Food Quality and Preference 43 (2015) 122–125

Contents lists available at ScienceDirect

Food Quality and Preference

journal homepage: www.elsevier.com/locate/foodqual

Sensory discrimination testing and consumer relevance

Benoît Rousseau *

The Institute for Perception, 2306 Anza Avenue, Davis, CA 95616, United States

ARTICLE INFO

Article history:
Received 25 January 2015
Received in revised form 28 February 2015
Accepted 1 March 2015
Available online 10 March 2015

Keywords:
Discrimination testing
Consumer relevance
Sensory difference
Power
Sample size

ABSTRACT

In order to ensure consistency in the decision making process over time, a discrimination testing program must take into account all of five elements: The testing protocol, the sample size, the Type I error (α), the Type II error (β) (power = 1 − β) and a measure of the threshold above which the scientist has established that the difference is meaningful to the consumer (δ_R). Two putatively different products will always be found to be different provided that the sample size is large enough. This fact underscores the need to set δ_R. The concept of discriminators is attractive but flawed, as the same underlying sensory difference will result in different proportions of distinguishers depending on the method used. Prescott, Leslie, Kunst, and Kim (2005) proposed the idea of consumer rejection threshold which avoids the pitfalls of the proportion of discriminators concept. However, it is limited to differences that can be linked to a specific compound, such as one responsible for a product defect or off-flavor. In this manuscript two alternative approaches are discussed. The first one uses a special feature of the same-different protocol which permits the estimation of the size of the sensory difference above which consumers would call two products "different". The second one links the estimate of a standardized measure of sensory difference, d', to consumer hedonic response between the product pairs and finds the threshold above which a sensory difference results in a meaningful preference result. Experimental research is needed to study the suitability of these approaches. Ultimately, establishing δ_R is essential to ensure that results from a discrimination testing program are actually relevant to the consumers whose behavior it is trying to predict.

1. Introduction

The investigation of whether sensory differences exist between products is often conducted using discrimination methodologies such as the triangle, same-different, 2-alternative forced choice or tetrad tests. Typical research involves an ingredient replacement for cost saving or regulatory requirements ("matching" objective) or product modification where the scientist must confirm that a quality improvement has actually been achieved. Many tools are available to the sensory scientist, including rating scales (e.g., descriptive analysis) and studies involving consumers for hedonic investigations, but discrimination testing has the advantage of not requiring the same level of expertise (descriptive analysis) or large numbers of subjects (hedonic-based investigations).

A typical study would involve a number of panelists, trained or not, a discrimination protocol and a decision making process usually based on the statistical p-value of 5% (Type I error). For instance, a company might typically conduct triangle tests with 20 semi-trained employees. Using the binomial test, a minimum of 11 tests correct is required to reach a conclusion of significant difference at the 5% level. If fewer than 11 correct responses are obtained, the result is inconclusive, even though it is often wrongly assumed that no difference exists or that it is "small enough". In such a program the results are used to predict consumer behavior. Specifically, in the case of research conducted to match a reference product, a statistically significant outcome will result in the rejection of the alternative product as the difference is "too large" while a non-significant finding will usually provide assurances that the difference is "small enough". As will be shown in the remainder of this article, a significant difference will be meaningless unless the scientist has initially defined the size above which a sensory difference is meaningful to the consumer. Such difference will be thereafter labeled as δ_R.

A suitably set up sensory discrimination testing program will involve five components, as represented as the five fingers of a hand that are essential to its success (Fig. 1): the sensory protocol, the sample size, the size of the underlying difference of interest δ_R, the Type I error level (α, the likelihood of wrongly concluding a difference, usually set at 5%), and the Type II error (β, the likelihood of missing a difference of size δ_R, usually set at 10% or 20%; test power = 1 − β). All five components are intimately linked so that

* Tel.: +1 (530) 848 6008.
E-mail address: Benoit.Rousseau@IFPress.com

http://dx.doi.org/10.1016/j.foodqual.2015.03.001
0950-3293/

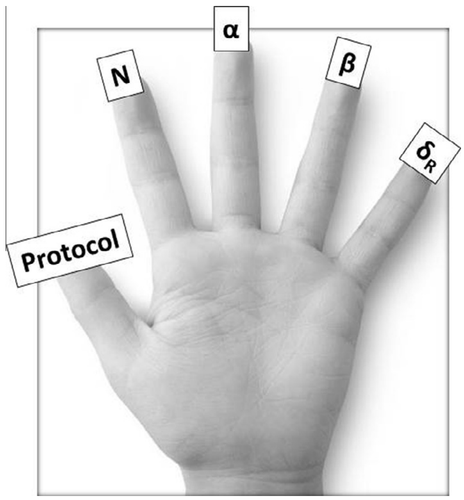

Fig. 1. Representation of the 5 elements in a sensory discrimination program. Protocol, N: sample size, α: Type I error; β: Type II error; δ_R: size of the difference no to be missed.

if the protocol and three others are chosen, the fifth and last one is automatically set. Tables summarizing these relationships have been published. They are generally set up so that the protocol, α, β (or power) and δ_R are chosen and the corresponding needed sample size is estimated (Ennis, 1993; Ennis & Jesionka, 2011; Schlich, 1993). For instance, using Ennis and Jesionka's Table 1, for a triangle test, at $\alpha = 5\%$, $\beta = 20\%$ (Power = 80%) and $\delta_R = 1.5$ (Thurstonian δ), the corresponding necessary experimental sample size is 57.

