one can also see cases in which having too many attributes causes a deflation in ratings. As in the dumping examples, their usual paradigm was to compare the sweetness ratings of a simple sucrose solution to the same concentration with a fruity odor added. When rating sweetness only, the rating is higher than when rating sweetness and fruitiness, the common dumping effect. But when total intensity and six additional attributes were rated, the sweetness rating was significantly lower than either of the other two conditions. In another example, including a bitterness rating (in addition to the sweetness and fruitiness ratings) lowered the sweetness rating compared to rating sweetness (the highest condition) and also compared to rating sweetness and fruitiness (an intermediate sweetness rating was obtained). This effect appears to be a deflation due to people over-partitioning their sensations into too many categories. The specific choices may be important in this effect. Adding only a bitter or a bitter and a floral rating had little or no effect and dumping inflation was still observed probably because there was no fruity rating.

In the study by Clark and Lawless (1994), the control condition (sweetener only) showed some evidence of a decrement when the attributes for volatiles were also available. Even more dramatic was the complete reversal of sweet enhancement to inhibition when a large number of response categories were provided for simple mixtures (Frank et al., 1993).

Although this effect has not been thoroughly studied, it serves to warn the sensory scientist that the number of choices given to untrained consumers may affect the outcome and that too many choices may be as dangerous as too few. Whether this effect might be seen with trained panels remains an open question. It is sometimes difficult to predetermine the correct number of attributes to rate in order to guard against the dumping effect. Careful pre-testing and discussion phases in descriptive training may help. It is obviously important to be inclusive and exhaustive, but also not to waste the panelists' time with irrelevant attributes.

9.6 Classical Psychological Errors and Other Biases

A number of psychological errors in judgment have been described in the literature and are commonly listed in references in sensory evaluation (e.g., Amerine et al., 1965; Meilgaard et al., 2006). They are only briefly listed here, as they serve primarily as empirical descriptions of behavior, without much reference to cause or any theoretical bases. It is important to distinguish between description and explanation, and not to confuse naming something with trying to explain why it occurred in terms of mechanism or larger theory. The sensory evaluation practitioner needs to be aware of these errors and the conditions under which they may occur.

9.6.1 Errors in Structured Sequences: Anticipation and Habituation

Two errors may be seen when a non-random sequence of products is presented for evaluation, and the observer is aware that a sequence or a particular order of items is going to be presented. The error of anticipation is said to occur when the subject shifts responses in the sequence before the sensory information would indicate that it is appropriate to do so (Mattes and Lawless, 1985). An example is found in the method of limits for thresholds, where an ascending sequence is presented and the observer expects a sensation to occur at some point and "jumps the gun." The opposite effect is said to be the error of habituation, in which the panelist stays put too long with one previous response, when the sensory information would indicate that a change is overdue. Obviously, the presentation of samples in random order will help to undo the expectations involved in causing the error of anticipation. Perseveration is a little bit harder to understand but may have to do with lack of attention or motivation

on the part of the observer or having an unusually strict criterion for changing responses. Attention and motivation can be addressed by sufficient incentives and keeping the test session from being too long.

9.6.2 The Stimulus Error

The stimulus error is another classical problem in sensory measurement. This occurs when the observer knows or presumes to know the identity of the stimulus and thus draws some inference about what it should taste, smell, or look like. The judgment is biased due to expectations about stimulus identity. In the old parlor game of trying to identify the origin and vintage of a wine, stimulus error is actually a big help. It is much easier to guess the wine if you know what the host is prone to drink or you have taken a peek at the bottles in the kitchen beforehand. In sensory evaluation, the principle of blind testing and the practice using random three-digit coding of samples mitigate against the stimulus error. However, panelists are not always completely in the dark about the origin or the identity of samples. Employee panels may have a fair amount of knowledge about what is being tested and they may make inferences, correctly or incorrectly. For example, in small-plant quality assurance, workers may be aware of what types of products are being manufactured that day and these same workers may serve as sensory panelists. In the worst possible scenario, the persons drawing the samples from production are actually doing the tasting. As much as possible, these situations should be avoided. In quality control panels, insertion of blind control samples (both positive controls and flawed samples) will tend to minimize the guesswork by panelists.

