

Consideration of Thurstonian Scaling of Ratings Data

A.R. Warnock^{a,b}, A.N. Shumaker^a, J.F. Delwiche^{a,*}

^a Department of Food Science and Technology, The Ohio State University, 2015 Fyffe Court, Columbus, OH 43210, United States

^b Unilever HPC, 40 Merritt Boulevard, Trumbull, CT 06611, United States

Available online 21 March 2006

Abstract

In the analysis of sensory data, the use of parametric statistics, specifically analysis of variance (ANOVA), is standard. However, such parametric analyses often make assumptions that are not valid. Recent advances have made available an alternative analysis that does not make these assumptions, specifically Thurstonian Scaling of Ratings Data (SoR). This study compared these two methods of analyses on a single dataset. To further clarify the differences between analyses, assessments of a subset of the stimuli were also made in a triangle task. Findings indicate that the differences between ANOVA and SoR are minimal, with SoR tending to be slightly more conservative. Regardless of analyses used, triangle tests were found to be superior at differentiating the stimuli. The d' values determined using Thurstonian SoR did not agree with those determined with the triangle tests. Possible reasons for the discrepancy include boundary variance, actual strategy used during triangle tests, as well as other possible sources of variance.

© 2006 Elsevier Ltd. All rights reserved.

Keywords: ANOVA; Thurstonian Scaling of Ratings; d' ; Boundary variance

1. Introduction

All parametric statistics, including ANOVA, assume that the data being analyzed has equal intervals (i.e., is either interval or ratio data) and that the dataset has homoscedasticity, i.e., equal statistical variances (O'Mahony, 1986). When humans are asked to rate intensities, responses are encoded as numbers and such ratings are typically analyzed using parametric statistics, especially analysis of variance (ANOVA), despite the fact that the two assumptions mentioned above are often violated. For example, it is well established that when judges rate intensity or liking, they treat the numbers as though they are not equally spaced (Ennis, 1999; Kim & Sung, 1985; O'Mahony, 1986). They often hesitate to use the extreme ends of the scale, and in effect, the numbers at the ends are treated as though they are further apart than those in the center of the scale. In addition, the assumption of homoscedasticity is typically violated by extremely weak or strong sam-

ples, whose sample distributions cannot spread equally in both directions (altering the variance of the sample distribution) when the mean of the distribution is at or near the end of the scale (Ennis, 1999; O'Mahony, 1986). Thus, although ratings data are typically analyzed with ANOVA, the underlying data often does not meet the assumptions of the analysis, which can result in erroneous conclusions.

Recent advances in software (i.e., IFPrograms™) have made an alternative analysis available: Thurstonian Scaling of Ratings Data (SoR). This analysis estimates the best fitting true scale values (of equal intervals) by using the method of maximum likelihood. This method pre-dates ANOVA (both were initially developed by Ronald Fisher), and it assumes the best estimate of a parameter is one that gives the highest probability that the observed set of measurements will be obtained. Just as in ANOVA, Thurstonian SoR assumes that the data are normally distributed. An advantage of SoR is that it does not start with the same assumptions of equal intervals or homoscedasticity.

In addition, SoR converts ratings into d' values, a measure of distance between sample means that is theoretically method-independent. The underlying premise of both

* Corresponding author. Tel.: +1 614 247 6756; fax: +1 614 292 0218.
E-mail address: delwiche.1@osu.edu (J.F. Delwiche).

