

Comparison of Two Psychophysical Methods for Image Color Quality Measurement: Paired Comparison and Rank Order

Chengwu Cui
Lexmark International Inc., Lexington, Kentucky

Abstract

This paper compares two popular psychophysical scaling methods, the method of rank order and the method of paired comparison, for measuring reproduced image color quality. Although there are reports supporting that the two methods produced virtually the same results when applied to some psychophysical scaling tests, given the complexity of image color quality perception, it is of interest to investigate the validity of the relationship between the two measurement methods. In this paper, an experiment was designed using three groups of color images to examine the potential relationship between the two methods. The results prove that the two methods produce similar but different results. Methods for deriving scaled values based on rank order data are also investigated. It is concluded that simpler data processing methods can give acceptable results. Some of the advantages, disadvantages, and error sources are also discussed.

Introduction

The method of rank order and the method of paired comparison are both popular psychophysical measurement methods for scaling a set of stimuli or samples on a given perceptual attribute.[1,2] Various forms of both methods are frequently resorted to in our daily decision making processes. When the two methods are used in psychophysical scaling measurements, the goal is to derive the perceptual quantities or the scaled values on a continuum based on certain modeling principles such as Thurstone's model, also known as the law of comparative judgment.[3] Thurstone's model places the perceptual quantities on a continuum by converting measured judgmental frequencies to z-scores based on the normal distribution. In the paired comparison process, each stimulus serves as the standard once against every other stimulus and the presentation of sample combinations can be randomized. Therefore, the paired comparison method is suitable for unbiased measurement. In principle, the method of rank order provides the same information as the method of paired comparison. The observer needs to compare each stimulus with every other stimulus in order to determine that stimulus's rank. Data from rank order measurement can be converted into paired comparison data to derive the scaled values based on paired comparison models. The method of rank order is straightforward, easy to administer and less time-consuming. Data of rank order tests can also be directly used to differentiate a set of stimuli. For example, the interest of the test can be only to identify the most preferred stimulus on the given attribute.

During a rank order test, the observer observes and arranges all the samples at the same time. The observer may form an opinion on Stimulus 1 when he or she compares Stimulus 1 to Stimulus 2. When the observer compares Stimulus 1 to Stimulus 3, he or she may carry that opinion on Stimulus 1 into its comparison to Stimulus 3. It is also difficult to verify that the observer actually compared each sample with every other sample to form the rank order. Further, the observer usually has to arrange all the stimuli at the same time; illumination non-uniformity can be a potential problem sometimes.

Practical situations require the use of one method over the other. For example, comparison of a CRT display image to an original image may be limited to displaying one image at a time, consequently the method of paired comparison can be the only option. If a large number of stimuli need to be measured in a short period, paired comparison may be impractical. Other factors being equal, it is of interest to investigate which method should be used for a specific application. Bartleson used a set of color sample chips to show that many scaling methods including the rank order and paired comparison produced very similar results. [1] However, his samples were distinctive uniform color samples. Hevner designed a specific experiment to test the relationship between the two methods. In the experiment, 370 subjects were employed to scale the degree of excellence of 20 handwriting specimens using the two methods (a third method was also compared, but is irrelevant here).[4] She concluded that the two methods gave virtually the same results. During the experiment, the subjects were asked to base their judgment on neatness, uniformity of the slant, and uniformity of the stems and ovals of the letter, which are all geometrical properties of handwriting. Perceived image quality is a combination of the color quality of all major colors used for the objects in the image. Color sensation is a rather complex process and affected by the presence of other colors in the visual field or scene. The observer's criteria regarding color quality can also vary depending on the specific image scene. For

example, adequately exaggerating chroma or saturation during the reproduction process is often preferred for colors of vegetables and fruits but disfavored for skin tones. These factors may all contribute to the psychophysical measurement validity. Therefore, it is the interest of this paper to investigate the difference between the method of paired comparison and the method of rank order in psychophysical measurement of image color quality. This paper describes an experiment designed to compare the two methods for measuring printed image color quality. During the experiment, the color quality of three types of images was scaled using both methods under controlled lab conditions. Simpler methods for computing scaled values based on rank order data are also discussed.

