Food Quality and Preference 32 (2014) 98–106

Contents lists available at SciVerse ScienceDirect

Food Quality and Preference

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

Precision of measurement in Tetrad testing

John M. Ennis [a,*], Rune H.B. Christensen [b]

[a] The Institute for Perception, Richmond, VA, USA
[b] The Technical University of Denmark, Lyngby, Denmark

ARTICLE INFO

Article history:
Received 30 September 2012
Received in revised form 18 February 2013
Accepted 11 May 2013
Available online 29 May 2013

Keywords:
Difference testing
Tetrad
Triangle
Statistics
Equivalence

ABSTRACT

Interest in the Tetrad test has increased recently as it has become apparent that this methodology can be a more powerful alternative to the Triangle test within the standard difference testing paradigm. But when products are tested following an ingredient or process change, a pressing question is whether a sensory difference is large enough to be meaningful. To this end, in this paper we examine the precision of measurement offered by the Tetrad test as compared to two other standard forced-choice discrimination testing procedures – the Triangle and 2-AFC tests. This comparison is made from a Thurstonian perspective. In particular, for all three methods we compare: (1) The variances in the maximum-likelihood estimates of the Thurstonian measure of sensory difference, (2) The expected widths of the corresponding likelihood-based confidence intervals, and (3) The power of the tests when used for equivalence testing. We find that the Tetrad test is consistently more precise than the Triangle test and is sometimes even more precise than the 2-AFC. As a result of this precision, we discover that the Tetrad test is typically more powerful than the Triangle test for equivalence testing purposes and can, under certain conditions, even be more powerful than the 2-AFC.

1. Introduction

Tetrad testing has received increased attention lately as business and political pressures have forced difference testing back to the forefront of sensory science. In particular, ingredient changes can cause such subtle changes to products that, in many cases, sensory scientists would prefer not to specify an attribute of interest. Yet a longstanding problem in sensory science has been the shortage of sensitive testing methods in the absence of a specified attribute (Ennis, 1990, 1993; Frijters, 1979; Gridgeman, 1970). In recent years, this topic has been of active interest (Hautus, Shepherd, & Peng, 2011; Kim & Lee, 2012; Lee, Van Hout, & Hautus, 2007; van Hout, Hautus, & Lee, 2011), with one proposed solution being the increased use of Tetrad testing (Ennis, 2012).

In the Tetrad test, respondents are presented with four samples – two samples from one group and two from another – and are asked to group the samples into two groups of two based on similarity. Importantly, these instructions are not the same as to select the two samples that are most similar – these latter instructions can lead to two intermediate samples being selected as a pair, leaving the two samples that are most different from each other to be called the second pair. Although the idea of the Tetrad test is not new (Frijters, 1984; Gridgeman, 1956; Lockhart, 1951), it was not explored until the mid-1990's (Delwiche & O'Mahony, 1996a; Ennis, Ennis, Yip, & O'Mahony, 1998; Masuoka, Hatjopoulos, & O'Mahony, 1995; O'Mahony, Masouka, & Ishii, 1994). Even following this exploration, the theoretical power of the Tetrad test was not noted in the sensory literature until 2011 (Ennis & Jesionka, 2011). In fact, the Tetrad test is predicted to have relatively high power for an unspecified testing method – this high power has both led to a large-experimental comparison of the Tetrad test with the Triangle test (Garcia, Ennis, & Prinyawiwatkul, 2012) and to a discussion of circumstances under which the Tetrad test might reasonably serve as an alternative to the Triangle test (Ennis, 2012).

In this paper, we continue the development of the Tetrad test by considering the precision with which it measures sensory differences[1]. This consideration is important as ingredient or process changes can lead to positive but non-meaningful sensory differences (Ennis, 1990; MacRae, 1995). In addition, while Thurstonian methods allow us to estimate the sensory effect size, the estimator has variance (Bi, Ennis, & O'Mahony, 1997)[2]. Thus we consider precision from three perspectives. First, we compare the variance in the estimate of δ, the Thurstonian measure of sensory effect size, as measured by the Tetrad test to the corresponding variance of the Triangle and 2-AFC tests. We choose the Triangle test as it is a widely

[1] For a valuable discussion of other considerations of sensory measurement, most notably reliability and validity, see (Bi & Kuesten, 2012).

