Thurstonian Scaling for Difference Tests

Daniel M. Ennis

Background: Suppose that an ingredient supplier has changed a key ingredient used in the manufacture of your company's chocolate chip cookies. Based on the baking chemistry of this new ingredient, there is reason to suspect that the new ingredient will make your cookies harder. Additionally, recent market research data has connected cookie hardness with consumer liking. For these reasons, you decide to conduct a test to determine what effect the change in ingredient will have on your cookies. After some deliberation, you decide to use the duo-trio method to determine whether cookies made using the new ingredient are perceptibly different from cookies made using the current ingredient.

One hundred experienced panelists are assembled, and these panelists are divided into two groups. Panelists in the first group each receive a sample from the current production to use as a reference, while panelists in the second group each receive a sample from the new production for the same purpose. Panelists are then presented with samples from the current and new productions, and are instructed to choose the sample most similar to the reference.

From this test you are unable to confirm a difference between the cookies from the new and the current productions. Given your knowledge of the new ingredient's baking chemistry, you are perplexed by this result, and decide to conduct a second, smaller test. You suspect that by specifying the attribute in question, namely hardness, a difference between the new and current production cookies will be detected. For this reason you decide to use the 2-alternative forced choice (2-AFC) method. Thirty of the previous 100 panelists participate in the test, and in a counterbalanced design each panelist is presented with a sample from the new production and a sample from the current production. The panelists are asked to identify which sample is harder. When the test is analyzed, you find that the new ingredient makes your cookies significantly harder.

Table 1. Results of difference tests using the duo-trio and the 2-AFC methods.

Method	Proportion of Correct Responses (P_c)	Sample Size	Test of the Guessing Model
Duo-trio	0.55	100	NS
2-AFC	0.71	30	S
NS: Not Significant at $\alpha = .05$		S: Significant at $\alpha = 0.05$	

Gridgeman's Paradox: These experiments and results, typical of routine tests conducted in consumer product testing, contain a profound message for the interpretation of product testing results. The two methods employed have exactly the same null hypothesis, or guessing model, but show large differences in sensitivity to the products tested. This type of result, referred to as Gridgeman's Paradox[1,2], has been extensively discussed over the last 45 years. A common misconception is that the difference between the 2-AFC and the duo-trio methods occurs because an attribute has been specified in the 2-AFC test. Experience demonstrates, however, that Gridgeman's Paradox would arise even if an attribute were specified in the duo-trio test. For this reason other resolutions of the paradox must be sought.

Perceptual Variability and Decision Rules: The duo-trio and the 2-AFC methods share a common guessing model with a probability of correct response = 0.5. The difference between the two methods lies in the decision rules that are used to produce responses. Returning to the cookie example, a 2-AFC correct response occurs when a sample from the new production is harder than a sample from the current production. How often this occurs determines the proportion of correct responses, P_c. Contrastingly, a duo-trio test with a sample from the new production as reference yields a correct response when the new production sample is more similar to the reference than the current production sample. How do we measure "more similar?" One approach is to base the decision on perceptual distances so that a correct response occurs when the distance between the new production sample and the reference is less than the distance between the current production sample and the reference. How do incorrect responses arise? We suppose that the perceptual magnitudes (hardness in this case) for the two products follow normal distributions with different means but unit variances. The difference between the means of these distributions is called ι and its estimate, dℜ The units of ι are perceptual standard deviations. Variances in perceptual magnitudes explain why sometimes the current production sample may appear harder than the new production, even when the new production hardness mean is higher.

Thurstonian Models and Psychometric Functions: Thurstonian models require the two assumptions we have just made: 1) Perceptual variability exists and can be modeled using the normality assumption, and 2) methods have associated decision rules.

Figure 1. Psychometric functions for the 2-AFC and duo-trio.

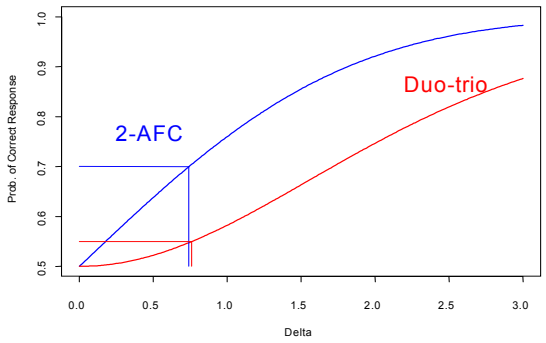


Figure 1 shows the psychometric functions for the 2-AFC and the duo-trio methods. Psychometric functions link the probability of a correct response to ι. For instance, if ι is 1, the probability of a correct response for the 2-AFC is 0.76. Assuming that one perceptual dimension is used in the duo-trio decision, we can use Figure 1 to see that for ι = 1, the probability of a correct response for the duo-trio method is 0.58. Marked on Figure 1 are the results from Table 1 for each of the methods. Although the P_c values for these methods are dissimilar, the estimated ι values are almost identical. The difference between the P_c values for the duo-trio and the 2-AFC methods for similar or the same ι values is due to the difference between the two methods' decision rules. Note that consideration of the decision rules has yielded this result; specification of attributes has not been mentioned.