Thurstonian modeling and signal detection theory is that the momentary perception of a set stimulus can vary over time. With repeated sampling, sometimes the momentary perception will be higher and sometimes it will be lower. Most of the time, the perceived intensity will be at a particular value and the more extremely the momentary perceptual intensity deviates from this norm (by becoming lower or higher), the more the frequency will decline. It is further assumed that if the momentary perceived intensity is plotted against frequency of occurrence, it approximates a normal distribution. If two different stimuli are compared to one another, in both Thurstonian modeling and signal detection theory, the intensity of each stimulus can be represented by plotting a normal distribution along an intensity X -axis and a frequency Y -axis. The closer the two stimuli are to one another, the closer they will be to each other on the intensity axis. If the stimuli are confusable, then there will be overlap between these two distributions. The distance between the means of these distributions is a measure of the intensity difference between the stimuli and, when measured in terms of standard deviation, is known as d' (O'Mahony, 1992). Using the same logic, Thurstonian models can also be applied to discrimination tests (including triangle tests) to estimate d' (Ennis, 1999). As with Thurstonian Scaling of Ratings Data, these latter estimates rely upon the method of maximum likelihood. Thus it is theoretically possible to compare data collected from the different methods using d' as the common unit.

The main objective of this investigation was to compare the analyses of ANOVA with that of Thurstonian Scaling of Ratings Data. The similarity of conclusions from the two approaches was compared and the accuracy of each considered (i.e., which had results that most closely match perceptual differences as measured with triangle tests). A secondary objective that arose in the course of the investigation is to compare the d' values estimated from different methods, specifically from ratings and from triangle tests.

2. Experiment 1

2.1. Materials and methods

2.1.1. Materials

The solutions were made with 0.2% w/w cherry-flavored Kool-Aid™ containing different levels of sucrose (Sigma–Aldrich, St. Louis, MO), specifically: 4%, 6.5%, 9%, 10%, 11%, 13.5%, and 16% w/w sucrose. All solutions were prepared with room temperature spring water (Ice Mountain Spring Water Co., Hilliard, OH) at least 24 h prior to testing and used within 72 h of preparation.

2.1.2. Subjects

One hundred volunteer subjects (60 female, 40 male; 18–65 years of age; 91 non-smokers, 9 smokers) were recruited from the food science building and surrounding areas at The Ohio State University. In accordance with the approval of procedures by The Ohio State University

Office of Responsible Research Practices, all subjects gave informed consent.

2.1.3. Procedures

Testing was conducted in individual booths equipped with computer monitors, keyboards, and a mouse at each of 10 stations. Data was collected using Compusense® five version 4.6 software (Compusense, Inc., Guelph, Ontario, Canada).

Each panelist was asked to rate the sweetness intensity of the 7 different stimuli (varying in sucrose concentration) on an 18-point category scale where “1” was labeled “weak” and “18” was labeled “strong”. This unusual scale was chosen since preliminary investigations had shown that when a 10-point category scale was used, low intensity samples were crowded at the bottom of the scale (Hubbard & Delwiche, 2004; Shumaker, Warnock, & Delwiche, 2005). Thus, an expanded scale was used and sample concentrations were selected to avoid the extreme ends of the scale. Samples were presented in 20 mL aliquots in ~30 mL (1 oz) Plastic Soufflé Cups (Solo Plastic Soufflés, P100, Solo Cup Company, Baltimore, MS) affixed with random 3-digit codes and arranged in mini muffin trays. Instructions were given via on-screen displays. All seven samples appeared on the screen at once allowing panelists to change the sweetness ratings for all samples before continuing. Retasting was allowed. All samples were presented at room temperature (23–25 °C). Samples were counterbalanced across subjects. Panelists were provided with room temperature spring water (Ice Mountain Spring Water Co., Hilliard, OH) for rinsing; no specific rinsing instructions were given.

2.1.4. Statistical analyses

The sweetness ratings were analyzed in two different manners. The sweetness ratings were subjected to analysis by Thurstonian SoR Data using IFPrograms™. The best fitting d' values, or true scale values, were estimated from the rating data using the method of maximum likelihood. The d' values were then subjected to Bonferroni pairwise comparisons. Alternatively, the sweetness ratings were analyzed using ANOVA. Since Thurstonian SoR cannot account for subject differences, in order to make a fair comparison of the analysis methods a one-way ANOVA was used. Findings were followed by Bonferroni pairwise comparisons when ANOVA indicated a significant difference.