[2] In this paper, we refer to the variance of the estimator as the variance in the estimate, to be consistent with previous literature.

* Corresponding author. Tel.: +1 5305636647.
E-mail address: john.m.ennis@ifpress.com (J.M. Ennis).

0950-3293/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.foodqual.2013.05.003

used method for unspecified testing and we choose the 2-AFC test for comparison purposes as it is sometimes taken as the "gold standard" test that should be used when it is possible to name an attribute of interest (Dessirier & O'Mahony, 1998; Ennis & Jesionka, 2011). In this comparison, the Tetrad test will be shown to be typically more precise than the Triangle test, and sometimes more precise than the 2-AFC, at measuring sensory differences within the range of values likely to be meaningful for product testing purposes. Next we consider the expected widths of likelihood-based confidence intervals for the three methods – this consideration follows recent research proposing the use of likelihood-based confidence intervals in difference testing (Brockhoff & Christensen, 2010; Christensen & Brockhoff, 2009; Christensen, Lee, & Brockhoff, 2012). We then consider the power of the three tests for use in equivalence testing. Here we find that the Tetrad test is typical more powerful than the Triangle test, as expected, but also that there exist circumstances under which the Tetrad test is more powerful than the 2-AFC.

2. Variance in the Thurstonian estimate of sensory discriminal distance

Classically, difference testing has occurred in one of two ways – results are used to test whether or not there is a statistically significant difference (e.g. Peryam & Swartz, 1950) or, less often, to test whether or not there is statistical support for equivalence (c.f. Castura, 2010, for a recent review). Recently, however, an increasing acceptance of Thurstonian analysis (Bi et al., 1997; Brockhoff & Christensen, 2010; Ennis, 1990; Frijters, 1979; Lee & O'Mahony, 2007; Thurstone, 1927) has given sensory scientists a framework in which to consider difference testing as a form of sensory measurement[3].

According the Thurstonian perspective, the purpose of difference testing is not to detect whether or not a difference is present – if the samples are different then there will be a sensory difference. But the difference may not be consumer-relevant[4], and the goal of difference testing is to measure sensory differences accurately so as to support informed risk analyses.

In the Thurstonian framework, we suppose that the probability of a correct response is a function of a sensory effect size δ. The probability of a correct response is assumed to be a function of the this effect size, and the function that relates these two quantities is known as a psychometric function. See Fig. 1 for a comparison of the psychometric functions for the Tetrad, Triangle, and 2-AFC tests. The expressions for the functions were derived by Ennis et al. (1998), David and Trivedi (1962), and Thurstone (1927), respectively.

For the same products, a higher proportion of correct responses from the Tetrad test than from the Triangle test has been observed (Delwiche & O'Mahony, 1996b; Garcia et al., 2012; Masuoka, Hatjopoulos, & O'Mahony, 1995), and the superior ability of the 2-AFC to return correct answers in the presence of a sensory difference is well-established (c.f. Dessirier & O'Mahony, 1998). Since the guessing probability for both of these methods is 1/3, more correct responses means an increased probability of statistically significant difference and hence greater power in a difference test (c.f. Ennis, 1990, 1993; Ennis & Jesionka, 2011). Note that because the 2-AFC has guessing probability 1/2, it is not correct to compare the psychometric function of the 2-AFC to the psychometric functions for the other methods directly for the purposes of computing power.

In order to estimate sensory effect sizes from data, we typically use the method of maximum-likelihood and choose as our best estimate the estimate for the effect size that gives the highest likelihood given the data actually observed. For instance, suppose that, in a comparative experiment between the Tetrad, Triangle, and 2-AFC tests with 60 respondents, we obtain the data shown in Table 1.

