Qualifying Image Quality from Pairwised Comparisons

Tuo Wu*, Joyce Farrell**, and Amnon Silverstein**
****Hewlett-Packard Company, San Diego, California***
*****Hewlett-Packard Laboratory, Palo Alto, California***

Abstract

The method of paired comparisons is one of the most common experiment techniques to quantify image quality. The common data analysis is the Thurstone's traditional procedure that assumes normal distribution and equal variance of observer judgments across the test samples. When the distributions of observer judgments are apparently unequal, less restrictive method should be used. However, most existing statistical procedures for the purpose are limited in practice to the data with complete pairwised comparisons and with no unanimous judgments. This paper describes a new procedure that can overcome the limitations. It utilizes linear regression and iterative error reduction techniques to directly estimate the standard deviations of observer judgments and the scale values of test samples without approximation.

Introduction

The method of paired comparisons generates subjective data about the relative quality of two test samples. After repetitive trials by a number of observers or judges, the proportion of the times one sample is preferred over the other can be used to derive quantitative measurement about the difference between the two samples with some assumptions.

Though the observed proportions can be used directly as the predicted probabilities of preferences of the observer population, the data are most useful to derive quantitative image quality measurement about the test samples. This process is commonly known as one-dimensional psychophysical scaling. It is a statistical data reduction process that converts the ordinal paired comparison data into quantitative measures at least on an interval scale. In practice, the most common scaling procedures for the paired comparisons are those based on the *law of comparative judgment by* Thurstone.[1,2]

The Law of Comparative Judgment

The law of comparative judgment is presented as a mathematical model for relating the image quality scale values of a set of test samples to the observed proportions from the experiment. The law is constructed based on the following assumptions:

1. Each test sample gives rise to a value on the image quality scale during the comparative judgment.
2. The value of a test sample may be higher or lower on repeated presentations to observers. It is postulated that these values are normally distributed with mean value R_k and standard deviation σ_k, respectively.
3. For test samples *j* and *k*, the joint distribution of the difference is also normally distributed with a standard deviation as

$$\sigma_{jk} = \sqrt{\sigma_j^2 + \sigma_k^2 - 2r_{jk}\sigma_j\sigma_k} \qquad (1)$$

where r_{jk} is the correlation coefficient between test samples *j* and *k*.

4. The difference in scale values between samples *j* and *k* can, thus, be expressed in the following form:

$$R_k - R_j = z_{jk}\sqrt{\sigma_j^2 + \sigma_k^2 - 2r_{jk}\sigma_j\sigma_k} \qquad (2)$$

where z_{jk} is the standard normal deviate corresponding to the proportion that image *j* is judged greater than image *k*.

Eq. (2) is the mathematical form of the law of comparative judgment. Thurstone classifies the law into five cases under different restrictive assumptions.

Table 1. The law of comparative judgment case models[3]

Case	Constrain	Expression
I, II	σ_k , σ_j unrelated $0 < r_{jk} \le 1$	$R_k - R_j = z_{jk}\sqrt{\sigma_k^2 + \sigma_j^2 - 2r_{jk}\sigma_j\sigma_k}$
III	σ_k , σ_j unrelated $r_{jk} = 0$	$R_k - R_j = z_{jk}\sqrt{\sigma_k^2 + \sigma_j^2}$
IV	$\sigma_k \cong \sigma_j$ $r_{jk} = 0$	$R_k - R_j = 0.707 z_{jk}\left(\sigma_j + \sigma_k\right)$
V	$\sigma_k = \sigma_j = \sigma$ $r_{jk} = 0$	$R_k - R_j = \sqrt{2} z_{jk}\sigma$

Since the scale values are really affective values estimated on a psychophysical continuum which is usually an interval scale without absolute zero and absolute scale unit, there are additional postulates to the law of comparative judgment:

a) The unit of the scale is proportional to the mean of all the above-mentioned standard deviations σ_k.
b) The mean of the standard deviations for all test samples is defined as unity:

$$\frac{1}{n}\sum_{k=1}^{n}\sigma_k = 1 \qquad (3)$$

Thurstonian One Dimensional Scaling

According to the normal assumption, we can convert the proportions to normal deviates or z-scores. The data are usually arranged in a matrix format. Note that the principal diagonal terms of the z-score matrix are filled with zeros assuming each test sample were also compared with itself and resulted in a proportion of 0.50.

