Evaluation of Incomplete Paired-Comparison Experiments

Yuan LI, Stephen Westland and Vien Cheung
School of Design, University of Leeds, Leeds (UK)

Abstract

Incomplete paired comparison is an important technique for color-imaging problems because it can avoid observers to compare every possible pairs since the number of paired comparisons for n stimuli is n(n-1)/2 which becomes prohibitive for large values of n. However, the experimental designer often struggles with questions such as what is the smallest limit the proportion of paired comparisons included that will still allow reliable estimations of scale values? Fortunately a Monte-Carlo computational simulation is carried out with a model of an ideal observer and the results shows that the proportion of paired comparisons that is included is more critical than the number of observers who make those observations [1]. This work aims to test the results from computational simulation with 25 real observers and 10 stimuli from the gray scale. The work suggests when each observer estimates the same proportion of paired comparisons included the more proportion of pairs and number of observers, the more accurate scale values will be produced and the proportion of pairs is more critical than the number of observers who make those observations, which quite agrees with the findings from the computational simulation. The work also suggests when the each observer estimates a different proportion of paired comparisons the more proportion of paired comparisons will not always produce a more accurate scale values.

Introduction

Previous work presented at CIC used a computational model to simulate virtual observers taking part in an incomplete paired-comparison experiment [1]. This explored the impact on the number of observers and of the proportion of the total possible number of paired comparisons on the accuracy of the scale values that could be estimated from the experimental data. The simulation required a number of assumptions that may or may not be justifiable in a given experimental situation. The work presented in this study analyses raw data from a real paired-comparison psychophysical experiment in order to further understand the parameters of incomplete paired-comparison experiments.

One of the fundamental problems in psychophysics is the assignment of scale values to the individual members of a set of stimuli, with respect to some physical attribute of the stimuli, and with respect to the mental responses, which they evoke [2]. To obtain interval scale values, which have equal spaced units between each pair of neighbor scales, the paired-comparison technique is widely used [3-6]. The basic process of the paired-comparison method consists of serially presenting pairs of samples to an observer; the observer is asked to indicate which one of the two samples has the most characteristics the study administrator is investigating. The raw data are used to construct a table of preference ratios [7]. Table 1 shows how the table of preference ratios and the standard normal deviate matrix are constructed for three stimuli. The upper table shows the frequency matrix F. The middle table shows preference ratios matrix P generated from matrix F. The lower table shows the response differences in units of standard normal deviates corresponding to matrix P. By example, if a pair is viewed 10 times and one stimulus is preferred 9 times out of 10, then the preference ratio would be 0.9 (see Figure 1); this would correspond to a response difference of 1.28 in units of standard normal deviate (similarly, if the preference ratio was 0.5 then the response difference would be zero).

Thurstone constructed a model to generate scale values from paired-comparison data; he specified five cases for this model and also identified the assumptions needed [3, 4]. According to Thurstone's Law of Comparative Judgment, the means of the columns in the lower table are estimates of the scale values for the three samples. However, there are two limitations with this Summation method. Firstly, the method requires that the complete matrix of comparisons. Secondly, if all observations agree that one stimulus is preferred over another there is no information available so that some of the preference ratios will be 1 or 0 [8, 9].

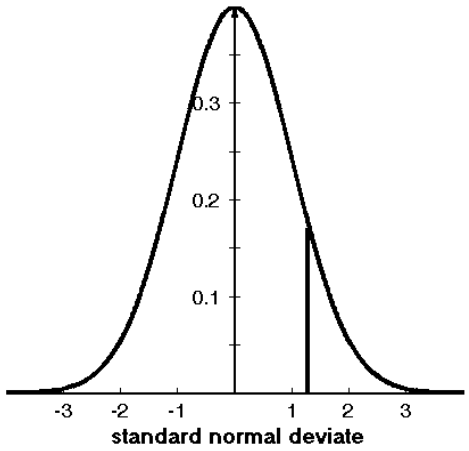

Figure 1: Relationship of the preference ratio to response difference in units of standard normal deviate. The solid vertical line represents the situation where the preference ratio is 0.9; the area under the curve to the left of the solid line is 90% and this corresponds to 1.28 standard normal deviate units.

In 1955 and 1956, Morrisey and Gulliksen separately proposed the least squares solution to solve this problem [5, 6]. The only difference of these two methods is that in Gulliksen's work an iterative procedure was suggested which can markedly reduce the computation time. Based on the data collected from paired comparison method, for a p column and q row standard normal deviate matrix paired we then construct matrices **A** and **d** such that

$$As = d \tag{1}$$

where **d** is a $(q+1) \times 1$ matrix of the summation of response differences of each row and **A** is a $(q+1) \times p$ matrix that defines the pair-wise comparisons that are made. Again, for example, on basis of the lower table in table 1, Equation 1 can be written in full as

$$\begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \\ 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix} = \begin{bmatrix} d_{1,2} \\ d_{1,3} \\ d_{2,3} \\ 0 \end{bmatrix} \quad (2)$$

where s_i are the scale values and $d_{i,j}$ are the response differences between s_i and s_j for i, $j \in \{1,2,3\}$. The last row in matrices **A** and **d** imposes the constraint that the sum of all scale values is zero. Equation 1 can be solved using MATLAB's backslash operator, thus **s** = **A\d**. The advantage of Morrisey-Gulliksen's method over the Summation method is that it can be solved even when every possible paired comparison is not carried out.

