Reprinted from IFPress (1999) 2(3)2,3

Thurstonian Models for Intensity Ratings

Daniel M. Ennis

Background: Products and concepts are often evaluated on rating scales to quantify degree of liking, level of purchase interest, intensity of an attribute, degree of difference or level of agreement with a statement. Rating scales are constructed in many different ways, but usually involve options labeled with numbers, words or symbols. In some cases a particular option has a special label, such as "just right" or "same as reference." Scales may involve labels for every option, such as the words on the 9-point acceptability scale; other scales have labels only at the end points.

Rating responses are encoded as numbers. These numbers correspond to perceived intensities and express the order of these intensities, but ratings data may not possess the properties of the numbers used to encode them. For instance, if we encode the categories on the 9-point acceptability scale as integers from 1 to 9, we may assume that "3" corresponds to a higher liking value than "2" and that "9" corresponds to a higher liking value than "8." Nevertheless we cannot assume that the difference between a "3" response and a "2" response is the same as the difference between a "9" response and an "8" response. We characterize this deficiency by saying that ratings data do not exhibit equal interval properties. One reason that the equal interval assumption cannot be applied to ratings data is that certain rating options may be avoided or favored. For example, on the 9-point acceptability scale the 3rd and 7th categories corresponding to "dislike moderately" and "like moderately" are favored and the 5th category, "neither like nor dislike" is avoided. This means that there is a smaller intensity range corresponding to a "5" response than to a "3" or to a "7." Consequently, the difference between an average "7" intensity and an average "6" intensity will be larger than the difference between an average "6" intensity and an average "5" intensity. Moreover, since we are only interested in differences between values; adding a fixed number to all values on an equal interval scale produces an equivalent scale. It follows that the zero point of an equal interval scale is arbitrary.

Scenario: You are interested in the relative sweetness intensities of two beverage products; one of these products is your regular brand and the other is a modification. Two experiments have been conducted. In one experiment a numerical 7-point rating scale was used, and in the second experiment a 5-point word-category scale was used. Both experiments were conducted with 100 heavy users of your brand. The numerical scale consisted of the numbers "1" to "7" with "1" labeled as "not sweet" and "7" labeled as "extremely sweet." The remaining numbers were not labeled. The word-category scale consisted of 5 options: "not sweet," "slightly sweet," "moderately sweet," "very sweet" and "extremely sweet." The results of these two experiments are shown in Tables 1a and 1b.

Your question is: Do these two experiments reflect the same sweetness intensity difference?

Table 1a. Rating frequencies for two beverage products on a 7-point numerical scale.

Category Label	"1"	"2"	"3"	"4"	"5"	"6"	"7"
Current Brand	4	10	40	34	9	2	1
Modification	0	2	21	41	23	9	2

Table 1b. Rating frequencies for two beverage products on a 5-point word category scale.

Category Label	"Not Sweet"	"Slightly Sweet"	"Moderately Sweet"	"Very Sweet"	"Extremely Sweet"
Current Brand	7	24	62	4	3
Modification	1	9	66	13	11

The Problem: To answer your question it is important to realize the distinction between perceived sweetness values and rated sweetness values. Perceived intensity can be placed on an equal interval scale while, as previously noted, rated intensities may not have equal interval properties. A problem arises: How can we measure sweetness intensities on an equal interval scale when all we have are non-equal interval ratings?

A Solution: Represent the sweetness values that the consumer experiences as values from a normal distribution with mean 0 and standard deviation 1. The zero mean is consistent with the arbitrary zero point of an interval scale, and the perceptual standard deviation is the unit of measurement. This representation leads to an equal interval intensity scale for sweetness. Figure 1a illustrates how consumers would produce the ratings given in Table 1a. For a 7-point scale, the consumer uses 6 points on the sweetness axis to decide how to rate a particular product. These points are marked by vertical lines and are called decision boundaries. If a sweetness value falls between two of these decision boundaries, the consumer responds with a rating appropriate to that interval. For example, the area between the first decision boundary and the second illustrates the probability of a "2" response.

Rating Means and Scale Means: Rating means are the means of the rating values, while scale means are the means of the intensity distributions, like sweetness, on an equal interval scale. Scale means are expressed relative to the scale means of a particular product. Accordingly, this product will have a scale mean of zero. Since the zero of an interval scale is arbitrary, it is immaterial which product we choose as reference. Differences between scale means are called δ (delta) values. Rating variances depend on perceptual variances but are also affected by the location of decision boundaries. For instance, it is well known that rating variances are generally lower for large rating means than at more intermediate values. This occurs because the number of rating options, determined by the decision boundaries, becomes reduced as the mean increases above intermediate values. The size of the rating variance may have little or no relationship to the size of the perceptual variance.

Figure 1a. Product distributions and decision boundaries on an interval scale modeled from 7-point numerical ratings.