Comparison of d′ Values From Different Methods: From psychometric functions or tables, we can convert P_c values into estimates of ι values, or d′s. These estimates have variances that can be obtained either from tables[3] or by direct computation using the method of maximum likelihood. Table 2 shows the results of the cookie example in terms of d′ values.

Table 2. Results of difference tests in terms of d′ values.

Method	Proportion of Correct Responses (P_c)	d′	Variance of d′
Duo-trio	0.55	0.76	0.16
2-AFC	0.71	0.74	0.12
Test for equivalence of d′ values: Not significant at $\alpha = 0.05$			

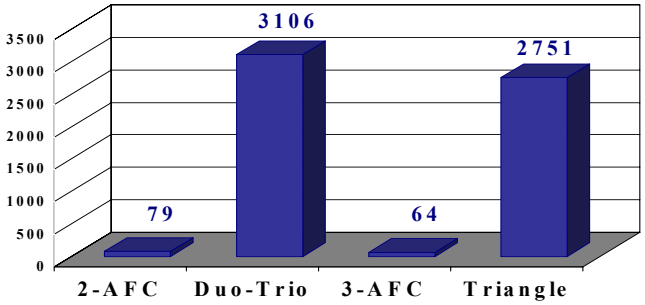
From these results we can determine whether there is a significant difference between the d′ values obtained using the two methods. This table shows no significant difference between the two d′ values. Combining the results of the two tests, we find that the new ingredient imparts a difference, probably hardness, and the perceptual difference is 0.75 ± 0.51 using 95% confidence intervals

Power: Given that there was a difference between the cookies from the new production and the cookies from the current production, why did the duo-trio method fail to detect the difference? To answer this question, we consider Table 2. Although the ι values for the duo-trio and the 2-AFC methods are similar, the P_c values are 0.55 and 0.71 respectively. As the two methods share a common guessing model, the P_c values for each method are compared to the same value to decide significance. Given this information, we expect the 2-AFC method to declare a significant difference in this situation more often than the duo-trio method.

This result holds in general, and we are able to say that the 2-AFC method is more powerful, or will declare a given d′ difference significant more often, than the duo-trio method.

Figure 1 verifies this result. For any ι value greater than 0, the probability of a correct response for the 2-AFC method is greater than that for the duo-trio method.

Figure 2. Sample size needed for 80% chance of detecting a d′ of 0.5 (64:36 split) at an ζ of 0.05.



Sample Size Requirements for a Given Power: Figure 2 shows the sample sizes required to be 80% certain of detecting a ι of 0.5 at an ζ = 0.05 for four different methods. In addition to the 2-AFC and duo-trio methods, Figure 2 shows the 3-AFC and the triangular methods. The 3-AFC method is similar in instructions to the 2-AFC, but presents the subject with two products that are the same and one that is different. In the triangular method, there are two products that are the same and one different and the instruction is to choose the odd sample.

Conclusion: The ideas discussed here apply to a situation in which there is a single attribute that subjects attend to in making decisions. Both the duo-trio and the triangular methods are not restricted in their applications to unidimensional perceptions. For this reason, the excellent agreement among the methods shown here may not always occur. A lack of agreement among the methods may, in fact, suggest multidimensionality. Sequence effects may also cause disagreement among methods. To deal with these issues, multidimensional Thurstonian models[4] and models for sequence effects[5] have been developed. These more sophisticated models are often unnecessary, however. Numerous practical applications of Gridgeman's Paradox have been observed, and the resolution of the paradox using Thurstonian models has underscored the importance of decision rules and variability in modeling choice experiments.

References:

1. Frijters, J.E.R. (1979). The paradox of the discriminatory non discriminators resolved. *Chemical Senses and Flavour*, **4**, 355.
2. Ennis, D.M. (1993). The power of sensory discrimination methods. *Journal of Sensory Studies*, **8**, 353-370.
3. Bi, J., Ennis, D.M., and O'Mahony, M. (1997). How to estimate and use the variance of d′ from difference tests. *Journal of Sensory Studies*, **12**, 87-104.
4. Ennis, D.M. (1998). Foundations of sensory science and a vision for the future. *Food Technology*, **52**, 78-89.
5. Ennis D.M. and O'Mahony, M. (1995). Probabilistic models for sequential taste effects in triadic choice. *Journal of Experimental Psychology: Human Perception and Performance*, **21**, 1-10.