Deriving Scaled Values from Rank Order Data

Assuming each stimulus is compared to every other stimulus during the rank order process with the same criteria used in the paired comparison process, the two methods should produce the same measurement results. Rank order data can be converted to paired comparison data to compute scaled values based on the comparative judgment modeling. Converting the rank order data to paired comparison data and computing scaled values based on comparative judgment models can be time consuming without the assistance of computer, therefore, rank order data reduction methods are often used. [1] The simplest method is to compute a rank score for each stimulus. If stimulus i is ranked at j^{th} rank for k_j times, its rank score can be computed by,

$$RankScore_j = \frac{\sum_{j=1}^{n}(n-j)k_j}{N(n-1)}, \qquad (1)$$

where n-j is a weighting factor (the lowest in the rank has a weighting factor of 0).

Rank scores can be used to represent scaled values. They can also be further converted to z score values.[1] However, scaled values produced by this method is different from that produced by applying the law of comparative judgment. To understand the inherent difference, we need to examine the computational flows of the two methods. For convenience, we can assume the paired comparison data set (paired comparison matrix) is complete. For stimulus i, its scaled value s_i is given by,

$$s_i = \frac{1}{n-1}\sum_{j=1}^{n} z\left(\frac{\sum_{k=1}^{N} p_{jki}}{N}\right) \qquad (2)$$

where z() represents the operator to convert the proportion of choice to z score; p_{kij} is a binary number (a value of 1 represents that stimulus i is greater than stimulus j on the given attribute, and 0 for vice versa, from the k^{th} observation).

Converting rank order data into paired comparison matrix, we have

$$RankScore_i = \frac{1}{(n-1)N}\sum_{k=1}^{N}\sum_{j=1}^{n} p_{kij} = \frac{1}{n-1}\sum_{j=1}^{n}\left[\frac{\sum_{k=1}^{N} p_{kij}}{N}\right] \qquad (3)$$

Applying z() operator to Eq. 6, we have,

$$s_i = z\left(\frac{1}{n-1}\sum_{j=1}^{n}\left[\frac{\sum_{k=1}^{N} p_{kij}}{N}\right]\right) \qquad (4)$$

Eq. 4 is equivalent to computing the average proportion of choice of comparison for one stimulus with all other stimuli and then computing the z score. Eq. 2 computes the scaled value by computing the z score from the proportion of choice with every other stimulus and then computing the average z score.

The differences produced between the above different computations can be demonstrated by simulation. Assuming there are five stimuli with hypothetical scaled values of 0, 0.5, 1.0, 1.5 and 2.0, respectively, we can compute the expected proportion or choice when compared with each other. Models based on the standard normal distribution such as Thurstone's law of comparative judgment deal with scaled values in a small numerical range (z score ranges from –3 to 3 for reasonable proportion of choice values). Therefore, the choice of the above theoretical values covers a reasonable range of practical values. Table 1 shows the theoretical proportions of choice for the five stimuli. For example, the value of 0.69 at Column 3 and Row 2 of Table 1 represents the proportion of choice of Stimulus 2 over 1 on a given attribute.

Table 1. Theoretical proportion of choice for each paired combination of the five stimuli.

	1 (0)	2 (0.5)	3 (1.0)	4 (1.5)	5 (2.0)
1 (0)	-	0.69	0.84	0.93	0.98
2 (0.5)	0.31	-	0.69	0.84	0.93
3 (1.0)	0.16	0.31	-	0.69	0.84
4 (1.5)	0.07	0.16	0.31	-	0.69
5 (2.0)	0.02	0.07	0.16	0.31	-

Table 1 can be used to demonstrate the difference between above computation equations. Table 2 gives the computed scaled values, the normalized rank score and the scaled value computed by the data reduction method given by Eq. 4. The normalization of the rank scores was done by adjusting the rank score so that Stimulus 5 has a value of 2.0. The normalization factor is 2.77. Fig. 1 shows

normalized rank scores and the scaled values computed by Eq. 4 versus the hypothetical values, respectively.

