Empirical formula for creating error bars for the method of paired comparison

Ethan D. Montag
Rochester Institute of Technology
Chester F. Carlson Center for Imaging Science
Munsell Color Science Laboratory
54 Lomb Memorial Drive
Rochester, New York 14623
E-mail: montag@cis.rit.edu

Abstract. *The method of paired comparison based on Thurstone's case V of his law of comparative judgments is often used as a psychophysical method to derive interval scales of perceptual qualities in imaging applications. However, methods for determining confidence intervals and critical distances for significant differences have been elusive, leading some to abandon the simple analysis provided by Thurstone's formulation. Monte Carlo simulations of paired comparison experiments were performed in order to derive an empirical formula for determining error. The results show that the variation in the distribution of experimental results can be well predicted as a function of stimulus number and the number of observations. Using these results, confidence intervals and critical values for comparisons can be made using traditional statistical methods. © 2006 SPIE and IS&T.* [DOI: 10.1117/1.2181547]

1 Introduction

There are many psychophysical techniques for creating scales of perception and sensation. Common techniques such as the classical psychophysical threshold methods or ratio scaling methods can be used to create scales of sensory magnitude.[1–3] However, for judgments of image quality or image preference, these techniques may not be applicable. For example, if one wants to determine which of several printers produces prints that are of highest image quality, there is no one continuous physical parameter that is being manipulated. In addition, the scale that is to be created is not a ratio scale because no concept of "zero quality" exists. What is desired is a scale that assigns numbers to the psychological percept that allows comparisons of the different stimuli. Such techniques for creating interval scales include different ranking, sorting, and comparison procedures and statistical analyses that transform the subjects' ratings into interval scales.

The method of paired comparison has become a popular tool for evaluating the effect of various algorithms or treatments on image quality or for quantifying the change in a perceptual characteristic (Engeldrum's "-nesses"[4]) such as perceived contrast, sharpness, graininess, etc. Specifically, reference is made to Thurstone's law of comparative judgment.[5–7] In its original formulation,[6] Thurstone presents a simple theory of the discriminal process and how its nature allows the construction of an interval scale based on comparisons of pairs of stimuli. This has been the starting point for much research in the application and analysis of paired comparison data. In its original formulation with its simplifying assumptions, the analysis of paired comparison data is computationally easy and straightforward.

This said, we have had some problems in implementing Thurstone's law because the ability to compute confidence intervals is missing in the formulation. Later research has used modifications of Thurstone's law to produce error metrics,[8] but this seems to be at a cost of losing the inherent simplicity of Thurstone's original formulation. In this paper we use Monte Carlo simulations to estimate an empirical formula for constructing error bars based on Thurstone's law.

2 Thurstone's Law of Comparative Judgment

Thurstone presents the law of comparative judgment[6] based on the following propositions:

1. Each stimulus gives rise to a discriminal process, which has some value on the psychological continuum of interest.
2. Due to momentary fluctuations (occurring within or between observers), the value of a stimulus may vary on repeated presentations. The distribution of this fluctuation can be characterized by a postulated normal distribution.
3. The mean and standard deviation of the distribution associated with a stimulus are its internal scale values and discriminal dispersion, respectively.
4. Therefore the distribution of the difference between two stimuli is also normally distributed and it is a function of the proportion that one stimulus is chosen as greater than the other.
5. The difference in scale values R between two stimuli i and j is:

$$R_i - R_j = z_{ij}\sqrt{\sigma_i^2 + \sigma_j^2 - 2r_{ij}\sigma_i\sigma_j} \tag{1}$$

where R_i and R_j represent the scale values of stimuli i and j, σ_i and σ_j are the standard deviations of the respective discriminal dispersions, r_{ij} is the correlation between the two discriminal processes, and z_{ij} is the normal deviate (the z-score) corresponding to the proportion of times stimulus j is judged is judged greater along the psychological continuum than stimulus i.

For Thurstone's case V, we assume that: (1) the evaluation of one stimulus along the continuum does not influence the evaluation of the other in the paired comparison ($r_{ij} = 0$) and (2) the dispersions are equal for all stimuli ($\sigma_i = \sigma_j$). These assumptions lead to this formulation of the law:

$$R_i - R_j = z_{ij}\sigma\sqrt{2}. \tag{2}$$

Thurstone[5] argued that these assumptions apply even in cases where Weber's and Fechner's law apply and outlined a procedure for testing the validity of the assumption of equal variance. Mosteller's χ^2 test of goodness of fit[9] can also be used to test these assumptions. For our purposes, in paired comparison along continua such as image quality or preference we accept these assumptions, as has been done

Paper 05150LR received Aug. 12, 2005; revised manuscript received Dec. 12, 2005; accepted for publication Jan. 30, 2006; published online Mar. 9, 2006.
1017-9909/2006/15(1)/010502/3/$22.00 © 2006 SPIE and IS&T.