9.6.3 Positional or Order Bias

Time-order error is a general term applied to sequential effects in which one order of evaluating two or more products produces different judgments than does another order (Amerine et al., 1965). There are two philosophies for dealing with this problem. The first approach is to provide products in all possible orders, counterbalanced orders, or randomized orders so that the sequential effects may be counterbalanced or averaged out in the group data. The second approach is to consider the order effects of interest. In this case, different orders are analyzed as a purpose of the experiment and, if order effects are observed, they are duly noted and discussed. Whether order effects are of interest will depend upon the circumstances of the product evaluation and its goals. If counterbalanced orders or randomization cannot be worked into the experimental design, the experimenter must consider whether product differences are true sensory differences or are artifacts of stimulus order. Purposeful experimentation and analysis of order effects in at least some of the routine evaluations may give one an appreciation for where and when these effects occur in the product category of interest.

Another well-known order effect in acceptance testing is the reception of a higher score for the first sample in a series (Kofes et al., 2009). Counterbalancing orders is of course appropriate, but one can also give a "dummy" product first to absorb the first product's score. With monadic (single product) tests, such inflation could be misleading (Kofes et al., 2009). Positional bias was also of concern in early paired tests and also the triangle procedure (Amerine et al., 1965). Another bias was seen when preference questions were asked following the triangle difference test (Schutz, 1954). Following a difference test with a preference test is not recommended in good sensory practice, in part because these early studies showed a bias against the sample that was considered the odd one in the triangle test. Recent research indicates that this effect may not be so robust—a replication of Schutz's original experiment but using the words "different" rather than "odd" did not find much evidence for this bias (El Gharby, 1995). Perhaps the meaning of the term "odd" in the 1950s was in itself sufficiently biasing.

9.7 Antidotes

9.7.1 Avoid or Minimize

At first glance, one way to avoid contextual effects would seem to be only to present products as single items, in other words to perform only monadic tests. This may be appropriate in some consumer testing situations where the test itself changes the situation so

dramatically that future evaluation would be unduly influenced by the first product tested. Examples occur in testing consumer products such as insecticides or hair conditioners. However, monadic testing is rarely practical for analytical sensory tests such as descriptive analyses. It would be highly inefficient both financially and statistically to have trained panels evaluate only single samples in a session. More importantly, monadic testing does not take any advantage of the inherent comparative abilities of human observers. Furthermore, because of transfer bias (panelists, after all, have memories), this solution may be illusory for an ongoing testing program. Even without an immediate frame of reference in the experimental session, people will evaluate products based on their memory of similar items that have been recently experienced. So they will adopt a frame of reference if none is explicitly provided.

Poulton (1989) asserted that in most Western cultures, this baseline would be fairly constant (perhaps for consumers) and neutral based on the comparable experiences of experimental participants. He went on to suggest monadic testing as a way to avoid frequency and centering biases. However, experimental evidence for this assertion is lacking. Also, given the high degree of idiosyncratic food preferences and food habits, a constant baseline across individuals seems rather unlikely in sensory evaluation of foods. So monadic testing could actually add noise to the data. Furthermore, monadic test designs necessitate the use of between-subject comparisons and lose the statistical and scaling advantages inherent in using subjects as their own controls or baselines for comparison.

The problem can be rephrased as to how sensory evaluation specialists can control for contextual biases or minimize them. There are four approaches to dealing with context effects: randomization (including counterbalancing), stabilization, calibration, and interpretation. Stabilization refers to the attempt to keep context the same across all evaluation sessions so that the frame of reference is constant for all observers. Calibration refers to the training of a descriptive panel so that their frame of reference for the scale is internalized through training with reference standards. Interpretation is simply the careful consideration of whether ratings in a given setting may have been influenced by experimental context, e.g., by the specific items that were also presented in that session. Each of these approaches is considered below.

