On the Repeatability of Paired Comparison Based Scaling Methods

Chengwu Cui
Lexmark International, Inc.
Lexington, Kentucky

Abstract

Paired comparison based scaling method such as the method of paired comparison and the method of ranking, etc, are often used for subjective image quality measurement. The scaled values obtained are often used to predict customer preferences and are used for modeling purposes. Knowing the repeatability of these measurements is therefore important. This paper shows that the theoretical repeatability of paired comparison based subjective measurement can be simulated numerically. Also, in this paper, paired comparison based image quality measurement or scaling data are used to examine the potential repeatability of these measurements by sub-sampling the measured observer pool. In practice, it is strongly desirable to have as small a number of observers as possible. This type of analysis can help to reveal the potential risk when a small number of observers are used.

Introduction

Psychophysical scaling is often referred to as "mind measuring". However, the use of the word "measuring" has been criticized by physicists since the beginning of the subject, because data obtained by psychophysical scaling generally do not possess properties of their physical counterparts. Nonetheless, psychophysical scaling has been used successfully to provide useful information for the understanding of senses and the mind. In modern industrial production, psychophysical scaling has been constantly resorted for evaluating the quality aspect of a visual product, such as a reproduced digital image. [1]

Paired comparison based scaling methods have been widely used in image quality evaluation. These methods are based on the L. L. Thurstone's law of comparative judgment.[2,3] (To simplify the problem and follow the general trend in practice, we assume Thurstone's case V conditions in this paper unless specified otherwise.) Thurstone advanced psychophysics after Fechner's days. Fechner founded psychophysics but was obsessed in finding the direct one to one relationship between response and stimulus.[4] Thurstone hypothesized each stimulus produces subjective response value describable by random variables normally distributed on the sensory continuum. Given paired wise proportions of choice data, scaled values can be derived based on an appropriate level of Thurstone's modeling. The underpinning of the law of comparative judgment is the random distribution of the subjective response to the stimulus. Derived scaled values are therefore of merely statistical meanings. According to Bock and Jones, Thurstone's research involved data of large sampling sizes, therefore "there is little evidence of interest in the statistical bases of the methods". [5] When moderate sampling size is involved, which is much likely in image quality evaluation, statistical stability of the results can be problematic.

In imaging quality evaluation, the derived scaled values are often used in much the same way as their physical counterparts. Scale precision and measurement repeatability become important. Thurstone did not provide a method to evaluate the uncertainty of derived scaled values, he merely proposed the use of the average error of prediction on proportion of choice values. When the average error of prediction is small, the modeling would be regarded as a success. Because measurement data were transformed into z score and the final scaled values were optimized regression to all proportion of choice data (through z score) for the specific model, estimation on scaled value error become complicated.

In physical quantity measurement, it is important to know the measurement capability of a measurement method. Such capability is often represented by measurement precision and accuracy. Because the scaled values interested here are only of relative meaning (on the so-called ratio scale), accuracy is irrelevant here. On the other hand, the precision of the scaled values is important. Knowledge on the variance of the scaled values will reveal the validity of the data. In practice, variance is often related to a more fundamental measurement parameter, the measurement repeatability. For physical measurements, measurement repeatability can be obtained simply by repeating the measurement. In psychophysical scaling measurement, to repeat a certain scaling test is often remarkably prohibiting due to cost and time. Therefore, repeatability information is often given by estimating the 95% confidence intervals based on the estimated standard error of the scaled values.

