Comparing Perceptual Noise in Rating Scales

Daniel M. Ennis

Background: In the Southern District of New York in 2011, Church & Dwight Co. Inc. (C&D) sought a preliminary injunction barring The Clorox Company (Clorox) from continuing to air a commercial that implied that Clorox's cat litter product was superior to C&D's at eliminating cat waste malodor[1]. This preliminary injunction was granted in early 2012. Among a number of legal and technical issues in the case was the implausibility that eleven trained panelists would report a score of zero in each of four replications to a previously sealed jar containing a carbon and cat waste sample. The rating instrument used was a 15-point malodor scale with 0.1 unit increments, effectively making it a 151-point scale and previous research from Clorox had shown that human subject ratings of cat litter with or without cat waste are quite variable. C&D argued that such a result is "highly implausible" given that the test used humans, which are known to be "noisy instruments" when it comes to olfactory perception. The court explained that this is because humans "for neurological reasons, perceive the exact same thing differently at different times and report the presence of olfactory stimuli even when they do not exist."

Sources of Noise: In the C&D vs. Clorox case, the court correctly identified neural noise, one of at least three sources of noise that occur when chemosensory stimuli are evaluated. These sources include: a) stimulus noise, which arises from variation in the stimulus or its presentation; b) peri-receptor noise such as in the composition of saliva or the activity of enzymes that act on an odorant; and c) neural noise from within and among activated neurons. In addition to these three sources, the instructions to subjects and the presence of categories on a rating instrument themselves may be interpreted differently at different times leading to noise associated with the decision categories. These sources of noise are integrated to produce perceptual variation. Models that separately account for stimulus and neural noise have been developed[2]. If the data are from multiple subjects, such as when a panel is used, there may be additional variation due to individual differences in sensitivity or scale usage. This source of noise is typically confounded with other sources of perceptual variance, although the use of concurrent ratings, such as liking on the basis of an attribute combined with intensity on that attribute, may provide a basis for separating them[3].

Scenario: You are interested in evaluating the difference in perceptual variance between a 15-point ordered word-category-scale and a 7-point numerical-category scale on malodor intensity. Using the same stimuli and the same subjects, you hypothesize that the 15-point word-category scale will exhibit greater perceptual variance than a numerical-category scale due to differences in the interpretation of the words by different subjects or the same subjects at different times. Three hundred subjects evaluated two of your experimental products on the 7-point and 15-point rating scales, with appropriate balancing of the order in which the testing was conducted. Tables 1 and 2 show the results of the experiment using the 15-point word-category scale and a 7-point numerical-category scale, respectively.

Estimating Perceptual Variance: Models for converting categorical data from choice tasks such as 2-alternative forced choice, the method of tetrads, or ratings data are well developed[4,5,6]. In these models it is assumed that there is a perceptual difference between the items tested and a variance associated with each item as shown in Figure 1.

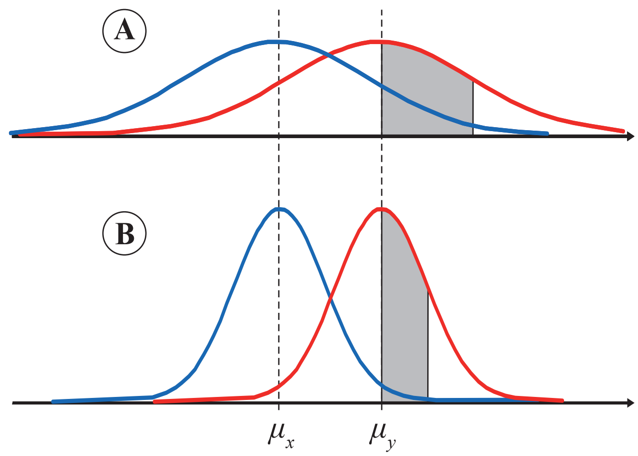

Figure 1. (A) Two perceptual distributions separated by one standard deviation, δ = 1. (B) Two distributions centered on the same means, but separated by two standard deviations, δ = 2. The only difference between (A) and (B) is the size of the perceptual standard deviation.

In order to create a unit of measurement, the perceptual standard deviation is used and the perceptual difference is expressed in terms of that standard deviation as shown in Figure 1. It is important to distinguish between the perceptual standard deviation and the standard deviation calculated from the ratings data or associated with the choice proportion in a binary task. The perceptual standard deviation is inferred from the data, it is latent and not directly observable. When we model or scale categorical data and determine d', we are estimating δ, the true standardized difference between the items tested, and δ depends on the perceptual variance to determine its unit of measurement. Hence δ is a form of signal to noise ratio. If we assume that the unstandardized sensory difference is constant for two comparisons involving the same pair of stimuli, then ratios of the two δ values may be assumed to reflect the inverse of the ratio of the perceptual standard deviations. In other words,

$$\frac{\delta_B}{\delta_A} = \frac{\sigma_A}{\sigma_B}\,.$$

This idea will now be used to evaluate the difference in perceptual variance between the 7-point numerical-category scale and the 15-point word-category scale.