9.7.2 Randomization and Counterbalancing

The use of different random or counterbalanced orders has long been a principle of good practice in applied sensory testing. Simple order effects, sequential dependencies, and contrast between any two items can be counteracted by using sufficient orders so that the immediate frame of reference for any one product is different across a group of respondents. Using an order that mixes up products from different positions on the scale range will also help avoid the occurrence of a local frequency bias. That is, if too many samples are given from the high end of the sensory continuum to be rated, it may give the respondent the impression that there is bunching at that end, even though there may not be in the product set as a whole. Poulton (1989) noted that a "controlled balanced order" might help to avoid this. As in the discussion of the classical time-order effect above, there are two philosophies here. Whether one randomizes and ignores the local sequential effects, or systematically counterbalances orders and analyzes the order dependencies as effects of experimental interest will depend upon the experimental objectives, the resources of the experimenter and the information needed by the end users of the experimental results.

However, using random or counterbalanced orders will not in itself undo the broader effects of context that develop during an experiment with repeated ratings. The broader frame of reference still exists and is still used by judges to frame the range of products and to map the products onto the known scale range. Thus the use of randomization or counterbalancing of orders does not get around the problem of altered context when results from two different experimental sessions are compared.

Note that the examination of multiple stimulus ranges is similar philosophically to the approach of randomization or counterbalancing. One approach to avoiding the centering bias in just-right scaling is to use different ranges so that the stimulus set that would be centered on the true just-right point can be found by interpolation (Johnson and Vickers, 1987; McBride, 1982) (see Fig. 9.8). Multiple contexts become part of the experimental design. The goal is to purposefully examine the range effects, rather than averaging or neutralizing them through randomization. As a general

principle, the greater the number of contexts within which a product is evaluated, the greater the understanding and accuracy of the final interpretation by the sensory scientist.

9.7.3 Stabilization and Calibration

The second approach to dealing with context effects is to try and hold the experimental context constant across all sessions containing products that will be compared. This can be difficult if the products are fatiguing or if adaptation or carryover effects are likely to restrict the number of products that can be given. In its simplest form, this strategy takes the form of a simple comparison of all test products to one control item (e.g., Stoer, 1992). Difference scores may then be constructed and serve as the primary form of the data. Alternatively, difference-from-reference ratings (Larson-Powers and Pangborn, 1978) or magnitude estimation with a constant reference item may be used. The presentation of a warm-up or practice sample probably has some stabilizing effect, another good reason to use such "throw-away" samples if the product is not too fatiguing. Another approach is to attempt to stabilize the endpoints by using reference standards for high and low stimuli that appear in every session. Also, high and low examples can be given as blind "catch trials" and judges suitably motivated to expand their scale usage if they are caught exhibiting contraction bias, end-of-scale avoidance, or simply gravitating toward the middle of the scale as sometimes occurs in repeated testing. In magnitude estimation and also in the use of the labeled magnitude scale, the reference standard used for comparison may have a stabilizing effect and reduce some of the contrast effects seen with all scaling methods (Diamond and Lawless, 2001; Lawless et al., 2000).

Calibration can refer to the use of bracketing reference standards in the experimental session, or reference standards given in training. Considering the context within the evaluation session is an important part of collecting good sensory judgments. Anderson (1974), for example, in discussing the use of category scales gives the following advice: "Several precautions have been standard with ratings scales in functional measurement. First, is the use of preliminary practice, which has several functions. The general range of stimuli is not known to the subject initially, and the rating scale is arbitrary. Accordingly, the subject needs to *develop a frame of reference* for the stimuli and correlate it with the given response scale." (emphasis added; pp. 231–232). Anderson notes that the effect of such practice is a decrease in variability. His practice in the psychological laboratory for stabilizing scale usage is similar to the training of descriptive panelists with examples of products to be evaluated. Anderson goes on to describe end-anchor stimuli, which serve as low- and high-intensity standards on his 20-point scale: "Stimulus end-anchors are extremely important. These are additional stimuli that are more extreme than the experimental stimuli to be studied. One function of the end-anchors is to help define the frame of reference." (p. 232). In this view then, the proper use of rating scales includes a definition of the context in which the sample is to be judged. In practice, this is achieved by presentation of specific examples that bracket the sensory range. An explicit (if extreme) example of this is the relative scaling method of Gay and Mead (Gay and Mead 1992; Mead and Gay, 1995) in which the samples are inspected, then the highest and lowest placed at the endpoints, and all others distributed along the scale. This does insure usage of the whole scale but renders the data totally relative and totally specific to that context and set of samples.