2.2. Results

The findings of the two analysis methods were quite similar (see Fig. 1). ANOVA indicates that there was a significant difference between samples in sweetness ($p < 0.01$). Bonferroni comparisons on the ratings after one-way ANOVA indicates most stimulus pairs were significantly different from one another ($p < 0.05$), with no significant differences being found between the comparisons of 9% and 10%, 10% and 11% and 13.5% and 16% ($p > 0.05$).

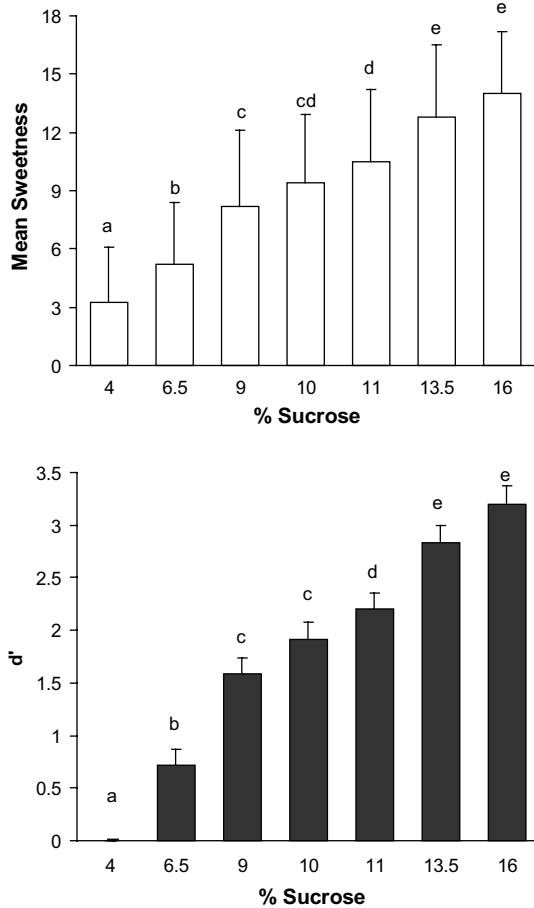


Fig. 1. Experiment 1, (Top) Mean sweetness ratings (+ standard deviation) of Kool-Aid solutions varying in sucrose concentration (Bottom) Scale sweetness values in d' (+ standard deviation) of the same stimuli. Bars within each plot with the same superscript are not significantly different (Bonferroni comparison, $p > 0.05$).

Thurstonian Scaling of Ratings Data was quite similar, with the Bonferroni comparisons of the scale values indicating that all but two stimulus pairs were significantly different ($p < 0.05$), with no significant differences being found between 9% and 10% or 13.5% and 16% ($p > 0.05$).

2.3. Discussion

Despite the differences in underlying assumptions, the findings of the two analyses were quite similar with only one discrepancy: SoR indicated a significant difference between 10% and 11% that one-way ANOVA did not. However, if one either uses a repeated-measures design or adds “subjects” as a second factor (Bonferroni comparisons for both is identical), the differences in the findings for SoR and ANOVA changes (see Table 1). When a repeated-measures ANOVA (rmANOVA) was conducted or when subject was used as a second factor (with no interaction), all but one stimulus pair, 10% and 11%, were found to be significantly different ($p < 0.05$). This means that rmANOVA (and two-way ANOVA) found significant

Table 1
Summary of Bonferroni comparisons for ANOVA and SoR analyses

Sucrose level (%)	One-way ANOVA	Repeated measures ANOVA	Thurstonian Scaling of Ratings
4	a	a	a
6.5	b	b	b
9	c	c	c
10	cd	d	c
11	d	d	d
13.5	e	e	e
16	e	f	e

differences between 9% and 10% and between 13.5% and 16% that SoR did not while SoR found a significant difference between 10% and 11% that rmANOVA did not. Currently, it is not possible to use a repeated-measures design with SoR, but it seems unlikely that the results of such an analysis would differ markedly from that of repeated-measures ANOVA.

It is difficult to interpret the meaning of these discrepancies. With these stimulus pairs, does the failure to find a significant difference indicate the method of analysis is overly conservative and thus prone to misses? Or does finding these significant differences indicate that the analysis method is overly sensitive and thus prone to false alarms? In other words, are there measurable perceptual differences between all stimulus pairs?