Table 2. Comparison of computed scaled values by three different methods.

Stimulus ID	1	2	3	4	5
Normalized Rank score	0	0.46	1.00	1.54	2.0
Scaled values by Eq. 4	0.00	0.58	1.08	1.59	2.17
Scaled values by Eq. 2	0	0.5	1.0	1.5	2.0

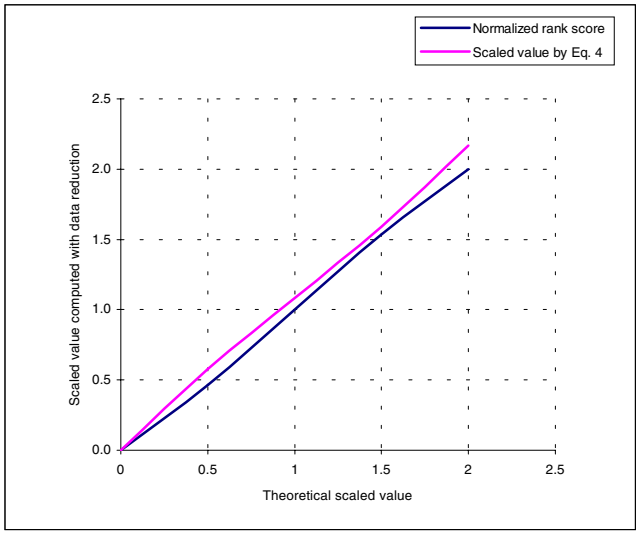

Figure 1. Relationship between the theoretical scaled value and the scaled value computed by Eq. 7 (solid) and the normalized rank score (broken line), respectively.

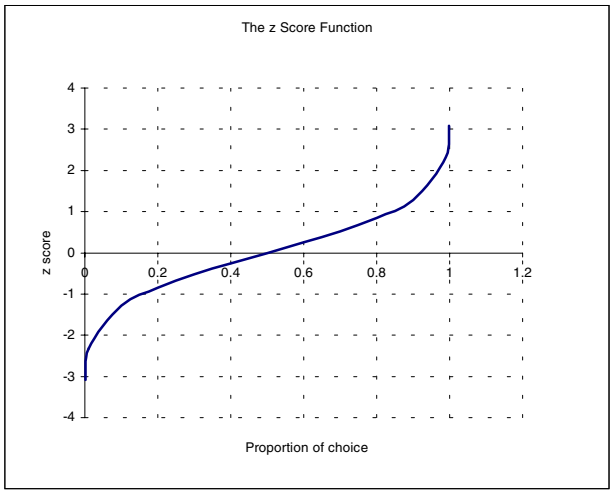

Figure 2. The z score function.

Fig. 1 shows that both the two curves are close to the 45°-line, indicating the three methods of computation produced very close results. To explain the relationship, we only need to examine the z or inverse accumulative normal distribution function as shown in Fig. 2. The rank scores used ranges from 0.14 to 0.86, which covers a linear segment of the z score function. This result implies that the rank score value can be used to substitute the scaled values for a stimuli set of close scaled values.

Experiment

Stimuli

As aforementioned, observers may look for different quality attributes and apply different criteria in regard to color quality of an image, depending on the type of image. Three test images representative of three different types of pictorial images were used in this study. Each of the three images represented a specific type of pictorial image. They were categorized according to their corresponding content as "People", "Places" and "Things", respectively. For each image, five copies were printed using a Lexmark color laser printer (model C710™) with a slightly different color correction algorithm, respectively. These different algorithms were intended to fulfill various color reproduction goals to various extents. Therefore, five copies of a different color reproduction quality were printed for each test image. The prints were pasted on a piece of white cardboard of the same size for handling convenience.

Viewing Conditions

The test was conducted in a windowless room with walls painted to a neutral gray (N7). The illuminator was a GretagMacbeth™ overhead fluorescent daylight D50 luminaire with a built-in diffuser. On the sample plane, the illumination was about 900lux. The variation of illumination intensity on sample viewing plane is less than 10%.

