(Harwood *et al.*, 2012) as well as for high cocoa flavanol containing cocoa powder in semisweet chocolate (Harwood *et al.*, 2013).

Threshold tests, specifically detection thresholds, have also been historically recommended in sensory evaluation texts as a means to determine the sensory acuity of individuals who are being considered for trained panels or other tasting roles. While these approaches have some limited utility in identifying individuals with gross defects in chemosensory function, modern understanding of human sensory systems suggests threshold-based approaches to screening panelists may have been an intellectual dead-end. Specifically, the relationship between detection threshold and suprathreshold intensity is not a simple one (see Keast and Roper, 2007; Bartoshuk and Klee, 2013), so knowing an individual's detection threshold reveals little about their responses at higher levels, Additionally, given the wide diversity of human bitter receptors and ligands that bind to them (Meyerhof *et al.*, 2010), as well as the genetic variation across receptors and individuals (Hayes *et al.*, 2011; Hayes *et al.*, 2013), screening individuals with a single bitterant like caffeine does not reveal any meaningful information about responses to other bitter compounds. For example, we recently demonstrated that the bitterness of the sweetener Acesulfame potassium (AceK) is totally unrelated to that of the stevia derived sweetener RebaudiosideA (RebA; Allen *et al.*, 2013). Accordingly, best practice when screening participants is to use the stimulus of interest at suprathreshold concentrations, rather than exemplars like caffeine at perithreshold levels.

21.2.2 Discrimination tests

Discrimination tests can be defined as "any method to determine if differences among stimuli are perceptible" (ASTM, 2013). These include but are not limited to tests such as triangle tests, tetrad tests, duo–trio tests, paired comparisons, difference from control tests and so on; these methods have been successfully carried out in chocolate products. For example, Aguilar and colleagues (1995) utilised triangle tests to determine if perceptible differences existed between chocolate samples processed via two different conching methods.

Discrimination tests can be set up to determine either difference (i.e. are the products different enough for the consumer to tell them apart?) or similarity (i.e. is a change in recipe small enough that it is not noticeable to the consumer?) between samples, depending on the statistical basis of analysis. Discrimination testing can also be directional. That is, the assessor can be asked to differentiate between samples based on a specific attribute. For example, directional paired comparisons (i.e. which sample is more sweet?) or *n*-alternative forced choice (*n*-AFC) tests can be carried out (i.e. of these samples, please identify which is the most sweet). This can be a particularly useful tool for claim substantiation [see ASTM E1958-06 (2014) for more information].

Discrimination testing is most useful when the differences between samples are subtle. That said, this also however increases the risk of type II errors in

judgment, where a real difference may be missed (a false negative; Lawless and Heymann, 2010). This is particularly concerning when dealing with a complex product such as chocolate, where the increasing complexity of the product is generally accompanied by an increase in noise in the samples, making true differences more difficult to detect. In this case, the sensory scientist must consider the relative signal to noise ratio for the specific product when determining the appropriate sample size and size of α (control of type I error) to ensure the power of the test to correctly identify differences.

21.2.3 Affective testing

Hedonic testing can be a useful tool for understanding how consumers perceive and react to specific products affectively. Liking, preference and/or acceptability can all be measured, both for products as a whole and for specific attributes. There is an abundance of tools that can be used to accomplish these types of goals. For acceptability and appropriateness, common tools could include scales, such as the nine-point hedonic scale (Peryam and Pilgrim, 1957) or Just About Right (JAR) scales. Additionally, liking can also be measured using the Army Quartermaster nine-point hedonic scale. Preference implies the comparison of at least two samples, and it can be measured for example via paired preference testing (i.e. 2-AFC methods) or a ranking task with three or more samples. Hedonic testing is generally carried out with untrained assessors, as the goal is usually to capture and quantify the consumer experience. Hedonic testing has been carried out successfully in chocolate (e.g. Bordi *et al.*, 2002; Lee *et al.*, 2002). However caution should be used when interpreting these types of results because ratings of chocolate are often skewed towards the higher/more well liked/more acceptable end of a given scale. When this is the case, seemingly small differences between samples may carry more weight and the interpretation of scores may come down to understanding the difference between statistical significance and consumer relevance for that particular attribute/those particular samples.

21.2.4 Descriptive analysis

There are many different methods that have been developed and described in detail in the literature for descriptive analysis and which have been applied to chocolate. These types of tests allow for the construction of a sensory profile of one or numerous products using words that are often linked to precise definitions and representative standards or references. Additionally, these methods provide quantitative information about the relative intensities of the attributes within the product. Descriptive analysis is generally performed by extensively trained panels of 8–12 assessors. Descriptive analysis results can be paired with discrimination testing or consumer testing to give further insight in interpreting results found in other types of testing. For example, if consumers prefer one chocolate over another, descriptive panel data quantifying how "waxy" each