The maximum-likelihood estimates of δ for the three methods, together with the variances in these estimates, are shown in Table 2. Even though all three methods return similar estimates of δ, the Tetrad method has the lowest variance in its estimate.

To understand the rank order of the variances shown in Table 2, we consider the corresponding relative likelihood functions. The likelihood functions are given by

$$L(\delta) = \binom{n}{k} P(\delta)^k (1 - P(\delta))^{n-k} \tag{1}$$

where $n = 60$ is the number of trials, k is the number of correct answers, and $P(\delta)$ is the psychometric function of the method under consideration. The relative likelihood functions, which are rescaled from the likelihood functions to have maximum value one, are shown in Fig. 2.

In this figure, we see that a δ value of approximately 1 maximizes the relative likelihood function given by the data in each case. But the relative likelihood function for the Triangle test shows a relatively wide range of possible values for δ values that have close to the same likelihood as the maximum-likelihood estimate. The relative likelihood functions for the 2-AFC and Tetrad test are more similar, with the Tetrad likelihood function being slightly more concave at its peak. Thus, in this case, we have greater confidence in this example in the estimate of δ from the Tetrad test than we do in the estimates from the Triangle or 2-AFC tests.

More generally, for each method, we compute (c.f. Pawitan, 2001)

$$Var(P(\delta)) = \left(\frac{\partial P}{\partial \delta}\right)^2 Var(\delta). \tag{2}$$

so that, if n is the sample size of the experiment, we compute the variance in the estimate of δ directly as

$$Var(\delta) = \left(\frac{\partial P}{\partial \delta}\right)^{-2} Var(P(\delta)) = \left(\frac{\partial P}{\partial \delta}\right)^{-2} \frac{p(\delta[1 - p(\delta)])}{n} \tag{3}$$

Thus the variance in the estimate depends only on a multiplier given by δ and the psychometric function, divided by the sample size (c.f. Bi et al., 1997). Fig. 3 shows these multipliers, or B values, for the Tetrad, Triangle, and 2-AFC tests as a function of δ. Note that tables for the B values used in this figure are given by Bi et al. (1997), for the Triangle and 2-AFC methods, and Ennis (2012) for the Tetrad method.[5] Since the B values for the Tetrad and Triangle test tend to infinity as δ goes to zero, we focus on the range $0.5 \leqslant \delta \leqslant 2.5$, i.e. the range in which consumer-relevant differences are likely to be found.

By comparing these B values, we see that the variance in the estimate of δ is uniformly less for the Tetrad test than it is for

[3] Thurstonian models and techniques for their statistical evaluation have been developed in a variety of settings (Bi, 2011a, 2011b; Bi & Ennis, 1998; Bi et al., 1997; Bockenhölt, 1992; Christensen & Brockhoff, 2009; Christensen, Cleaver, & Brockhoff, 2011; Christensen et al., 2012; David & Trivedi, 1962; Ennis & Ennis, 2013; Ennis et al., 1998, 1988; Ennis & Jesionka, 2011; Rousseau & Ennis, 2001) and the details of these models have been explored experimentally (Braun, Rogeaux, Schneid, O'Mahony, & Rousseau, 2004; Dessirier & O'Mahony, 1998; Garcia et al., 2012; Kim, Jeon, Kim, & O'Mahony, 2006; Lee & O'Mahony, 2007; Masuoka et al., 1995; Rousseau & O'Mahony, 1997; Tedja, Nonaka, Ennis, & O'Mahony, 1994).

[4] The problem of determining the size of consumer-relevant differences is one of the most important open problems in discrimination testing. See Ishii et al. (2007), Rousseau (2010), and Bi (2011b) for progress towards solving this problem.

[5] See (Bi & O'Mahony, 2013) for a compendium of B values for a variety of forced-choice sensory difference testing methods, including both the Specified and Unspecified Tetrad tests.