Thus, additional work was conducted to determine if, in fact, the stimuli are perceptually distinguishable. Triangle tests, which like all difference tests can detect smaller differences than can ratings (O'Mahony, 1995), were used to assess the perceptual distance between the sample pairs from this first study in a second study.

3. Experiment 2

3.1. Materials and methods

3.1.1. Materials

The solutions were made as they were in Experiment 1, although only 5 of the 7 sucrose levels were used: 9%, 10%, 11%, 13.5%, and 16% w/w sucrose. Again, all solutions were prepared with room temperature spring water (Ice Mountain Spring Water Co., Hilliard, OH) at least 24 h prior to testing and used within 72 h of preparation.

3.1.2. Subject selection

One hundred volunteer subjects (46 female, 54 male; 18–65 years of age; 87 non-smokers, 13 smokers) were recruited from the food science building and surrounding areas at The Ohio State University and all subjects gave informed consent.

3.1.3. Procedure

Half of this testing was conducted in same setting as was Experiment 1, using the same computerized real-time data collection system. However, data from 50 subjects was

collected at a remote location using paper ballots. Regardless of setting, each panelist was presented with six triangle tests, two replications of three triangles comparing 0.2% w/w cherry-flavored Kool-Aid™ at 9% and 10%, 10% and 11%, and 13.5% and 16% sucrose. The panelists were told two samples were the same and one was different in each triangle test, and then asked to choose the different sample. All samples were presented at room temperature (23–25 °C). Samples were counterbalanced across subjects and the design was blocked so that each panelist received the replicated triangles in sets. For example a panelist might have seen both of the 9% and 10% triangle tests in succession, then both of the 13.5% and 16% triangle tests in succession, and then finally both of the 10% and 11% triangle tests in succession.

As in Experiment 1, the samples were presented in 20 mL aliquots in ~30 mL (1 oz) Plastic Soufflé Cups affixed with random 3-digit codes arranged in mini muffin trays. Instructions were given via on-screen displays or via written instructions in the paper ballots. The random 3-digit codes affixed to the cups corresponded to written instructions indicating which sample to taste next. The panelists saw each triangle test either separately on different display screens or all at once on the paper ballot. Panelists were provided with room temperature spring water for rinsing; no specific rinsing instructions were given.

3.1.4. Statistical analysis

The triangle test data for each comparison was analyzed using the beta-binomial model to determine significance and which model, binomial or beta-binomial, was more appropriate (according to chi-square). This assessment was done using IFPrograms™, which was also used to determine the d' values (with the method of maximum likelihood) for each assessed stimulus pair.

In addition, the d' values from the triangle tests in the second experiment were compared to those from the ratings in the first experiment. To compare the d' values from the ratings (Experiment 1) with those from the triangles, the difference between ratings d' values of the stimulus pairs were taken. The triangle and SoR d' values were compared to each other with the d' test (which relies upon chi-square) using IFPrograms™.

3.2. Results

According to Chi-square tests, the comparison of 9% and 10% and of 10% and 11% were best fit by the binomial model, while the comparison of 13.5% and 16% was best fit by the beta-binomial model. No significant difference was found between 9% and 10% in the triangle test ($p = 0.0571$), which is in agreement with the findings of SoR and one-way ANOVA and in contrast with the findings of rmANOVA. A significant difference was found between 10% and 11% in the triangle test ($p = 0.0110$), which is in agreement with the findings of SoR and in contrast with one-way ANOVA and rmANOVA. A significant

Table 2

SoR d' values (covariance) and triangles d' values (variance) with d' test p -values

Stimulus pair	SoR d' values	Triangle d' values	d' test p -values
9% and 10%	0.332 (0.015)	0.810 (0.067)	0.10
10% and 11%	0.280 (0.016)	0.950 (0.054)	0.01
13.5% and 16%	0.364 (0.018)	0.910 (0.056)	0.04

difference was found between 13.5% and 16% in the triangle test ($p = 0.0243$), which is in agreement with the findings of rmANOVA and in contrast with the findings of SoR and one-way ANOVA.