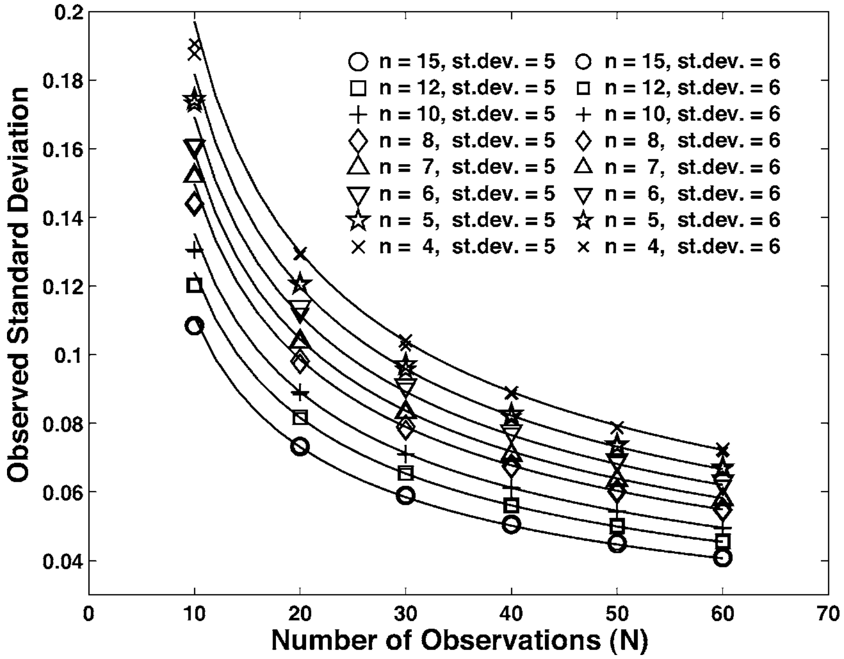

Fig. 1 Results of the simulations of paired comparison experiments showing the observed average standard deviation of the scale values as a function of the number of observations (N) and stimuli (n) with fit of Eq. (3).

in the literature,[2,10] and these assumptions are used as the basis of our simulations.

The analysis of data, according to this equation, is as follows. With n stimuli, $n(n-1)/2$ stimulus pairs are presented N times for judgment where N is the product of the number of presentations of each pair by each observer by the total number of observers. A frequency matrix (where the i'th column entry is chosen over the j'th row entry) of these judgments is constructed and then converted into proportions by diving by N. In turn the proportion are converted into normal deviates and the average of the columns create the interval scale values for the stimuli.

There seems to be no simple solution in the literature for determining the confidence intervals based on the case V solution. Bock and Jones[8] present an equation for calculating confidence intervals based on the arcsine transformation rather than the normal deviate. David[11] presents other methods for determining the statistical significance of scale value differences but these too rely on different analyses of the data. In the past, we have used Eq. (2) to calculate confidence. The 95% confidence intervals were $CI = R \pm 1.96(0.707/\sqrt{N})$ where the term $(0.707/\sqrt{N})$ reflects an estimate of the standard error, but we have never quite been satisfied with this because we surmised that the number of stimuli must also play a role in determining the error. Fortunately, the confidence intervals that result from this equation are quite conservative so that our previous experimental findings are not put into jeopardy.

3 Monte Carlo Simulation

To determine how scale values can vary and then empirically determine the standard deviation of these values for the construction of confidence intervals and tests of significance a Monte Carlo simulation of the paired comparison experiment was performed. By assuming an underlying psychological continuum that conforms to Thurstone's case V and randomly sampling from normal distributions along this continuum we can recreate the distribution of results that are expected over many repetitions of the experiment.

3.1 *Simulation*

For n stimuli, we chose n values as the means of normal distributions all with equal standard deviations. Samples were chosen from each of the $n(n-1)/2$ pairs of distributions and compared and the results were tallied. This was done N times for the N observations in the experiment. This would constitute one experiment in which the scale values were calculated as described above. This experiment was then repeated many times so that the means and standard deviations of the scale values could be calculated. The mean of the standard deviations of each of the scale values was then calculated to determine the overall standard deviation expected from an experiment with n stimuli and N observations.

So, for example, for $n=6$ stimuli, 6 normal distributions with means located at [5 6 7 8 9 10] and a standard deviation of 5 were used for sampling of the stimulus pairs. (A large standard deviation along the psychological continuum was chosen to reduce the number of results that contained unanimous judgments.) This was repeated for, say, $N=30$ observations. For each pair of n and N, the experiment was repeated 10,000 times and the means and standard deviations of the 6 scale values were calculated. The mean of the 6 standard deviations was taken as the overall observed standard deviation seen in an experiment.