Calibration of observers is a common practice in descriptive analysis, especially in techniques with intensity reference standards such as the texture profile and the Spectrum method (Meilgaard et al., 2006). Anderson (1974) warned that endpoint calibration was in fact a necessary practice with category scales in order to fix the linear usage of the scale by experimental subjects. What appears to be a problem in obtaining unbiased scale usage may be turned to an advantage if a stable frame of reference can be induced in subjects through training. Evidence shows that contextual effects do not appear all at once but are contingent upon and strengthened by experience (e.g., Lawless, 1983). So the question arises whether it is possible to "inoculate" a trained panelist against the day-to-day context of the experimental session by sufficient training. That is, is it possible to calibrate and stabilize judges' frame of reference to make them immune to contextual effects? Some of the examples of transfer bias cited by Poulton (1989) certainly make this seem reasonable. However, a recent study using extensive training on a 15-point sweetness scale failed to

eliminate simple contrast effects (Olabi and Lawless, 2008). Reference standards were similarly ineffective. Perhaps it is asking too much of human nature to get panelists to act like absolute measuring instruments.

9.7.4 Interpretation

The last approach to context effects and biases is to be aware that they are operating and to draw conclusions about product differences with suitable caution. It is not appropriate to conclude that two products evaluated in different sessions are in fact different unless they were examined in similar contexts. The sensory professional must look at the whole experimental context and not just the summary statistics for the products in question in order to draw accurate conclusions. In drawing conclusions about product differences, it is necessary to question whether the observed difference could possibly have arisen from contextual effects or whether it is likely to be a true sensory-based difference. In situations of apparent enhancement or synergistic influence of one flavor upon another, one must always question how the data were gathered. Were there sufficient and appropriate response categories or was there the possibility that the apparent enhancement occurred due to dumping into the available but restricted set of response attributes?

9.8 Conclusions

> Notes make each other sweeter in a chord, and so do colors when harmoniously applied. A certain amount of skin dipped in hot water gives the perception of a certain heat. More skin immersed makes the heat much more intense although of course the water's heat is the same. (James, 1913, p. 25).

William James reminds us that complex patterns of stimulation alter the perception of things around us. What something looks like, feels like, or tastes like depends upon the other patterns of stimulation that are present and that have come before the stimulus to be judged. Context effects are present in all scaling methods, and human behavior appears to be a sort of compromise between a completely relative system of judgment and one that retains some absolute or calibrated properties (Ward, 1987). A main point of this chapter has been to show the importance of a frame of reference in influencing people's judgments about products. We have seen that other products in the same session can have a marked effect, usually one of contrast. The word anchors or phrases on the scale can influence the frame of reference (what you ask and how you ask it) and, from the dumping effect, we find that even what you *do not ask* can also influence responses. One goal of the sensory evaluation specialist should be to minimize unwanted effects that endanger the accuracy of results. Poulton (1989) discussed range biases as potential problems and stated that they were unavoidable in category scaling. Another perspective is to consider them interesting human phenomena, worthy of study for what they tell us about the judgment process. A third approach is to embrace the relative nature of human judgment and reduce all data to difference scores or similar explicit comparisons. This would be difficult for most practitioners of descriptive analysis methods.

From a practical perspective, a very real danger exists in sensory evaluation when people try to compare ratings given to products in different settings or from different experimental sessions. Unless the context and frame of reference is the same in both sessions, it is not possible to say whether differences between the products arose from true sensation differences or from differences in ratings due to contextual effects. A difference in the data set might occur merely because the two items were viewed among higher or lower valued items in their test sessions. Conversely, two items might appear similar in ratings across two sessions, but their ratings might be similar only due to range or centering effects. How can a sensory practitioner know whether the scale value of 5.3 for this week's product is actually superior to the value of 4.9 given to the prototype evaluated last week?