As predicted, the d' test found that for the 9% and 10% sucrose Kool-Aid™ solutions the d' value of the triangle test was not significantly different from the d' value from the SoR ($p = 0.10$). However, in contrast to initial expectations, the d' test found that the d' values of the triangle tests were significantly different from the d' values from SoR for the 10% and 11% and the 13.5% and 16% sucrose Kool-Aid™ solutions ($p = 0.01$ and $p = 0.04$, respectively, see Table 2).

3.3. Discussion

The results of the triangle test were equivocal, with the overall results of the triangle tests differing from those of SoR, one-way ANOVA, and rmANOVA. Thus, the results did not elucidate which method of analysis was the most accurate. In addition, d' did not appear to be method independent, with the values of d' showing significant differences between those determined using SoR and those determined from triangle tests showing significant differences. However, upon further consideration, it was realized that this apparent discrepancy could be explained by the difference in subjects participating in the first and second studies. Since the subject pools were not identical, their inherent sensitivities also would have differed, potentially accounting for the different values found between the first and second studies.

A third study was therefore conducted in an attempt to (1) further compare the accuracy of ANOVA and SoR (i.e., which had results that most closely match true perceptual differences as measured with triangle tests) and (2) investigate the interchangeability of d' across methods given a stable subject pool.

4. Experiment 3

4.1. Materials and methods

4.1.1. Materials

The stimuli were made as they were in Experiments 1 and 2. Again, all solutions were prepared with room temperature spring water and 0.2% w/w cherry-flavored Kool-Aid™ at least 24 h prior to testing and used within 72 h of preparation.

4.1.2. Subject selection

One hundred volunteer subjects (58 female, 42 male; 18 – greater than 65 years of age; 97 non-smokers, 3 smokers) were recruited from the food science building and surrounding areas at The Ohio State University and all subjects gave informed consent.

4.1.3. Procedure

Testing was conducted in the same setting as was Experiment 1, using the same computerized real-time data collection system. Each panelist first rated the sweetness of the 7 stimuli (4%, 6.5%, 9%, 10%, 11%, 13.5% and 16% sucrose cherry-flavored Kool-Aid™) and then completed six triangle tests, two replications of three triangles at the same levels used in Experiment 2 (comparing cherry-flavored Kool-Aid™ at 9% and 10%, 10% and 11%, and 13.5% and 16% sucrose).

4.1.4. Statistical analysis

Analyses matched those used in Experiments 1 and 2. Ratings were assessed with both ANOVA and SoR. Triangles were assessed with the beta-binomial test. The *d'* values from SoR and triangles were compared using the *d'* test in IFPrograms™.

4.2. Results

For this dataset, the findings of one-way ANOVA, repeated-measures ANOVA and SoR were in complete agreement with one another. Significant differences were found between all levels except between 6.5% and 9% and between 13.5% and 16% (see Fig. 2). In contrast, the triangle test found a significant difference between all stimulus pairs tested: 9% and 10% (binomial model, $p < 0.0001$), 10% and 11% (binomial model, $p < 0.0001$), and 13.5% and 16% (beta-binomial model, $p < 0.0001$).

The *d'* values determined from SoR were compared with *d'* values from triangles (see Table 3). Similar to what was found earlier, the *d'* test indicated significant differences across methods for 9% and 10% and for 13.5% and 16% ($p < 0.01$).