Given n as the number of stimuli and N as the number of observers, Braun, Fairchild, and Alessi used

$$\sigma = \sqrt{\frac{1}{2(n-1)N}} \tag{1}$$

as the estimate of standard error of the scaled values.[6] Morovic used

$$\sigma = \sqrt{\frac{1}{2N}} \tag{2}$$

as the standard error estimate of his scaled values.[7] Because the variances are related to both the number of stimuli and the number of observers, Morovic's method therefore may overestimate the variance. According to Braun et. al.'s method, the number of stimuli will have the same effect on variance as the number of observers, which is unlikely given the way paired comparison data are obtained and modeled. Bock proposed the use a three-component scaled value model (true scaled value, individual difference, experimental noise) to analyze the variance in Thurstone's case V, and he suggested a formula to estimate the standard error of the scaled values,

$$\sigma = \sqrt{\frac{2[1+\rho(n-2)]}{nN}} \tag{3}$$

based on statistical inference and some statistical assumptions.[8] His goal was to explain the often too small χ^2 values produced by Mosteller's method.[9] The coefficient ρ is used to account for correlation between choices made on two comparisons of a common stimulus or the replication effect of a specific stimulus. The value of ρ was proven to be between 0 and 1/3. In the simplest case, ρ can be set to 0, corresponding to the case of no replication effect, the standard error is given by

$$\sigma = \sqrt{\frac{2}{nN}} \tag{4}$$

When ρ takes the largest value of 1/3, the standard error is given by

$$\sigma = \sqrt{\frac{2}{3}\frac{n+1}{n}\frac{1}{N}} \tag{5}$$

In the case of no replication effect, Bock's formula also implies that the number of stimuli and the number of observations have the same effect on variance. On the other hand, if the replication effect cannot be discounted, the number of stimuli effect reduces to $\frac{n+1}{n}$. When n becomes large, the number of stimuli will have little effect on variance. Bock's interesting evaluation deserves further verification.

There are alternative models to treat paired comparison data. The most well known model is the Bradley and Terry model.[10] Lately, Zhou and Cui proposed a maximum likelihood format of Thurstone's model.[11] Both these models are solvable through generalized linear methods that are often readily available in statistical and math software packages. These numerical methods normally provide variance information on the scaled values. Scaled value variance obtained in such a fashion reflects the tolerance of the fit to the proportion of choice data. Whether variance obtained in this way is equivalent to the true scaled value variance propagated from variance on measured proportions of choice is not clear. Further, for small numbers of observers as often desirable and used in industrial image quality evaluation, statistical inference can become unreliable. Because all models that were used for paired comparison based scaling essentially deal with the same information, the relative variance should all reflect the uncertainty of the scaled values that are derived from the original paired comparison data.

In this paper, we attempt to investigate the variance of the scaled values based on the Thurstone model by computer simulation. In reality, procedural and test condition errors can also give problems that can even invalidate some fundamental modeling assumptions. We further use a set of actual measurement data to examine measurement repeatability in hope to gain more insight into the capability of the paired comparison based scaling methods for practical image quality assessment.

Standard Error Estimation By Simulation

Given two physically identical images, when they are compared against each other in a paired comparison test, the chances of which one being chosen over the other shall be 50%. That is true only on the condition of a relative large number of observers. When the number of observers is small, say if only ten observers participated, if two observers chose one over the other while eight observers chose the opposite, the outcome can be perfectly normal, similar to the case when we toss a coin ten times. The outcome of this process follows the binomial distribution. For a set of assumed scaled values, each representing a stimulus or a sample image, when pair compared, the probability of one being chosen over the other for a specific pair combination can be computed. If all the probability values of all possible pairs are used to compute scaled values, we shall get back to the same scaled values because the process should be mathematically invertible. The key step of the simulation is to introduce random experimental proportion of choice errors. Given a probability value, the sampling size, and the number of trials, potential outcomes can be simulated with a random number generator. With these potential "noise added" probability or proportion of choice value, we can compute the scaled values. These scaled values will be slightly different from the originally assumed scaled values. Repetition of the simulation process for sufficient times should provide the variances of the scaled values. The flowchart for the simulation is given in Fig. 1. For further details, see Reference 12.

Because conversion of the proportion of choice to z score (normal deviate) becomes numerically unstable when it approaches 0% or 100%, simulated proportion of choice values close to 0% or 100% produce large z score fluctuation, which causes large scaled value errors. Therefore, the simulation is sensitive to the scaled value

range assumed. Using the case of 5 stimuli and 50 observers as example, Fig. 2 shows the dependency of standard error by simulation on the range of scaled values.