Unless the sensory professional is aware of context effects and guards against them, inappropriate conclusions may be drawn from evaluation sessions, especially if products evaluated in different contexts are to be compared. Sometimes the effects can be subtle and insidious. Consider the context effect discussed above in optimization (Mattes and Lawless, 1985). Ascending in concentration results in the estimation of a preferred sensory optimum that is too low relative to the peak that would be obtained from a randomized

order of samples. Yet adding a flavor ingredient "to taste" is what product developers commonly do at the benchtop or chefs do in a research kitchen when they come up with a seemingly optimized ingredient level. However, even an informal tasting has its own context. We cannot assume that the results of such informal tastings are accurate—only properly randomized or counterbalanced sensory testing using acceptability ratings or a just-right scale with precautions against centering bias would give accurate direction as to the appropriate flavor level.

It is sometimes difficult to pin down exactly which of the many biases discussed in this chapter may be operating in an experiment. For example, the contrast effect and response contraction effects can work against each other so that there may be little evidence that these tendencies are present in a particular evaluation session (Schifferstein and Frijters, 1992). Communication with test panelists, careful examination of the data, and looking at historical records can all give hints about what may be going on with a panel that does repeated testing. Is there evidence of response contraction or gravitation toward the center of the scale? If these conservative tendencies are creeping in, they should show up through examination of data sets over time. The contraction effect is insidious since the trend may appear as reduced standard deviations that may give the false impression that the panel is achieving higher levels of calibration or consensus. However, the reduced error will not be accompanied by a higher rate of significant differences among products, since the product means will also gravitate toward the middle of the scale and differences between means will be smaller. This type of insight highlights the depth of analysis that is needed in good sensory practice and that a sensory professional must be more than a person who merely conducts tests and reports results. They must be in touch with the trends and finer grain of the data set along with the psychological tendencies of the respondents that may arise as a function of changes in frame of reference and other biases.

References

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

Anderson, N. 1974. Algebraic models in perception. In: E. C. Carterette and M. P. Friedman (eds.), Handbook of Perception. II. Psychophysical Judgment and Measurement. Academic, New York, pp. 215–298.

Baird, J. C. and Noma, E. 1978. Fundamentals of Scaling and Psychophysics. Wiley, New York.

Beebe-Center, J. G. 1932. The Psychology of Pleasantness and Unpleasantness. Russell & Russell, New York.

Bingham, A. F., Birch, G. G., de Graaf, C., Behan, J. M. and Perring, K. D. 1990. Sensory studies with sucrose maltol mixtures. Chemical Senses, 15, 447–456.

Birnbaum, M. H. 1982. Problems with so called "direct" scaling. In: J. T. Kuznicki, A. F. Rutkiewic and R. A. Johnson (eds.), Problems and Approaches to Measuring Hedonics (ASTM STP 773). American Society for Testing and Materials, Philadelphia, pp. 34–48.

Bonnans, S. and Noble, A. C. 1993. Effects of sweetener type and of sweetener and acid levels on temporal perception of sweetness, sourness and fruitiness. Chemical Senses, 18, 273–283.

Boring, E. G. 1942. Sensation and Perception in the History of Experimental Psychology. Appleton-Century-Crofts, New York.

Cardello, A. V. and Sawyer, F. M. 1992. Effects of disconfirmed consumer expectations on food acceptability. Journal of Sensory Studies, 7, 253–277.

Cardello, A. V., Lawless, H. T. and Schutz, H. G. 2008. Effects of extreme anchors and interior label spacing on labeled magnitude scales. Food Quality and Preference, 21, 323–334.

Clark, C. C. and Lawless, H. T. 1994. Limiting response alternatives in time–intensity scaling: An examination of the halo-dumping effect. Chemical Senses, 19, 583–594.

Diamond, J. and Lawless, H. T. 2001. Context effects and reference standards with magnitude estimation and the labeled magnitude scale. Journal of Sensory Studies, 16, 1–10.