Eq. (4) is the Thurstone general solution for scaling paired comparison data by the law of comparative judgment. It is known as the method of column means. Many researchers have demonstrated that the method is equivalent to minimize $\left(z_{jk} - z_{jk}\right)$ errors of observed and estimated z-scores.

$$R_k = \frac{1}{n}\sum_{k=1}^{n} z_{jk}\sqrt{\sigma_j^2+\sigma_k^2-2r_{jk}\sigma_j\sigma_k} \qquad (4)$$

The column mean method is only adequate when the data matrix is complete with no missing values. If the data matrix is incomplete, the scale values should be estimated using the traditional procedure.[4] Since r_{jk} value is difficult to be determined, the law is only solvable only in under Case III, IV and V conditions in practice.

Case V: $$R_k = \frac{\sqrt{2}\sigma}{n}\sum_{k=1}^{n} z_{jk} \qquad (5)$$

Case IV: $$R_k = \frac{0.707}{n}\sum_{k=1}^{n} z_{jk}\left(\sigma_j+\sigma_k\right) \qquad (6)$$

Case III: $$R_k = \frac{1}{n}\sum_{k=1}^{n} z_{jk}\sqrt{\sigma_j^2+\sigma_k^2} \qquad (7)$$

where j, k = 1, 2, …, n, and n is the number of samples.

The Case V solution is the simplest but the most restrictive form of the law of comparative judgment. Because of its simplicity, a Case V solution is the most widely used in practice. As shown by Eq. (5), once the observed z-scores are obtained, the scale values about test samples are readily estimated.

A Case V solution is only adequate if both the normality and equal variance assumptions about observer judgments across test samples are met. If the variances of observer judgments are apparently unequal, a solution for less restrictive Case IV or Case III that accounts for the unequal variances should be used. If the number of trials is sufficient and the experiment is well under control. In practice, unequal variances are most likely caused by the non-homogeneity of test samples. Since it is inevitable to have some non-homogeneity with test samples, the accuracy of the scaling results can almost always be improved if a less restrictive Case IV or III solution is used.

In order to scale paired comparison data under the Case III or Case III assumptions, the standard deviation σ_k about each test sample must be estimated. Many procedures have been proposed for this purpose since the law was introduced. The most well known are probably the Thurstone procedure for Case IV,[5] the Burros procedure for Case III,6 the Burros-Gibson procedure for Case III,[7] and the Torgerson procedure for Case IV[4]. One thing in common about these procedures is that they all are based on some sort of Case IV approximations. Also, except for the Torgerson procedure, they are only adequate when the observed z-score matrix is complete without any missing values either caused by incomplete design of experiment, or due to unanimous judgments. Though the Torgerson procedure is capable of dealing with incomplete data, the procedure is found not always reliable in practice.

A Proposed Iterative Regression Procedure for Case III

When describing the Case IV scaling procedure, Thurstone[9] pointed out that the plot of the columns of the observed z-scores and the fit of regression lines can be used to examine the model assumptions. If the fits are definitely linear, it supports the normality assumption. If the slopes of the straight lines are also approximately the same, it further supports the equal variance assumption. If this is the case, a Case V solution can be used advantageously. If not, a better scaling results can be obtained by a Case IV or III solution. This suggests that the regression slopes are possible to be used directly to estimate the standard deviations. Following this suggestion, Torgerson[4] developed the Case IV regression approximate procedure.