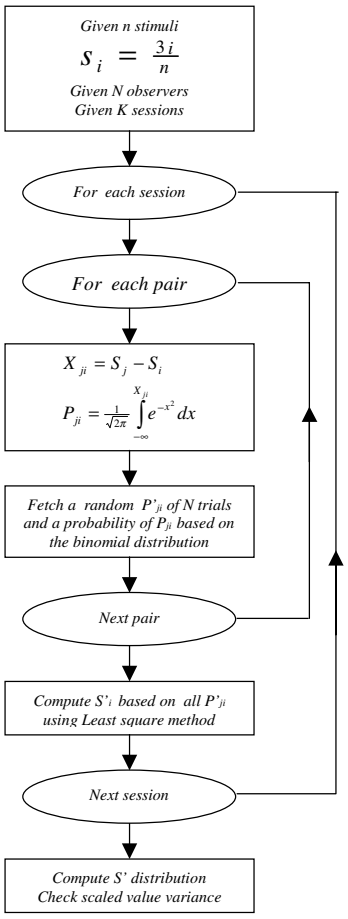

Figure 1. Simulation flowchart.

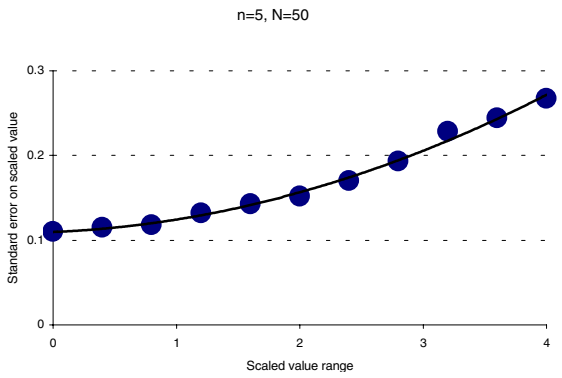

Figure 2. Dependency of the standard error on range of the scaled values assumed for the case of n=5 *and* N=50.

If a maximum scaled value range of 3 is used, it corresponds to a highest proportion of choice value of 99.9%. The simulated variances can be considered as assuming the worst-case scenario. For example, for three reproduced images, one of the images is clearly separable from the other two on a statistical level as high as 99.9%. The simulated standard errors for such a stimulus range can be fitted by,

$$\sigma(n,N) = \frac{1.85}{N^{0.42}(n-1)^{0.5}} \frac{n+1}{n} \qquad (6)$$

again, n is the number of stimuli and N is the number of observers. For a more realistic scaled value range, the simulation was also done for a maximum scaled value of 2 when the number of stimuli is equal or higher than 5; when the stimuli number is less than 5, the maximum scaled value was determined by 0.5 multiplied by the number of stimuli. A scaled value difference of 2 corresponds to a highest proportion of choice value of 97.7%. The simulated standard errors for such a scaled value range is shown in Fig. 3.

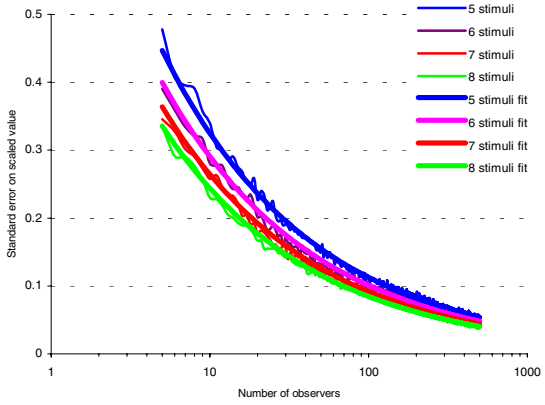

Figure 3. Simulated standard errors for various combinations of number of stimuli and number of observers, and the corresponding fits.