Diehl, R. L., Elman, J. L. and McCusker, S. B. 1978. Contrast effects on stop consonant identification. Journal of Experimental Psychology: Human Perception and Performance, 4, 599–609.

Dijksterhuis, G. 1993. Principal component analysis of time–intensity bitterness curves. Journal of Sensory Studies, 8, 317–328.

Dolese, M., Zellner, D., Vasserman, M. and Parker, S. 2005. Categorization affects hedonic contrast in the visual arts. Bulletin of Psychology and the Arts, 5, 21–25.

Eimas, P. D. and Corbit, J. D. 1973. Selective adaptation of linguistic feature detectors. Cognitive Psychology, 4, 99–109.

El Gharby, A. 1995. Effect of Nonsensory Information on Sensory Judgments of No-Fat and Low-Fat Foods: Influences of Attitude, Belief, Eating Restraint and Label Information. M.Sc. Thesis, Cornell University.

Eng, E. W. 1948. An Experimental Study of the Reliabilities of Rating Scale for Food Preference Discrimination. M. S. Thesis, Northwestern University, and US Army Quartermaster Food and Container Institute, Report # 11–50.

Engen, T. and Levy, N. 1958. The influence of context on constant-sum loudness judgments. American Journal of Psychology, 71, 731–736.

Frank, R. A. and Byram, J. 1988. Taste–smell interactions are tastant and odorant dependent. Chemical Senses, 13, 445.

Frank, R. A., van der Klaauw, N. J. and Schifferstein, H. N. J. 1993. Both perceptual and conceptual factors influence taste–odor and taste–taste interactions. Perception & Psychophysics, 54, 343–354.

Gay, C. and Mead, R. 1992 A statistical appraisal of the problem of sensory measurement. Journal of Sensory Studies, 7, 205–228.

Giovanni, M. E. and Pangborn, R. M. 1983. Measurement of taste intensity and degree of liking of beverages by graphic scaling and magnitude estimation. Journal of Food Science, 48, 1175–1182.

Green, B. G., Dalton, P., Cowart, B., Shaffer, G., Rankin, K. and Higgins, J. 1996. Evaluating the 'labeled magnitude scale' for measuring sensations of taste and smell. Chemical Senses, 21, 323–334.

Hanson, H. L., Davis, J. G., Campbell, A. A., Anderson, J. H. and Lineweaver, H. 1955. Sensory test methods II. Effect of previous tests on consumer response to foods. Food Technology, 9, 56–59.

Helson, H. H. 1964. Adaptation-Level Theory. Harper & Rowe, New York.

James, W. 1913. Psychology. Henry Holt and Company, New York.

Johnson, J. and Vickers, Z. 1987. Avoiding the centering bias or range effect when determining an optimum level of sweetness in lemonade. Journal of Sensory Studies, 2, 283–291.

Jones, F. N. and Marcus, M. J. 1961. The subject effect in judgments of subjective magnitude. Journal of Experimental Psychology, 61, 40–44.

Jones, F. N. and Woskow, M. J. 1966. Some effects of context on the slope in magnitude estimation. Journal of Experimental Psychology, 71, 177–180.

Kamenetzky, J. 1959. Contrast and convergence effects in ratings of foods. Journal of Applied Psychology, 43(1), 47–52.

Kofes, J., Naqvi, S., Cece, A. and Yeh, M. 2009. Understanding Presentation Order Effects and Ways to Control Them in Consumer Testing. Paper presented at the 8th Pangborn Sensory Science Symposium, Florence, Italy.

Larson-Powers, N. and Pangborn, R. M. 1978. Descriptive analysis of the sensory properties of beverages and gelatins containing sucrose or synthetic sweeteners. Journal of Food Science, 43, 47–51.

Lawless, H. T. 1983. Contextual effect in category ratings. Journal of Testing and Evaluation, 11, 346–349.

Lawless, H. T. 1994. Contextual and Measurement Aspects of Acceptability. Final Report #TCN 94178, US Army Research Office.