Guilford[8] further suggested that the linear regression slope of the row z-scores on the estimated scale values is inversely proportional to the standard deviation of the corresponding sample. He also noticed that the intercepts of the regression lines are approximately the same as the estimated scale values in magnitude. These findings serve the foundation of the proposed scaling procedure.

Assume we have observed z-scores from a paired comparison experiment. If the standard deviations about the test samples are known, we shall be able to construct an x-score matrix according to Eq. (8), and estimate the scale values under Case III assumptions according to Eq. (9). The scale values about the samples should be estimated on a psychological scale with the scale unit equal to the mean standard deviations of the test samples.

$$x_{jk} = z_{jk}\sqrt{\sigma_j^2+\sigma_k^2}\ ,\ j, k = 1, 2, \ldots, n \qquad (8)$$

$$R_k = \frac{1}{n}\sum_{k=1}^{n} x_{jk} \qquad (9)$$

Plot the rows of x-scores on the estimated scale values and fit with straight lines by linear regressions. If the paired comparison data are errorless under the law, we shall expect to have a perfect fit of a straight line to each series of data points and all the straight lines shall have the same slope. The slope should be equal to unity.

It is easy to see that the above statement is true under the Case V assumptions of the law. The properties about the regression slopes should be also true for the more general Case III with the x-scores and the estimated scale values if the model is consistent with the model assumptions. If this is not true, the image quality scale obtained will not be uniform. Therefore, the deviation between the regression slopes and a common slope value of unity are fundamental by the law of comparative judgment, and they can be used as measures for test of internal consistency with the model assumptions.

From the analytical geometric viewpoint, if the slopes of the regression lines are proportional to the standard deviations of the test samples, the distance between these lines should be proportional to differences in scale values between the test samples. Therefore, the intercepts of the regression lines can be used directly to estimate the scale values about the test samples.

According to the above analogy, the right scaling procedure under Case III assumptions is to estimate σ_k values in a way so that the variation between the regression slopes be minimized. Since σ_k and R_k are mutually dependent in the regression process, they have to be estimated simultaneously. One way is to use iterative error reduction technique. Based on this assumption that the regression slope is inversely proportional to the standard deviation of its corresponding test sample, we have

$$\sigma_k \sim \frac{1}{c_k}$$

where c_k is the regression slope of the *k*th row of x-scores on the estimated scale values. Based on this assumption, we can estimate the standard deviation according to the following iterative Eq. (10).

$$\sigma_k^{(i)} = \left[\frac{d^{(i)}}{c_k^{(i)}} \right] \sigma_k^{(i-1)} \qquad (10)$$

where k = 1,2, …, n; i = 1, 2, …, m, and $d^{(i)}$ is a constant. According to the additional postulate b), $d^{(i)}$ is defined as

$$d^{(i)} = \frac{1}{n} \sum_{j=1}^{n} \frac{\sigma_k^{(i-1)}}{c_k^{(i)}} \qquad (11)$$

The iteration procedure starts with estimates of R_k and σ_k under Case V assumptions. The deviation among the regression slopes is measured by computing the standard deviation of the regression slopes. We set up a tolerance $\Delta_{tolerance}$ for the deviation of the regression slopes. After each iteration, a new set of standard deviations σ_k are estimated, a new x-score matrix are formulated, and a new set of scale values R_k are estimated. We continue the iteration process until tolerance condition is met.

This iterative process can be easily achieved through a computer program. In order to have the program converge successfully, like an optimization routine, another limit should be set up about the precision of error reduction. In practical situation the deviation of regression slopes may never reach the tolerance.

Test of Goodness of Fit

There is no clear-cut statistical test if the scaling procedure is consistent with the model assumptions. The practical approach is rather to test the goodness of fit by examining how well the estimates serve to predict the observed proportions. The most widely test is to compute the average absolute deviates (AAD) between observed and predicted proportions. If the average is "small," one may conclude that the model fits the problem adequately. The Mosteller χ^2 test[9] might be considered as a clear-cut test, but it is advised to be used cautiously.[4,8] It is proposed that the deviation of the regression slopes be used as another method to test the goodness of fit of the models.