5. General discussion

In both Experiments 1 and 3, the differences in findings of SoR and ANOVA were minimal or non-existent, suggesting that these methods of analysis are roughly equivalent. The findings of the first study suggests that SoR may be more likely to find significant differences than one-way ANOVA, while repeated-measures ANOVA appears to be more likely to find significant differences than both. Comparison with the results of the triangle tests suggest that none of the significant differences revealed by SoR and ANOVA were false alarms; in fact, triangle tests detected more significant differences than all analysis of ratings of the same stimuli in Experiment 3. From these studies we can conclude that neither method of analysis,

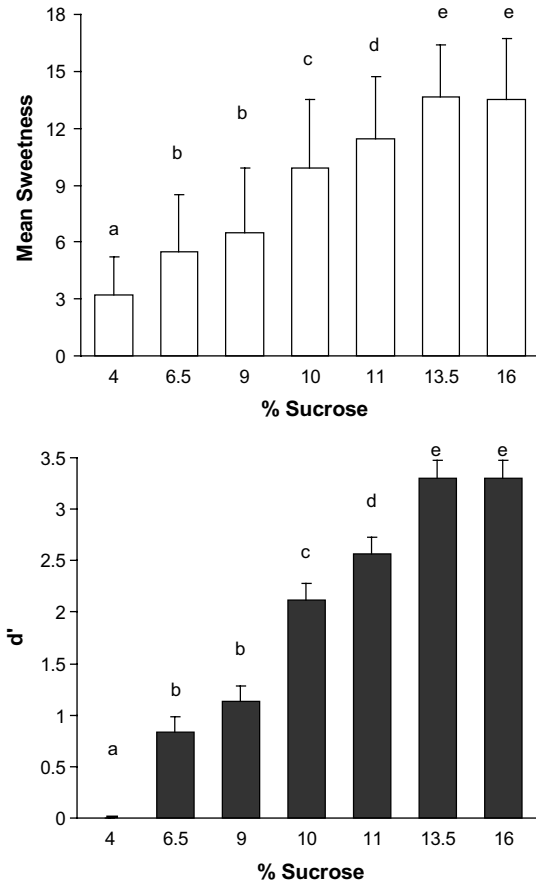


Fig. 2. Experiment 3, (Top) Mean sweetness ratings (+ standard deviation) of Kool-Aid solutions varying in sucrose concentration (Bottom) Scale sweetness values in *d'* (+ standard deviation) of the same stimuli. Bars within each plot with the same superscript are not significantly different (Bonferroni comparison, $p > 0.05$).

Table 3

SoR *d'* values (covariance) and triangles *d'* values (variance) with *d'* test *p*-values

Stimulus pair	SoR <i>d'</i> values	Triangle <i>d'</i> values	<i>d'</i> test <i>p</i> -values
9% and 10%	0.985 (0.013)	3.090 (0.040)	<0.01
10% and 11%	0.441 (0.016)	1.360 (0.360)	0.13
13.5% and 16%	0.001 (0.019)	2.000 (0.031)	<0.01

one-way ANOVA and Thurstonian SoR, is superior to the other – they are largely equivalent despite the different assumptions each makes. And while a repeated-measures ANOVA was able to detect more significant differences than either of these methods of analysis, the triangle test succeeded in detecting significant differences that the repeated-measures ANOVA of the sweetness ratings did not.

It is well established that given a particular set of subjects and stimuli, discrimination tasks are superior at distinguishing small differences between stimuli than rating tasks. Thus, it is not surprising that in Experiment 3 the triangle tests detected differences that the ratings did not (regardless of analysis method used). What was surprising

Table 4

SoR d' values (covariance) and 3-AFC d' values (variance) with d' test p -values

Stimulus pair	SoR d' values	3-AFC d' values	d' test p -values
9% and 10%	0.985 (0.013)	1.630 (0.017)	<0.01
10% and 11%	0.441 (0.016)	0.490 (0.013)	0.77
13.5% and 16%	0.001 (0.019)	0.900 (0.014)	<0.01

is that the d' values determined from the triangle tests differed significantly from those determined from SoR. A measure of perceptual distance between samples expressed in standard deviations, d' is theoretically independent of method used to assess perceptual distance (i.e., triangle test, ratings, 2-AFC, same-different test, etc.).

