Reprinted from IFPress (2005) 8(3)2,3

Relative Scales and Difference Testing Norms

Daniel M. Ennis

Background: Appropriate treatment of "no difference" responses in paired preference and difference testing is a persistent issue in product testing. Different approaches involving their inclusion or omission from subsequent testing using the binomial distribution have been proposed. In 1980, paired testing of four blind labeled brands of a major consumer products company was conducted to establish difference and preference testing norms for identical products. Products were manufactured at the company's main manufacturing plant and product from the same production run for each brand was divided into two samples for paired testing. Products and a ballot were mailed to 600 category users of each brand (a total of 2400 consumers) and the return rates were 69% to 81% (a total of 1787 completed ballots.) The results of that research on identical products showed a narrow range of expected preference results (%) from 39.7:39.7:20.6 (prefer A: prefer B: no preference) to 40.8:40.8:18.5, depending on the brand, with a mean of 40:40:20. Females showed a slightly higher likelihood of choosing the "no preference" response. Although norms can be established by testing identical products in this manner, they can also be predicted from modeling routine paired product testing data as will be discussed in this report.

The "no difference" version of the paired test is a special case of a relative rating method and the same model can be used to analyze data from relative-to-reference scales and just-about-right scales. It may seem surprising that three methods differing so much in instructions and objectives share a common underlying model. It will be shown how the model interprets data from these methods and how the results of the analysis can be used to provide guidance for the development of norms for future testing.

Relativity of Difference Responses: In a previous technical report[1], it was shown that the just-about-right (JAR) scale is a relative scale in which the reference point is an ideal product. From the analysis of JAR data it was shown how the expected distribution of responses to an ideal product could be obtained and used to make product testing decisions. In a difference or preference test, the instruction is to choose the product with the greatest (or least) intensity of some attribute or to express a preference for one or the other product. If a "no difference" option is available, the task can be viewed as one involving a choice among three categorical alternatives. In this case one of the products acts as a reference and the other product is rated higher, lower or not different from the reference. Conceptually this is similar to using the JAR scale, except that the reference product is explicitly provided in the test whereas, in using the JAR scale, the reference product is an implicit ideal value. If two products, A and B, are presented and the instruction is to choose the sweetest product or declare "no difference", the choice of A could correspond to a relative rating of 1 (B is less sweet), 2 (no difference) or 3 (B is sweeter than A.)

We assume that the reference product follows a normal distribution and that decision boundaries are used to decide the rated intensity of the alternative product (B in this example.) Figures 1a and 1b illustrate these ideas. In Figure 1a, evaluation of product A produces the sample value **a** and around this value two symmetric boundaries are formed. If a sample value from the B distribution falls within these boundaries, the products are declared to be not different as shown in Figure 1a. In Figure 1b, a new sample of the A product, **a**, is obtained and a value from the B distribution, **b**, is found to fall to the right of the uppermost boundary value. In this case, the B sample is declared to be sweetest. Notice that the B sample has the same intensity in these figures, but was rated differently in the two examples. The goal of modeling data obtained from this type of paired test with a "no difference" option is to find the location of the mean for the B distribution relative to the A distribution and to find the location of the decision boundaries. Once the boundary values are known, norms for products with identical means can be established so that a null hypothesis for difference or preference testing with the "no difference" option can be stated. The lack of a basis for this hypothesis is a fundamental statistical problem in the analysis of this type of data.

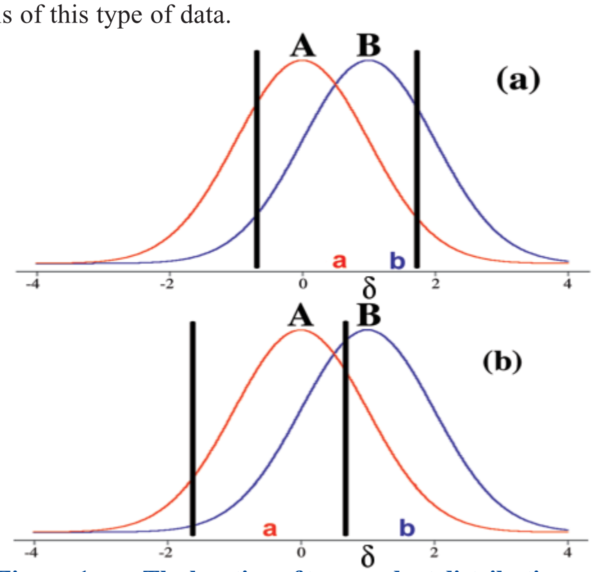

Figure 1. The location of two product distributions, A (at zero) and B at δ. The vertical lines are decision boundaries

Prefer A	Prefer B	No Preference
79	81	40
70	100	30
75	90	35
90	78	32
120	60	20

Table 1. Data from five preference experiments in which a prototype (B) is compared to a common control (A).