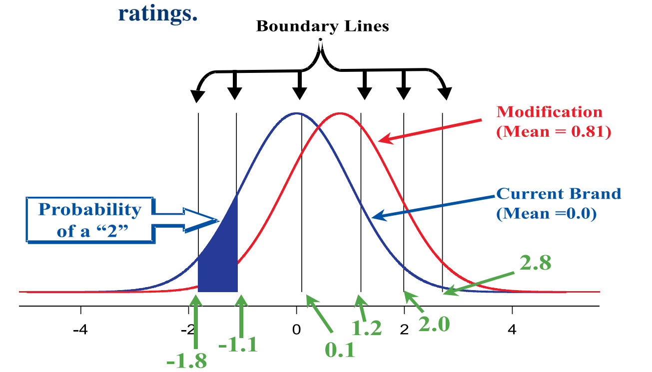
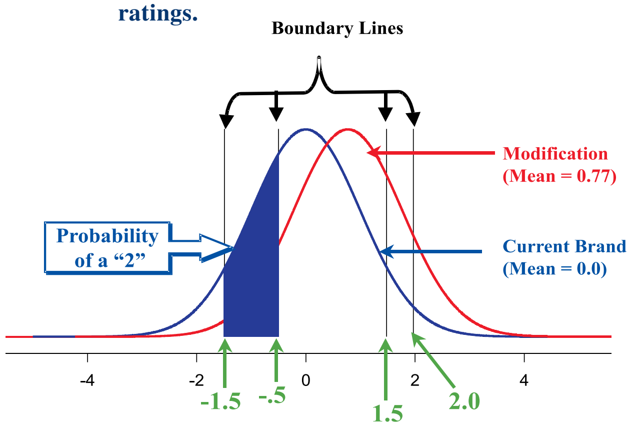


Figure 1b. Product distributions and decision boundaries on an interval scale modeled from 5-point word-category ratings.



Estimating Scale Means and Variances: Scale means do not depend on the location of decision boundaries, and scale means are not be affected by the method used to obtain them. In fact, the same scale means can be obtained from a 5-, 7- or 9-point scale; words, symbols, or numbers; or even ordinal measurement such as "choose the sweetest." The key to obtaining scaling information is to use a model that relates the rating results to the scale means and decision boundaries. Estimates of these parameters and their variances can be obtained using the method of maximum likelihood[1]. Using this procedure, we find values for the scale means and decision boundaries that best correspond to the results given in Tables 1a and 1b. It should be noted that this method can be applied to any number of products, not just to two as discussed here.

The means of the two distributions in Figures 1a and 1b show where the two products fall on the equal interval sweetness intensity scale. For the numerical ratings, the means are 0.0 and 0.81. For the word-category ratings, the means are 0.0 and 0.77. The units for each of these scales are the same: perceptual standard deviations. The fact that the intensity units for both scales are the same allows us to compare directly the results of the two experiments.

Relating Methodologies: Fundamental measures like scale means are not method specific. Consequently, these measures can be used to predict the results of other methodologies. Table 2 gives the results of modeling the data in Tables 1a and 1b.

Table 2. Scale means for the modified brand relative to the current brand in perceptual standard deviation units (δ values).

	7-Point Scale	5-Point Scale
Mean (δ)	0.81	0.77
Variance	0.026	0.029
Boundary Estimates	-1.8, -1.1, .1, 1.2, 2, 2.8	-1.5, .5, 1.5, 2

t-test on the equality of the δ values: not significant at $\alpha = 0.05$.

In this table, the results from the two rating experiments are comparable because they have been modeled using a measure that is independent of the rating method used. From this analysis you conclude that the 7-point scale and the 5-point scale produce similar estimates of the sweetness intensity difference between the current brand and the modification. Averaging these results, you find that the overall δ value is 0.79, and in a separate test you could determine whether this result is significantly larger than zero.

A valuable feature of Thurstonian models[2-4] is the ability to interlink the results of categorical methods, such as rating methods. There are many benefits to this aspect of Thurstonian modeling, one of which is the capability to make power comparisons among methods to determine optimum sample sizes. This point was made in a previous report on difference tests[5]. The ability to predict the results of other methodologies from scale means is quite general and applies to many types of categorical data.

Conclusion: Although rating methods may not produce interval scale data, a decision boundary Thurstonian model can extract interval scale information from ratings data. Hypothesis tests can be conducted and the results of different methods can be compared and combined.

References:

1. Dorfman, D.D. and Alf, E., Jr. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals - Rating method data. *Journal of Mathematical Psychology*, **6**, 487-496.

2. Thurstone, L.L. (1927a). A law of comparative judgement. *Psychological Review*, **34**, 273-286.

3. Thurstone, L.L. (1927b). Psychophysical analysis. *American Journal of Psychology*, **38**, 368-389.

4. Thurstone, L.L. (1927c). Three psychophysical laws. *Psychological Review*, **34**, 424-432.

5. Ennis, D.M. (1998). Thurstonian scaling for difference tests. *IFPress*, **1**(3), 2-3.