Again, the standard errors can be fitted by

$$\sigma(n,N) = \frac{2.5}{N^{0.46} n^{0.61}}, \qquad (7)$$

which is also shown in Fig. 3.

The result can be compared with those published. Fig. 4 shows the standard error as a function of N when $n=5$. Fig. 5 shows the standard error as a function of n when $N=30$.

Examination of Measurement Data

The above simulation assumes only random sampling errors. Still, it reveals that when the number of observers and the number of stimuli are small, scaled values based on paired comparison modeling can have large errors. In practice, there is often an array of other potential test error

sources, typical when surveying human subjects, will further worsen the scale precision. It is difficult, if not impossible, to narrow down and to quantify all these errors. Because the law of comparative judgment assumes "constant" inherent discriminal dispersions, the scaled values will be relative to the discriminal dispersions. In reality, the discriminal process may not have a stable distribution due to some of these unaccounted errors. For example, the outcome of the comparison of a pair of color image samples may change due to emotional changes within a single individual; we therefore may not have a stable discriminal dispersion distribution. Consequently, any modeling will likely to fail. Errors of such nature cannot be simulated easily. Here we will examine a real set of paired comparison data in hope to gain some insight of the derived scale precision or the repeatability of such methods when used for image quality assessment.

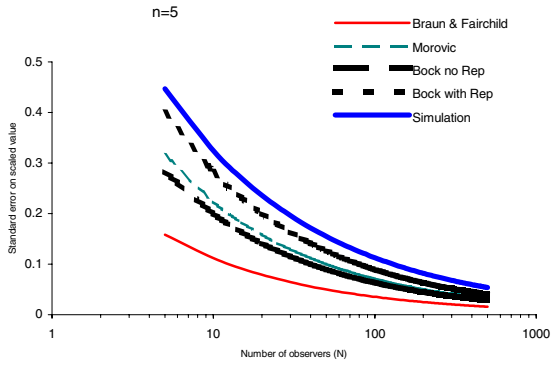

Figure 4. Comparison of estimated standard errors as a function of the number of observers for a set of five stimuli.

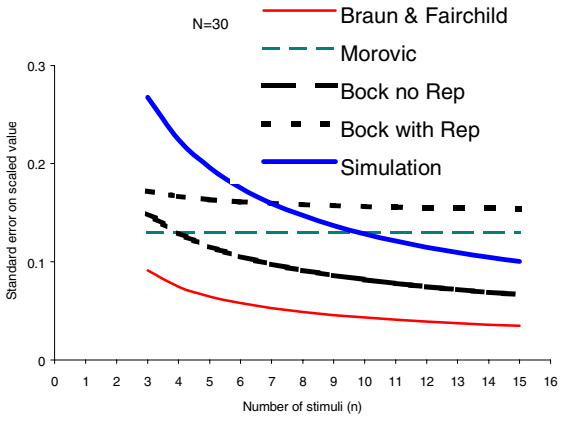

Figure 5. Comparison of estimated standard errors on scaled values as a function of number of stimuli with thirty observers.

The data used here are test results on printed image color quality. Three images ("people", "places", and "things") were printed using five different color mapping algorithms. Seventy-eight observers participated in the paired comparison test. A comparison table was obtained on each observer for each image. In total, 78 tables were obtained for each image. Details on the test can be found in Reference 13. Scaled values can be computed based on these data following a specific paired comparison model, variance on the scaled values due to random sampling error can be simulated following the simulation method given above.[12] Alternatively, the standard error can be estimated by Equation 7.

Because the test goal was to evaluate the color algorithm preference, it can be expected that there might be age, gender, and even cultural differences in color reproduction preference. We will assume that the data are representative of a certain population and sub-sample the 78 observers to examine the scaled values computed based on the smaller samples.