Lawless, H. T. and Malone, G. J. 1986a. A comparison of scaling methods: Sensitivity, replicates and relative measurement. Journal of Sensory Studies, 1, 155–174.

Lawless, H. T. and Malone, J. G. 1986b. The discriminative efficiency of common scaling methods. Journal of Sensory Studies, 1, 85–96.

Lawless, H. T., Glatter, S. and Hohn, C. 1991. Context dependent changes in the perception of odor quality. Chemical Senses, 16, 349–360.

Lawless, H. T., Horne. J. and Speirs, W. 2000. Contrast and range effects for category, magnitude and labeled magnitude scales. Chemical Senses, 25, 85–92.

Lawless, H. T., Popper, R. and Kroll, B. J. 2010a. Comparison of the labeled affective magnitude (LAM) scale, an 11-point category scale and the traditional nine-point Hedonic scale. Food Quality and Preference, 21, 4–12.

Lawless, H. T., Sinopoli, D. and Chapman, K. W. 2010b. A comparison of the labeled affective magnitude scale and the nine point hedonic scale and examination of categorical behavior. Journal of Sensory Studies, 25, S1, 54–66.

Lee, H.-S., Kim, K.-O. and O'Mahony, M. 2001. How do the signal detection indices react to frequency context bias for intensity scaling? Journal of Sensory Studies, 16, 33–52.

Marks, L. E. 1994. Recalibrating the auditory system: The perception of loudness. Journal of Experimental Psychology: Human Perception & Performance, 20, 382–396.

Mattes, R. D. and Lawless, H. T. 1985. An adjustment error in optimization of taste intensity. Appetite, 6, 103–114.

McBride, R. L. 1982. Range bias in sensory evaluation. Journal of Food Technology, 17, 405–410.

McBride, R. L. and Anderson, N. H. 1990. Integration psychophysics. In R. L. McBride and H. J. H. MacFie (eds.), Psychological Basis of Sensory Evaluation. Elsevier Applied Science, London, pp. 93–115.

McBurney, D. H. 1966. Magnitude estimation of the taste of sodium chloride after adaptation to sodium chloride. Journal of Experimental Psychology, 72, 869–873.

McGowan, B. A. and Lee, S.-Y. 2006. Comparison of methods to analyze time–intensity curves in a corn zein chewing gum study. Food Quality and Preference, 17, 296–306.

Mead, R. and Gay, C. 1995. Sequential design of sensory trials. Food Quality and Preference, 6, 271–280.

Meilgaard, M., Civille, G. V. and Carr, B. T. 2006. Sensory Evaluation Techniques, Third Edition. CRC, Boca Raton, FL.

Mellers, B. A. and Birnbaum, M. H. 1982. Loci of contextual effects in judgment. Journal of Experimental Psychology: Human Perception and Performance, 4, 582–601.

Mellers, B. A. and Birnbaum, M. H. 1983. Contextual effects in social judgment. Journal of Experimental Social Psychology, 19, 157–171.

Muñoz, A. M. and Civille, G. V. 1998. Universal, product and attribute specific scaling and the development of common lexicons in descriptive analysis. Journal of Sensory Studies, 13, 57–75.

Murphy, C. 1982. Effects of exposure and context on hedonics of olfactory-taste mixtures. In: J. T. Kuznicki, R. A. Johnson and A. F. Rutkeiwic (eds.), Selected Sensory Methods: Problems and Applications to Measuring Hedonics. American Society for Testing and Materials, Philadelphia, pp. 60–70.

Murphy, C. and Cain, W. S. 1980. Taste and olfaction: Independence vs. interaction. Physiology and Behavior, 24, 601–605.

Olabi, A. and Lawless, H. T. 2008. Persistence of context effects with training and reference standards. Journal of Food Science, 73, S185–S189.

Parducci, A. 1965. Category judgment: A range-frequency model. Psychological Review, 72, 407–418.

Parducci, A. 1974. Contextual effects: A range-frequency analysis. In: E. C. Carterette and M. P. Friedman (eds.), Handbook of Perception. II. Psychophysical Judgment and Measurement. Academic, New York, pp. 127–141.