The number of stimuli used in the experiment was $n = [4\ 5\ 6\ 7\ 8\ 10\ 12\ 15]$. The number of observation was $N = [10\ 20\ 30\ 40\ 50\ 60]$. Each experiment was repeated

10,000 times except when $n=15$, which was repeated 5,000 times because of limitations in computer memory. These simulations were repeated twice with the standard deviation of the underlying psychological distribution set at a value of 5 and 6. The locations of the means of the distributions along the psychological continuum and their standard deviations did not have an effect on the results, as would be expected from the theory.

Figure 1 shows the results of the simulation. The observed standard deviation is a function of both the number of stimuli n and the number of observations N.

3.2 *Estimates of the Observed Standard Deviation*

Abandoning an analytic approach, a number of equations were fit to the data to find an empirical equation that could capture the observed standard deviation as a function of the number of stimuli and observations. The following equation gave a good fit:

$$\sigma_{obs} = b_1(n - b_2)^{b_3}(N - b_3)^{b_5} \tag{3}$$

with $b_1=1.76$, $b_2=-3.08$, $b_3=-0.613$, $b_4=2.55$, and $b_5=-0.491$. The solid lines in Fig. 1 show the fit of this equation to the data. The rms of this fit is 87%. The fit is very good, which is not remarkable given that there are 5 parameters. It does seem to overestimate the standard deviation at $N=10$. This is not a bad feature considering that with such a small number of observers there is much less power in the experiment and the chance of unanimous judgments occurring by chance is increased. Repetition of the simulation with various values of n, N, and mean values on the psychological continuum verify that this equation gives an excellent approximation of the observed variability.

3.3 *Implementation*

To calculate 95% confidence intervals (CI) around the interval scale values determined in any particular experiment, one can use the expression $CI=\pm 1.96\sigma_{obs}$. In designs of paired comparison experiments that reduce experimental labor[7] the expression can also be used to adjust the size of the CIs for the various experimental partitions. Besides its use for determining confidence intervals, the value of σ_{obs} can be useful in determining more stringent critical values and confidence intervals for multiple comparisons of scale values. It is recognized that multiple comparisons of pairs of means in the analysis of variance leads to an increase in errors. It must also be true for the comparison of multiple scale values in paired comparison experiments. Keppel[12] enumerates a number of planned and post hoc tests such as the Scheffé, Dunnett, and Tukey tests for controlling error rate for multiple comparisons. Using an empirical value of the standard deviation in the place of the standard error in these tests may prove useful.

3.4 *Goodness of Fit*

In the course of doing the simulations, Mosteller's χ^2 test of goodness of fit[9] was performed on each of the individual experiments. This test compares the arcsine transform of the observed proportions to those predicted by the resulting scale values. The proportion of the experiments that were rejected based on this test for the different combinations of n and N were calculated. Surprisingly, these proportions were never less than 0.05 even though the underlying distributions from which the observations were drawn were normally distributed, conforming to case V assumptions, and unidimensional. This would lend one to think that this test is conservative, although Engledrum[2] cites Mosteller's report[9] that the model often appears better than it really is.

4 Conclusions

In this paper Monte Carlo simulations were used to estimate an empirical formula for the standard deviation of scale values and the constructing confidence intervals based on Thurston's law. It is recognized that a thorough understanding of the underpinnings of statistical distributions, experimental noise, and mathematical precision would lead to a much more satisfactory answer, but for now we are more interested in creating a tool for the analysis and determination of the significance of our data.

References

1. C. J. Bartleson and F. Grum, Eds., *Optical Radiation Measurements, Vol. 5, Visual Measurements*, Academic Press, New York (1984).
2. P. G. Engeldrum, *Psychometric Scaling*, Imcotek Press, Winchester, MA (2000).
3. G. A. Gescheider, *Psychophysics: The Fundamentals*, Lawrence Erlbaum Assoc., Publishers, Mahwah, NJ (1997).
4. P. G. Engeldrum, "A framework for image quality models," *J. Imaging Sci. Technol.* **39**(4), 312–318 (1995).
5. L. L. Thurstone, "Psychophysical analysis," *Am. J. Psychol.* **38**, 368–389 (1927).
6. L. L. Thurstone, "A law of comparative judgment," *Psychol. Rev.* **34**, 273–286 (1927).
7. W. S. Torgeson, *Theory and Methods of Scaling*, John Wiley & Sons, New York (1958).
8. R. D. Bock and L. V. Jones, *The Measurement and Prediction of Judgment and Choice*, Holden-Day, San Francisco (1968).
9. F. Mosteller, "Remarks on the method of paired comparison: III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed," *Psychometrika* **16**(2), 207–218 (1951).
10. C. J. Bartleson, "Measuring differences," in *Optical Radiation Measurements, Vol. 5, Visual Measurements*, C. J. Bartleson and F. Grum, Eds., pp. 441–489, Academic Press, New York (1985).
11. H. A. David, *The Method of Paired Comparison*, Hafner Press, New York (1969).
12. G. Keppel, *Design and Analysis: A Researcher's Handbook*, Prentice-Hall, Englewood Cliffs, NJ (1982).