Scenario: A demographically balanced sample of 200 consumers each evaluated five prototype package designs and a common control in a series of randomized paired preference tests with a "no preference" option. The results are given in Table 1. These results show that the proportion of "no preference" responses appear to vary depending on the degree of difference between the prototypes and the control. Your goal is to quantify and test this degree of difference while including the "no preference" responses in the analysis.

How Relative Scale Responses Arise: The numbers in Table 1 are frequencies which are used to estimate the probabilities of the three possible outcomes (A preferred, B preferred, and no preference.) These probabilities depend on the relative location of the means to each other and the location of the boundaries. A standard method for measuring the degree of difference between the means is called δ and its units are standard deviations. Since δ is measured on an interval scale, one of the means can be set to zero, so that δ is the mean of the alternative distribution. When two means differ by 0.5 units of δ, it means that the products differ by 0.5 standard deviations in perceptual intensity. The experimental estimate of δ is called d′. Once a particular sample of the A product has been taken, the response on a trial depends on where a sample of the B product falls relative to boundaries that are set up around the A sample. These boundaries are sometimes called "relative boundaries" because their location varies depending on the value drawn from the A distribution. For instance, if the **a** value (taken from a normal distribution with mean 0 and variance 1) is 0.5 and the boundaries are placed ± 1 unit from this value, the boundaries will occur at -0.5 and 1.5. Any **b** sample that falls between these two limits will be assigned a "no difference" or "no preference" response. Clearly, the likelihood of this occurring will depend on how close the mean for the B product is to the A product placed at zero. However, if the B product is exactly placed at zero, it is still possible to generate samples from the B distribution that would fall outside boundary values and produce "A preferred" or "B preferred" responses. In Table 1 the first paired test provides an almost equal number of "A preferred" and "B preferred" responses, but the "no preference" category includes only 40/200 or 20% of the responses.

Prototype	d'	Variance
1	0.018	0.013
2	0.274	0.014
3	0.135	0.014
4	-0.109	0.014
5	-0.569	0.014
Relative Boundary: 0.268		

Table 2. Results of modeling Table 1 in which d′ values, their variances and a relative boundary are estimated.

Analysis of Table 1 and the Development of Norms: Table 2 shows the relative locations of the prototype means to the control (d' values) and their variances. A multiple d' test[2] to check for differences among the values found a highly significant χ^2 value (29.79 compared to the critical value of 9.49 for 4 degrees of freedom.)

Table 2 shows that the location of the relative boundary is 0.286. Since we assume that the boundaries are symmetric about the reference product, only one value is needed for three categories. For instance, if a reference value (from the A distribution) is 0.5, then any values from the B distribution that fall between 0.214 and 0.786 will be declared to be not different from A. For this particular value from the A distribution, values from B greater than 0.786 will indicate that B is preferred over A and values less than 0.214 will indicate that A is preferred over B. If we assume that the means of A and B are equal, then this boundary value can be used to compute the expected proportions for identical products. In this case the best estimates of these proportions are 42%:42%:16% corresponding to prefer A: prefer B: no preference.

Using the Preference Norm: The value of positing a preference norm is that future test results can be compared to it. If, for instance, a preference test is conducted with 400 consumers and the preference results are 45% prefer A, 45% prefer B and 10% no preference, the traditional methods of analyzing this data would conclude that there is no evidence that one product was preferred to the other. However, if the expected outcome is 42%:42%:16%, a chi-square test of the results compared to this norm is highly significant ($\chi^2 = 10.7$, $p < 0.01$, df = 2). A likely cause of the difference between the two analyses is that there are latent subgroups or segments that differ with respect to preference for the products. For this reason, analysis using the testing norm can determine the existence of latent segments.

Conclusion: Using the idea that difference and preference tests are relative scales, results from these tests can be used to estimate and test product differences and preferences. In addition, the analysis can be used to provide an estimate of the norm for the case where products are identical. The norm provides a useful basis for comparing product test results to that expected under an hypothesis of no difference and opens up the possibility of providing guidance on the existence of latent segments. All of the modeling discussed in this report can be conducted using the *IFPrograms*™ software.

References (available at www.ifpress.com)

1. Ennis, D.M. (2003). Just-About-Right Scales. *IFPress*, **6**(3), 2- 3.
2. 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.