Parducci, A. and Perrett, L. F. 1971. Category rating scales: Effects of relative spacing and frequency of stimulus values. Journal of Experimental Psychology (Monograph), 89(2), 427–452.

Parducci, A., Knobel, S. and Thomas, C. 1976. Independent context for category ratings: A range-frequency analysis. Perception & Psychophysics, 20, 360–366.

Parker, S., Murphy, D. R. and Schneider, B. A. 2002. Top-down gain control in the auditory system: Evidence from identification and discrimination experiments. Perception & Psychophysics, 64, 598–615.

Poulton, E. C. 1989. Bias in Quantifying Judgments. Lawrence Erlbaum Associates, Hillsdale, NJ.

Rankin, K. M. and Marks, L. E. 1991. Differential context effects in taste perception. Chemical Senses, 16, 617–629.

Riskey, D. R. 1982. Effects of context and interstimulus procedures in judgments of saltiness and pleasantness. In: J. T. Kuznicki, R. A. Johnson and A. F. Rutkeiwic (eds.), Selected Sensory Methods: Problems and Applications to Measuring Hedonics. American Society for Testing and Materials, Philadelphia, pp. 71–83.

Riskey, D. R., Parducci, A. and Beauchamp, G. K. 1979. Effects of context in judgments of sweetness and pleasantness. Perception & Psychophysics, 26, 171–176.

Sandusky, A. and Parducci, A. 1965. Pleasantness of odors as a function of the immediate stimulus context. Psychonomic Science, 3, 321–322.

Sarris, V. 1967.Adaptation-level theory: Two critical experiments on Helson's weighted-average model. American Journal of Psychology, 80, 331–344.

Sarris, V. and Parducci, A. 1978. Multiple anchoring of category rating scales. Perception & Psychophysics, 24, 35–39.

Schifferstein, H. J. N. 1995. Contextual shifts in hedonic judgment. Journal of Sensory Studies, 10, 381–392.

Schifferstein, H. J. N. 1996. Cognitive factors affecting taste intensity judgments. Food Quality and Preference, 7, 167–175.

Schifferstein, H. N. J. and Frijters, J. E. R. 1992. Contextual and sequential effects on judgments of sweetness intensity. Perception & Psychophysics, 52, 243–255.

Schutz, H. G. 1954. Effect of bias on preference in the difference-preference test. In: D. R. Peryam, J. J. Pilgram and M. S. Peterson (eds.), Food Acceptance Testing Methodology. National Academy of Sciences, Washington, DC, pp. 85–91.

Stevenson, R. J., Prescott, J. and Boakes, R. A. 1995. The acquisition of taste properties by odors. Learning and Motivation, 26, 433–455.

Stoer, N. L. 1992. Comparison of Absolute Scaling and Relative-To-Reference Scaling in Sensory Evaluation of Dairy Products. Master's Thesis, Cornell University.

Teghtsoonian, R. and Teghtsoonian, M. 1978. Range and regression effects in magnitude scaling. Perception & Psychophysics, 24, 305–314.

Thorndike, E. L. 1920. A constant error in psychophysical ratings. Journal of Applied Psychology, 4, 25–29.

van der Klaauw, N. J. and Frank, R. A. 1996. Scaling component intensities of complex stimuli: The influence of response alternatives. Environment International, 22, 21–31.

Vollmecke, T. A. 1987. The Influence of Context on Sweetness and Pleasantness Evaluations of Beverages. Doctoral dissertation, University of Pennsylvania.

Ward, L. M. 1979. Stimulus information and sequential dependencies in magnitude estimation and cross-modality matching. Journal of Experimental Psychology: Human Perception and Performance, 5, 444–459.

Ward, L. M. 1987. Remembrance of sounds past: Memory and psychophysical scaling. Journal of Experimental Psychology: Human Perception and Performance, 13, 216–227.

Zellner, D. A., Allen, D., Henley, M. and Parker, S. 2006. Hedonic contrast and condensation: Good stimuli make mediocre stimuli less good and less different. Psychonomic Bulletin and Review, 13, 235–239.