The paired comparison data for each image was stored in 78 copies of paired comparison tally table, one for each observer. The software used to calculate the scaled values could be fed as many copies of the tally table as desired. For a specific sub-sampling size, e.g. a size of 10, the software was modified to randomly take 10 tally tables out of 78 to re-compute the scaled values, simulating that the test was done with the 10 of the 78 observers. Such random sampling was repeated for 300 times and a distribution of the scaled values were obtained. The scaled value distribution for the sub-sampling size of 10, 20, and 30 for the "people" image is shown in Fig. 6, 7, and 8, respectively.

The sub-sampling was done for the other two images. Table 1 shows the average standard deviation for each sub-sampling size and image combination.

Table 1. Average standard deviations of the scaled values calculated based on sub-sampling the 78 observers.

Sub-sample Size	People	Places	Things	Eq. 7
10	0.35	0.33	0.37	0.32
20	0.26	0.22	0.25	0.24
30	0.19	0.17	0.19	0.20

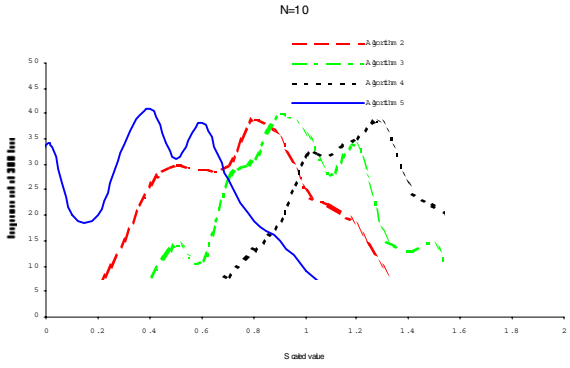

Figure 6. Distributions of scaled values calculated based on a sub-sampling size of 10 out of the 78 observers for the "people" image.

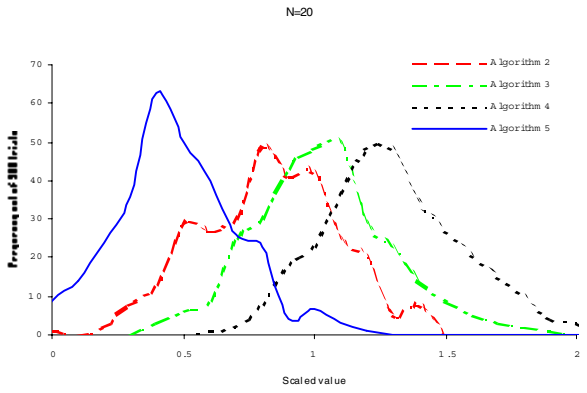

Figure 7. Distributions of scaled values calculated based on a sub-sampling size of 20 out of the 78 observers for the "people" image.

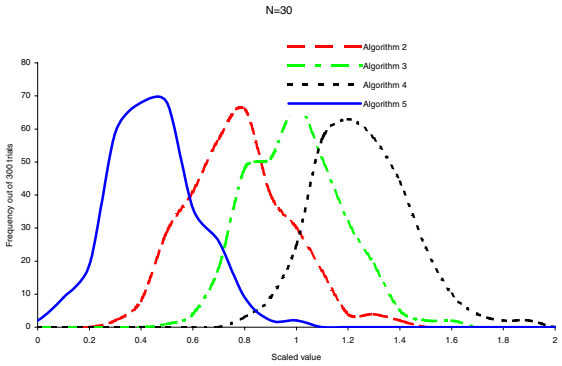

Figure 8. Distributions of scaled values calculated based on a sub-sampling size of 30 out of the 78 observers for the "people" image.

Discussion

The computer simulation provides an effective method for estimating the measurement errors of paired comparison based scaling methods. Fig. 4 and 5 prove that some often-used methods for estimating scaled value variance sometimes tend to under-estimate the errors. The errors of scaled values depend on the maximum scaled value difference because the large z fluctuation at the two extreme proportion of choices regions. For a reasonable maximum scaled value difference of 2, Eq. 7 can provide a good representation of the standard error for a certain number of observer and a certain number of stimuli.