Observers

Sixty-one observers volunteered and participated in the test. They were colleagues with various technical backgrounds at the Lexington, Kentucky location of Lexmark International, Inc. All subjects passed the Ishihara 100 Hue test for abnormal color vision screening administered prior to the test.

Procedure

A computer program was written to administer the paired comparison test. The program randomized the presentation order of the pair combinations. It also randomized the left-right presentation order of the pair to be compared. The program also drove a set of speakers via a sound card to play pre-recorded instructions for each step of the test. As directed by the program, the test administrator presented the samples and recorded the

observer's responses through the program. The recorded instruction for paired comparison test was:

You will be presented with a pair of prints. Please identify the one that you think has a better color reproduction quality based on your everyday experience of color. Please ignore other printing defects other than color related defects. If you are not sure, you are encouraged to make a guess, but you must make a choice as to which print has a better color reproduction quality.

For the rank order test, the observers were asked to use the same criteria to rank the samples from the best to the worst.

Immediately after the paired comparison test, the subjects were asked to perform the rank order test under the same condition. The experiment administer recorded the rank order for each set of prints. The time used for the test varied from around 20 minutes to 45 minutes depending on the individual subject.

Results

Data from the paired comparison test were in the form of paired comparison tables. They were processed by another module of the computer program that followed the data processing scheme for incomplete data set processing given by Torgerson(). Data from the rank order test was also processed by the same computer program. Another module of the program converted the rank order data into the corresponding paired comparison tables. The converted tables were then processed in the same way the paired comparison data were processed. The scaled values were tabulated in Table 1 and shown in Figs. 2-4.

Table 1. Measured color quality (scaled values) of the five different color correction algorithms by the method of paired comparison and the method of rank order, respectively.

Sample ID	Paired comparison			Rank order		
	People	Places	Things	People	Places	Things
1	0	0.62	1.09	0	0.38	1.32
2	0.71	0.83	0	0.37	0.85	0
3	1.0	0	0.99	1.15	0	1.09
4	1.24	1.21	0.65	1.23	0.7	0.85
5	0.44	0.1	0.12	0.65	0.15	0.67

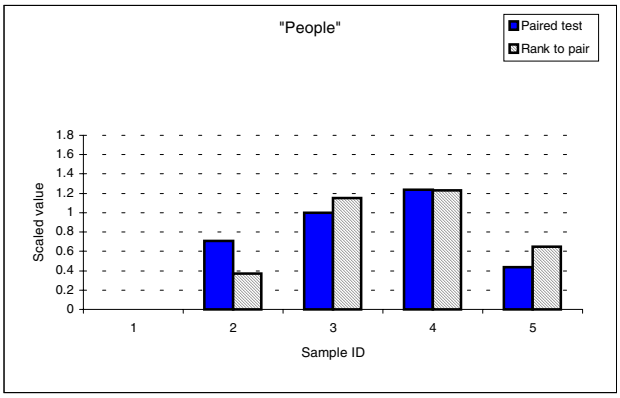

Figure 3. Comparison of measured preferred color quality (scaled value) of the "People" image with two measurement methods, respectively. Each sample ID represents a print sample printed using one of five different color correction algorithms. "Paired test" represents scaled values measured the method of paired comparison and "Rank to pair" represents scaled values measured by the method of rank order but converted to paired comparison data.

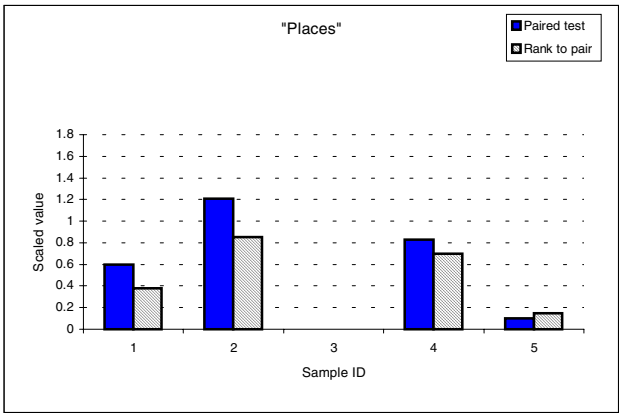

Figure 4. Comparison of measured preferred color quality (scaled value) of the "Places" image with two measurement methods, respectively. Each sample ID represents a print sample printed using one of five different color correction algorithms.