One possible reason for the discrepancy in d' is that the triangle d' values were estimated on a false assumption – that subjects were using a tau criterion in their judgments. Generally in a triangle test, the subject is attempting to determine which stimulus is furthest away from the other two stimuli, invoking a strategy that examines distances between stimuli and relies upon a tau criterion (O'Mahony & Rousseau, 2003; Rousseau, 2003). In contrast, for the 3-AFC (where one is instructed to find the sample strongest (or weakest) in a particular attribute), the subject is attempting to determine which of the stimuli is the strongest, invoking a strategy that looks at intensity and relies upon a beta criterion (O'Mahony & Rousseau, 2003; Rousseau, 2003). For a given d' and set of stimuli, subjects will give more correct responses for 3-AFCs than for triangle tests (O'Mahony & Rousseau, 2003; Rousseau, 2003).

In Experiment 3, subjects first rated sweetness and then conducted the triangle tests. They were likely already focused upon sweetness and despite their instructions to select the odd sample may have adopted a strategy that relied upon a beta criterion. Since d' is estimated from the number of correct responses, and they have been using a strategy that relied more upon a beta criterion, it is possible that the estimated d' values for the triangle test were inflated. Although it is not entirely clear what strategy the subjects were using, for comparison d' values also were estimated from the triangle data as though it came from 3-AFCs (Table 4). While it did decrease the estimated d' values, there was still a significant difference between these d' values and those estimated using SoR for 9% and 10% and 13.5% and 16%.

The difference between the triangle d' values and SoR d' values can also be accounted for by boundary variance. Simply put, the more boundaries there are, the more variance there will be (O'Mahony, Park, Park, & Kim, 2004). As variance (and thus standard deviation) increases, even if the perceptual distance between stimuli remains unchanged, d' will decrease since its unit of measure (standard deviation) is increasing. In this investigation, an 18-point category scale was used for the ratings, resulting in 17 boundaries. In contrast, the triangle test had two categories and thus only one boundary. Considered in this light, it is not surprising that d' values estimated from triangle tests were larger than those estimated from Thurstonian Scaling of Ratings. This difference in boundary variance also accounts for the general greater sensitivity of discrimination tests over ratings. So while in theory d' is independent of methods, it in fact is only method-independent when variance is equal across methods, which in practice is quite rare.

References

- Ennis, D. (1999). Thurstonian Models for Intensity Ratings. Richmond, VA. The Institute For Perception, 1–2.
- Hubbard, A. R., & Delwiche, J. F. (2004). Comparison of Thurstonian probabilistic modeling of rating data with standard parametric statistics. Abstract No. 48–6. IFT Annual Meeting Book.
- Kim, W. J., & Sung, H. S. (1985). Effects of temperature and sugar addition on the flavor of ginseng tea. *Korean Journal of Food Science*, 17(4), 304–310.
- O'Mahony, M. (1986). *Sensory evaluation of food*. New York: Marcel Dekker, Inc.
- O'Mahony, M. (1992). Understanding discrimination tests: A user-friendly treatment of response bias, rating and ranking R -index tests and their relationship to signal detection. *Journal of Sensory Studies*, 7, 1–47.
- O'Mahony, M. (1995). Sensory measurement in food science: fitting methods to goals. *Food Technology*, 4(4), 72–82.
- O'Mahony, M., Park, H., Park, J. Y., & Kim, K.-O. (2004). Comparison of the statistical analysis of hedonic data using analysis of variance and multiple comparisons versus an R -index analysis of the ranked data. *Journal of Sensory Studies*, 19(6), 519–529.
- O'Mahony, M., & Rousseau, B. (2003). Discrimination testing: A few ideas, old and new. *Food Quality and Preference*, 14(2), 157–164.
- Rousseau, B. (2003). Sensory difference testing. In B. Caballero, L. Trago, & P. Finglas (Eds.), *Encyclopedia of food science and nutrition* (pp. 5141–5147). Academic Press.
- Shumaker, A. N., Warnock, A. R., & Delwiche, J. F. (2005). Comparison of Thurstonian probabilistic modeling of rating data with parametric statistics: Part II. IFT Annual Meeting Technical Program, New Orleans, LA.