Figure 6 shows the dramatic variation on the scaled values caused by using different groups of observers 10 at a time with real test data. The variation magnitude shown are consistent with that predicted by the simulation although the variation distribution can hardly be regarded as "normal". As can be expected, the variations become smaller with lager sub-sampling sizes as shown in Fig. 7 and 8. The standard errors calculated with Eq. 7 are also included in Table 1. Because the sub-sampling was done to 78 observers, the degree of freedom drops when the sub-sampling size approaches 78. When the sub-sampling size is 78, there will be no variance for the sub-sampling. That is to say, the sub-sampling exhausts the "entire" population (all 78 observers). Therefore, the standard errors obtained by sub-sampling shown as shown in Table 1 under-estimate the true scaled value errors if we consider the population of interest. The true scaled value errors should also include the sampling error originated by using only 78 observers to represent the interested population. Nonetheless, the example here shows, with real measurement data, the large potential risk of using a small number of observers.

Table 1 also shows that, at least for this set of data, image content did not seem to show large effect on errors although it might be that these color correction algorithms were already balanced and optimized for all three types of images.

Summary

Variances or errors on scales derived by paired comparison based scaling methods were numerically simulated. The simulation results reveal the potential large errors when a small number of observers and a small number of stimuli are used. An error estimation equation was given based on the simulation results for a reasonable scaled value range. The equation can be used to help to determine the number of observers for designing paired comparison based scaling experiment. The results here also show that some methods used might have underestimated scale values erros. Examination with real image quality paired comparison data gave consistent results with the simulation and further shows the potential large errors when a small number of observers are used.

References

1. Engeldrum, Peter G, Psychometric scaling, A toolkit for imaging system development, Imcotek press, 2000.
2. Thurstone, L. L, A law of comparative judgment, Psychophysical Review, 34, 273, 1927.
3. Torgerson, W. S, Theory and methods of scaling, John Wiley & Sons, Inc. 1958.
4. Boynton, R. M, Psychophysics, Chapter 6 in Optical Radiation Measurement, Vol. 5, Visual measurements., C. J. Bartlson and F. Grum Eds. Academic Press, Orlando, FL 32887, 1984.
5. Bock, R. D. and Jones, L. V, The Measurement and Prediction of Judgment and Choice, Holden-Day Inc., San Francisco, CA, 1968.
6. Braun, K. M, Fairchild, M. D, Alessi P. J, Viewing environments for cross media image comparison, Color Research and Application, Vol. 21, p 6-17, 1996.
7. Morovic, J, To Develop a Universal Gamut Mapping Algorithm, Ph. D. Thesis, University of Derby, 1998.

8. Bock, R. D, Remarks on the test of significance for the method of paired comparisons, Psychometrika, Vol. 23, P. 323, 1958.
9. Mosteller, F. Remarks on the method of paired comparisons: III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed, Psychometrika, Vol . 16, P202, 1951.
10. Bradley, R. A, and Terry, M. E, Rank analysis of incomplete block design. I. The method of paired comparisons, Biometrika 39, 324-345, 1952.
11. Zhou, M. and Cui, C, New mathematical model for the law of comparative judgment, IS&T NIP16, Vancouver, B.C. Canada, P383-387, 2000.
12. Cui, C, Error simulation of paired comparison based scaling methods, Color Imaging: Device-Independence Color., Color Hardcopy, and Graphic Arts VI, Proceedings of SPIE, R. Eschbach and G. G. Marcu Eds. , San Jose, January, 2000. P192-198.
13. Cui, C, Comparison of two psychophysical methods for image quality measurement: paired comparison and rank order, Proceedings of 8th Color Imaging Conference, Scottsdale, Arizona, P222-227, 2000.

Biography

Chengwu Cui received his BS degree in optics from Shandong University, MS in color science from Chinese Academy of Science and PhD in vision science from the University of Waterloo. From 1995 to 1999, he worked for GretagMacbeth as a color scientist. He is currently with Lexmark International. His research interests include human vision, ocular optics, image quality, color measurement, daylight simulation, computer color formulation and psychophysics.