The proposed procedure has been tested empirically with both simulated data and data from experiments, and the results are compared with existing procedures. It has shown that the proposed procedure performs at least the same as the existing ones. In most cases, it performs even better. However, the major advantage of the proposed procedure is its capability to deal with incomplete data, which is of importance in practical applications.

An example

The following example is known as the Food data from Gulliksen.[10] As shown in Table 2, the data contain missing values due to the incomplete design of experiment as well as unanimous judgments. The experiment tested 15 food objects and surveyed 92 judges. Because the data are incomplete, the Thurstone solution for Case IV, the Burros solution for Case III, and the Burros-Gibson solution for Case III procedures can not apply. Though the Torgerson Case IV procedure can apply, it failed to yield valid estimates with this example.

Table 2. The Food paired comparison experiment data

	TP	T	TL	P	TB	PL	L	TS	PB	B	PS	LB	S	LS	BS
TP	x	-	-	-	-	-	18	-	-	2	-	5	2	4	0
T	-	x	-	24	-	16	13	-	8	1	9	3	0	1	1
TL	-	-	x	37	-	-	-	-	13	5	10	-	2	-	3
P	-	68	54	x	39	-	21	30	-	4	-	3	0	6	5
TB	-	-	-	53	x	37	40	-	-	-	17	-	16	9	-
PL	-	76	-	-	55	x	-	45	-	22	-	-	12	-	2
L	74	79	-	71	52	-	x	46	31	13	22	-	6	-	10
TS	-	-	-	62	-	47	46	x	31	32	-	24	-	-	-
PB	-	84	78	-	-	-	60	61	x	-	-	-	21	15	-
B	90	91	87	88	-	70	79	60	-	x	35	-	19	23	-
PS	-	83	82	-	75	-	70	-	-	57	x	43	-	-	-
LB	87	89	-	89	-	-	-	68	-	-	49	x	37	-	-
S	90	92	90	92	76	80	86	-	71	73	-	55	x	-	-
LS	88	91	-	86	83	-	-	-	77	69	-	-	-	x	-
BS	92	91	89	87	-	90	82	-	-	-	-	-	-	-	x

Note: (-) not compared in the experiment

The estimated standard deviations and scale values are listed in Table 3 and 4, respectively. Fig. 1 and 2 are the regression plots as the results of the Case V solution and the proposed Case III solution, respectively. Table 5 lists the results from tests for goodness of fit. For the Case V solution, the standard deviation is one for all food objects.

As shown in Fig. 1, it is apparent that some regression slopes are the different from others, which indicates that a Case V solution may not be adequate in this situation. As shown in Fig. 2, the proposed regression procedure has successfully come with the nearly same regression slopes by counting for the unequal variances. The test results are consistent with the argument about the regression slopes.

Table 3. Estimated standard deviations

	TP	T	TL	P	TB	PL	L	TS
Torgerson, Case IV	1.999	1.467	1.730	1.439	1.378	0.576	0.517	1.628
Proposed, Case III	1.017	0.958	0.988	0.914	1.123	0.828	0.870	1.291
	PB	**B**	**PS**	**LB**	**S**	**LS**	**BS**	**Mean**
Torgerson, Case IV	1.628	0.938	2.628	1.688	-0.841	-0.751	-1.024	N/A
Proposed, Case III	0.910	0.700	1.308	0.913	1.027	1.107	1.046	1.000