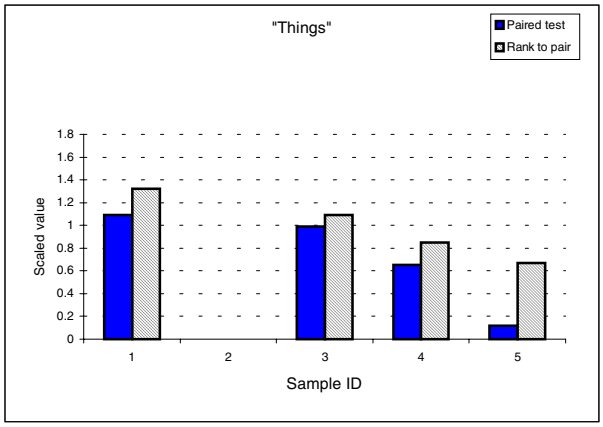

Figure 5. Comparison of measured preferred color quality (scaled value) of the "Things" image with two measurement methods, respectively. Each sample ID represents a print sample printed using one of five different color correction algorithms.

In addition to the method of computing the scaled values by converting the rank order data to paired comparison data, scaled data can also be computed by Eq. 1 and 4. The normalized rank order scores were computed using Eq. 1 and normalized by a factor of 2.77. Scaled values were also calculated using Eq. 4. Figs. 6-8 show the comparisons of scaled values calculated by three different methods.

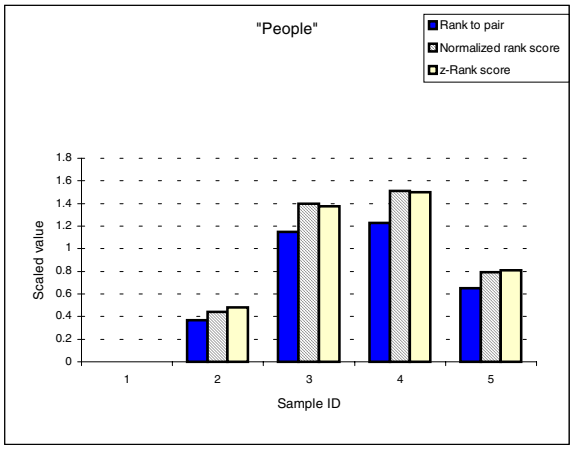

Figure 6. Comparison of the scaled values of the "People" printed images computed based on the rank order data with three different methods, respectively. "Rank to pair" represents scaled values computed by converting rank order data to paired comparison data and then applying the law of comparative judgment; "Normalized rank order score" represents scaled values computed using Eq. 1 with a normalization factor of 2.77; "z-Rank score" represents scaled values computed using Eq. 4.

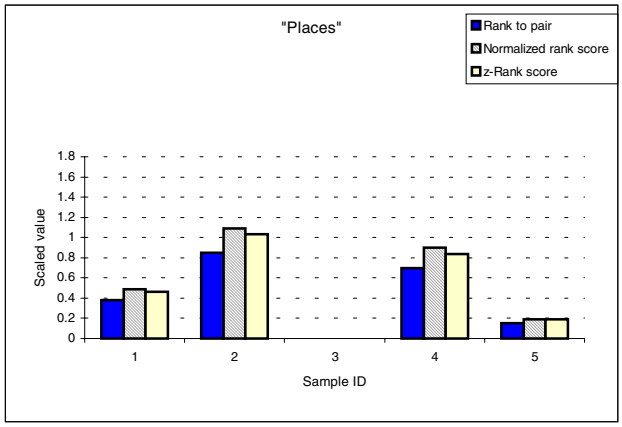

Figure 7. Comparison of the scaled values of the "Places" printed images computed based based on the rank order data with three different methods, respectively.