Analysis of the Ratings Data: Table 3 shows the values of d' for the two methods along with their variances obtained from IFPrograms. Note that these variances are d' estimate variances, not perceptual variances, and would change with sample size while the perceptual variances would not. The d' value for the 7-point scale is 0.995 with an estimate variance of 0.008 and for the 15-point scale it is 0.708 with an estimate

	Rating Categories (15-point scale)														
Products	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Product 1	12	19	28	31	9	28	53	25	27	37	10	14	0	4	3
Product 2	2	5	10	15	5	18	43	25	32	57	21	37	1	14	15

Table 1. Results from a 15-point word-category scale for two products rated by 300 assessors. The words corresponding to the 15 points are not shown.

	Rating Categories (7-point scale)						
Products	1	2	3	4	5	6	7
Product 1	8	19	56	88	93	27	9
Product 2	1	2	14	45	109	72	57

Table 2. Results from a 7-point numerical-category scale for two products rated by 300 assessors.

Scale Type	d'	Estimate Variance
15-Point	0.708	0.007
7-Point	0.995	0.008

Table 3. A Thurstonian model of the data in Tables 1 and 2 to determine d' values and their estimate variances.

variance of 0.007. The ratio of these two d' values is 1.405. The d' values can be assumed to be normally distributed for such a large sample size. Can we show that 1.405 is significantly greater than 1.0 (if the ratios were equal) with 95% confidence and conclude that the variance associated with the 15-point scale is greater than the 7-point scale? Using theory on the lower 95% confidence bound of ratios of normally distributed data[7,8], you find that this value is 1.1 . You therefore conclude that the perceptual variance associated with the 15-point scale in this experiment is greater than that for the 7-point scale. In fact, a point estimate of the ratio of the perceptual standard deviations is slightly greater than 1.4 and so the ratio of the perceptual variances is 1.96. This means that the perceptual variance present when the 15-point word-category scale is used is nearly twice that when the 7-point numerical-category scale is used. The result in this scenario is by no means a general result, but illustrates a methodology for checking differences in perceptual variance so that the most sensitive scales to detect differences can be found for a particular application.

Conclusion: Although perceptual variance cannot be directly measured with behavior tasks and must be inferred from a Thurstonian model of the data, it can have a profound effect on the sensitivity of measuring instruments such as rating methods since the perceptual standard deviation is used as a unit of measurement. In this technical report a method for quantifying and statistically testing the relative size of the perceptual variances under two different measurement stratagems was evaluated. This methodology may be useful to design highly sensitive rating instruments for particular applications. Given a neural component of perceptual variance, it should be expected that even if stimulus and peri-receptor noise were controlled, the probability of repeated zero ratings on a 151-point scale should be extremely small.

References

1. Church & Dwight Co. vs. Clorox Co., S.D.N.Y., No. 1:11-cv-01865-JSR, 1/3/12
2. Ennis, D. M. and Mullen, K. (1992). Probabilistic psychophysics with noisy stimuli. *Mathematical Social Sciences,* **23**(2), 221-234.
3. Ashby, F. G. and Ennis, D. M. (2002). A Thurstone-Coombs model of concurrent ratings with sensory and liking dimensions. *Journal of Sensory Studies,* **17**, 43-59.
4. Ennis, J. M., Ennis, D. M., Yip, D., and O'Mahony, M. (1998). Thurstonian models for variants of the method of tetrads. *British Journal of Mathematical and Statistical Psychology,* **51**, 205-215.
5. Ennis, D. M., Mullen, K. and Fritjers, J. E. R. (1988). Variants of the method of triads: Unidimensional Thurstonian models. *British Journal of Mathematical and Statistical Psychology,* **41**, 25-36.
6. Kim, K., Ennis, D. M., and O'Mahony, M. (1998). A new approach to category scales of intensity II: Use of d' values. *Journal of Sensory Studies,* **13**, 251-267.
7. Ennis, J. M. and Ennis, D. M. (2011). Confidence bounds for multiplicative comparisons. *Communications in Statistics - Theory and Methods,* **40**, 3049-3054.
8. Ennis, D., Ennis, J., Palen, J., and Lampe, R. (2008). Confidence bounds for positive ratios of normal random variables. *Communications in Statistics - Theory and Methods,* **37**, 307-317.