While one can relate to four of the factors fairly easily, namely α, β, the protocol and the sample size, setting δ_R might not be as straightforward. How does one decide or even measure the size of the sensory difference above which it becomes relevant especially to the consumer? In 1948, Roessler, Warren, and Guymon introduced the concept of Type I error in sensory discrimination as a way to study product differences. However, soon after this first piece of research the same group of authors (Baker, Amerine, & Roessler, 1954), citing the work of Lockhart (1951) called attention to the importance of taking the Type II error into account. Doing so requires establishing the size of the relevant sensory difference the scientist does not want to miss. The importance of this notion is illustrated further by Chew (1977) who points out that provided two putatively different products and a test sample size that is large enough a statistically significant difference will always be found. Therefore establishing consumer relevance for a sensory difference is essential. In this manuscript we review some existing options (proportion discriminators, consumer rejection threshold) as well as novel approaches available to set δ_R.

2. Proportion discriminators

Proportions discriminators can be seen as the first attempt to quantify the size of the underlying sensory difference between products. The idea is that following the results from a forced choice experiment, typically using a triangle, duo-trio or 2-AFC method, the proportion of the subjects who could discriminate between the samples can be estimated.

It is a very attractive concept as the effect size can take any value between 0% and 100% discriminators. Once a scientist has decided which proportion is acceptable, i.e., "small enough", power and sample size calculations can easily be made that link such values to the final decision of similarity or difference (e.g., Schlich (1993)). A detailed description of the idea behind the notion of discriminators can be found in Lawless and Heymann (2010, chap. 5.4).

Unfortunately, this concept has an inherent weakness that makes it unreliable and generally unusable for sensory discrimination testing power calculation. The issue is that there is no such thing as a "discriminator" and that the proportion will vary depending on the methodology used. For instance, a sensory difference corresponding to 13% discriminators in the triangle test will correspond to 24% discriminators in a tetrad test and 52% in a 2-AFC (Ennis, 1993; Jesionka, Rousseau, & Ennis, 2014). The reason behind this lack of consistency is that the proportion of discriminators is linearly linked to proportion of tests correct. Since proportion of tests correct varies with the method used as described by their psychometric functions (Ennis, 1993; Ennis & Jesionka, 2011), P_d cannot be used as a reliable unit of underlying sensory difference and should be avoided.

3. Consumer rejection threshold

The next concept for establishing consumer relevance was proposed by Prescott et al. (2005), namely the idea of consumer rejection threshold (CRT). This is the underlying idea behind survival analysis where the suitability of a sample is established against a time threshold above which it will be rejected. Prescott et al.'s research involved increasing concentrations of 2,4,6-trichloroanisole (TCA), a compound responsible for "cork taint" in white wine. The authors presented pairs of white wine, one being the untainted control and the other the same wine with a given TCA concentration. Subjects had to indicate which of the two they preferred. A curve could then be generated linking the TCA concentration in the tainted sample to the overall panel preference. As long as the TCA concentration does not reach the CRT value, it can be considered as "small enough", while it becomes relevant once it exceeds the value.

The authors defined the CRT as the level for which the untainted sample starts being significantly preferred (in their case at 66% preference for $p < 0.05$). This approach is not recommended as the threshold will be directly dependent on the sample size. Harwood, Ziegler, and Hayes (2012) used a similar approach to establish the CRT for sucrose octaacetate (SOA) which has bitter properties. In this research, they defined the CRT as 50% above chance, or at 75% preferring control. This definition of a threshold is again subjective. It is unlikely that a manufacturer of consumer goods would consider their product to be equivalent to a competitor if the latter were to be preferred 72% of the time. Other published research on the topic of consumer rejection threshold includes the work of Lee, Prescott, and Kim (2008) and Saliba, Bullock, and Hardie (2009).

The idea of consumer rejection threshold is very useful to establish consumer relevance and its underlying concept will be used in what will be described next. One of its limitations, as illustrated in the research literature referenced above, is that it works well if a physical ingredient can be added until a preference, usually for the sample without the additive, is detected. If a scientist works with products for which the difference cannot be described in terms of a single ingredient, such relationship cannot be built and the CRT cannot be quantified. It is also limited to product properties that can be generally related to "off-flavor" or "defects" and thus will have a monotonic relationship to consumer preference. Other more generalizable ideas are described next.

4. Relating consumer and trained panel sensitivities

Panel sensitivities can be measured using the Thurstonian δ. This index is a standardized unit of sensory difference based on