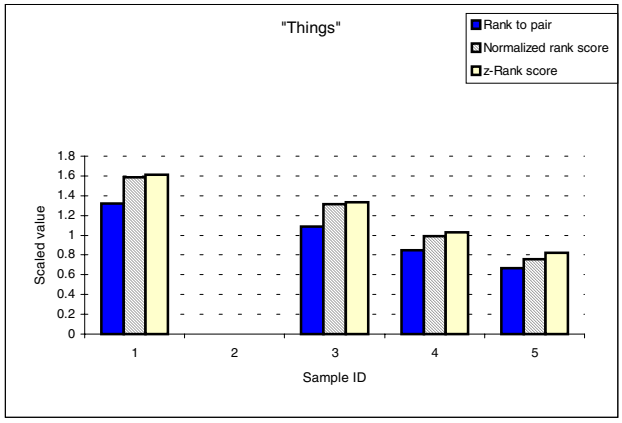

Fig. 8. Comparison of the scaled values of the "Things" printed images computed based on the rank order data with three different methods, respectively.

Discussion

Measurement errors are important factors to examine and they need to be analyzed in two different steps. The first step is to estimate the precision of the measured scaled values (the measured color quality preferences in this case). The second step is to estimate the error of fit or the difference between the predicted proportions of choice calculated based on the scaled values and the measured proportions of choice.

Error estimation of the measured scaled value is complicated. However, the estimation can be done with numerical simulation.[5] For the case of 61 subjects and 5 samples the estimated theoretical standard error of prediction is approximately 0.12 or a 95% confidence interval of 0.24. The simulated potential scaled value

distributions for five algorithms for the "People" group image is shown in Fig. 9.

For the majority of the measured scaled values by the two methods are not different at a 95% confidence level. There are a few exceptions such as the No. 5 algorithm in the "Thing" group. X squared tests of residues of the predicted proportion of choices did not show significant bias in the fit. Figs. 7-8 show that the methods of data reduction for the method of rank order produce the same results.

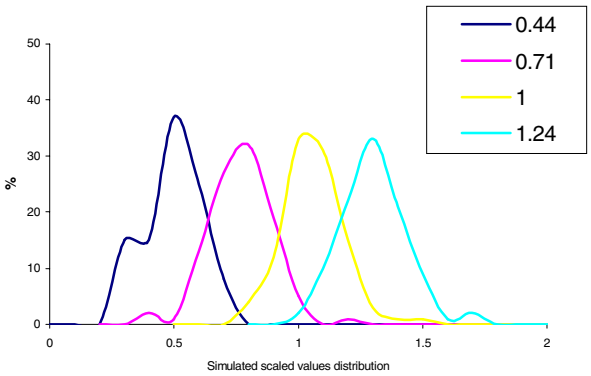

Figure 9 Simulated potential distributions of scaled values for the five color algorithms for the "People" group images.

Conclusions

With both theoretical analysis and actual experimental data, this paper proves that the methods of paired comparison and the method of rank order produced similar results when used to psychophysically scale image reproduction color quality. Data reduction methods for rank order data produce virtually the same results as that computed by more complicated comparative judgment models. However, it is important to use stimuli of similar scaled values (or highly confusable samples) in the test set as shown by the theoretical derivations given here.

Acknowledgements

The author would like to thank Mr. Russell Huffman, Mr. Ray Clark, and Mr. Paul Robinson at Lexmark International, for assisting the experiment, and all the subjects for volunteering.

References

1. Bartleson, C.J., in Optical Radiation Measurements, Vol. 5 Visual Measurements, C. J. Bartleson and F. Grum Eds. Chapter 8, Academic press, Orlando, 1984.
2. Peter, G. Engeldrum, Psychometric scaling, A toolkit for imaging system development, Imcotek press. (2000)
3. Torgerson, W.S. Theory and methods of scaling, John Wiley & Sons, Inc. (1958)
4. Kate Hevner, An empirical study of three psychophysical methods, J. Gen. Psychol., **4**, 191-212. (1930)
5. C. Cui, to be published elsewhere.