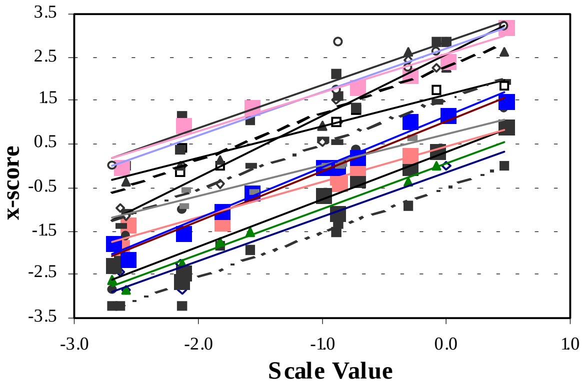

Figure 1. Case V solution regression plot

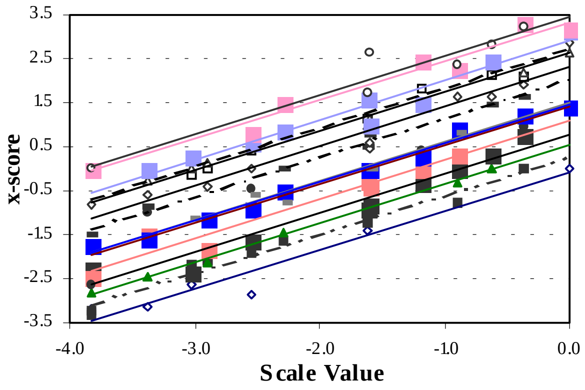

Figure 2. Proposed Case III solution regression plot

Table 4. Estimated scale values

	TP	T	TL	P	TB	PL	L	TS
Thurstone, Case V	0.000	0.465	-0.088	-0.300	-0.718	-0.864	-0.890	-0.992
Proposed, Case III	0.000	-0.366	-0.624	-0.893	-1.181	-1.602	-1.616	-1.605
	PB	**B**	**PS**	**LB**	**S**	**LS**	**BS**	
Thurstone, Case V	-1.587	-2.135	-1.816	-2.152	-2.581	-2.631	-2.696	
Proposed, Case III	-2.281	-2.550	-2.901	-3.025	-3.372	-3.829	-3.823	

Table 5. Test of goodness of fit

	Thurstone Case V	Torgerson Case IV	Proposed Case III
AAD	0.04	N/A	0.02
Mosteller χ^2	80.45	N/A	44.08
stdev of slopes	0.177	N/A	0.022

References

1. Thurstone, L. L., “A law of comparative judgment,” *Psychol Rev.*, **34**, p.273-286, 1927
2. Thurstone, L. L., “Psychophysical analysis,” *Amer. J. Psychol.*, **38**, p.368-389, 1927
3. Bartleson, C. J. and Grum, F., “Optical radiation measurements, **5**, Academic Press,
4. Torgerson, W. S., “Theory and Methods of Scaling,” John Wiley & Sons, Inc., 1958
5. Thurstone, L. L., “Stimulus dispersions in the method of constant stimuli,” *J. Exp. Psychol.*, vol. **15**, p.284-297, 1932
6. Burros, R. H., “The application of the method of paired comparisons to the study of reaction potential.” *Psychol. Rev*. **58**, No.**1**, p.60-66, 1951
7. Burros, R. H. and Gibson, W. A., “A solution for Case III of the law of comparative judgment,” *Psychometrika*, **19**, p.57-64, 1954
8. Guilford, J. P. “Psychometric Methods.” Second Ed., McGraw-Hill, 1954
9. Mosteller, F., “Remarks on the method of paired comparisons: III. A test of significance when equal standard deviations and equal correlations are assumed.” *Psychometrika*, **16**, p.207-218, 1951
10. Gulliksen, H., “A least squares solution for paired comparisons with incomplete data.” *Psychometrika*, **21**, p.125-134, 1956

Biography

Tuo Wu received B.S. degree in Mechanical Engineering from Beijing, China in 1984, and M.S. degree in Printing Technology from RIT in 1990. He jointed Hewlett-Packard Co. in 1997 working on color science and image and print quality. email: tuo_wu@hp.com