	1	**2**	**3**
1	0	3	1
2	2	0	4
3	4	1	0

	1	**2**	**3**
1	0.50	0.60	0.20
2	0.40	0.50	0.80
3	0.80	0.20	0.50

	1	**2**	**3**
1	0.00	0.25	-0.84
2	-0.25	0.00	0.84
3	0.84	-0.84	0.00

Table 1: Example table construction for three stimuli. The upper table shows the frequency matrix F for the three stimuli S_i ($i \in \{1,2,3\}$). The middle table shows the preference ratios Matrix P corresponding to the Matrix F in the upper table. The lower table shows the corresponding response difference obtained using the procedure outlined in Figure 1.

In 2009, a computational simulation experiment was carried to investigate what proportion of the matrix is required in order for the Morrisey-Gulliksen's methods to be valid and how robust the methods are as the matrix becomes sparser [1]. The study also considered the relationship between the sparseness of the matrix and the number of observers who take part in the paired-comparison experiment. The findings suggested that the number of observers who take part in the experiment is less critical than the proportion of possible paired comparisons that are carried out as Figure 2 shows and 40-50% of all the possible paired comparisons are suggested to be considered.

This work aims to test the model based on computational simulation experiment by the Morrisey-Gulliksen's methods with real observers and stimuli.

Experimental

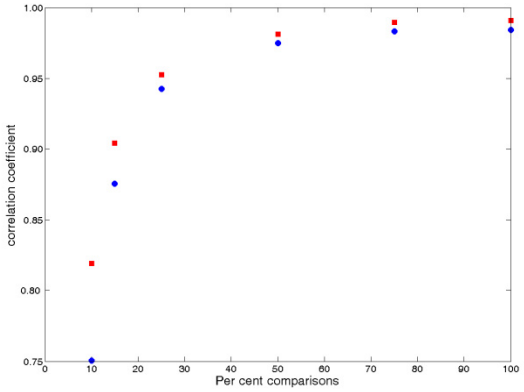

Figure 2: Mean correlation coefficient for various degrees of completion of the paired-comparison experiment for 10 (blue circles) and 20 (red squares) observers (n = 30).

In this work we aim to analyze data from a real psychophysical experiment that employs the paired-comparison technique. We stress that the actual nature of the experiment was relatively unimportant; what is required is that experimental data are available for a paired-comparison experiment in which each observer considered all possible paired comparisons. It is then possible to reanalyze the data by sub-sampling the complete experimental data. Since psychophysical data were not available, a new psychophysical experiment was carried out for the purpose of this study whereby observers were shown pairs of achromatic stimuli of varying Lightness and were asked to indicate which of each pair was darkest.

Color Stimuli

A set of 10 grey stimuli of varying Lightness values was selected for the study; a pilot experiment was used to specify the stimuli such that they formed a series in ascending Lightness with the difference between any two stimuli adjacent in the series being close to the just-noticeable difference.

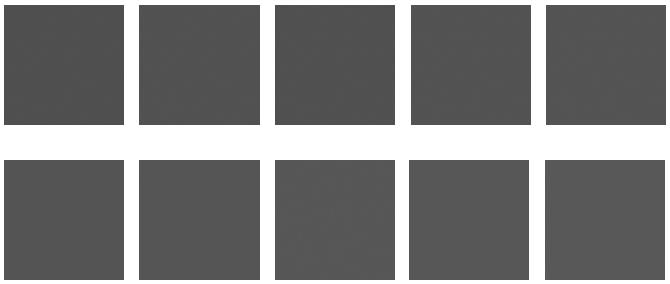

Figure 3: 10 grey stimuli. Each two of the adjacent stimuli have just noticeable difference.

For 10 stimuli there are 45 possible paired comparisons. Pairs were displayed against a grey background (L*=80) on a CRT monitor and observers were requested to indicate which of the stimulus in each pair was darker. Figure 3 illustrates the stimuli that were used.

Observers

Twenty-five observers participated in this experiment, including observers from China, UK, Iran, India, Pakistan and South Korea. All of these observers passed the Ishihara Test for Color Blindness before participating in the experiment.

Experimental Procedures

During the experiment, each observer was presented with colour stimuli on a CRT monitor at a viewing distance of 80 cm and a visual field size of 10º for each pair of stimuli. When observers were ready to conduct the experiment, the Start button was pressed to commence the experiment. Then, pairs of stimuli were presented in the centre of the monitor screen. Observers were asked to select one of the two stimuli each time according to their darkness and choose the darker one by pressing the button below it. By doing this, the next pair of images would be presented until all the 45 pairs of stimuli were estimated. A total of 1125 (45 pairs × 25 observers) observations were made. The rationale for this study was that these observations can be sub-sampled so that the results obtained with fewer than 25 observers and/or less than complete proportions of comparisons can be calculated.

For all conditions, the full data set was sampled 50 times. That is, if 10 observers were considered, each completing 90% of the comparisons, then for each trial 10 observers would be chosen at random and each would consider 90% of the paired comparisons. For each trial the scale values were calculated and compared with the true scale values. In this study it is assumed that the Lightness scale is psychophysically correct and that therefore the scale values obtained from the experiment can be compared with the L* values of the stimuli. The r^2 value between scale values and L* values is used as the performance metric. The r^2 values were averaged over all 50 trials for each set of conditions.

Note, however, that if observers undertake, say, 50% of the comparisons there are two ways of doing this. Firstly, each observer could undertake the same 50% of comparisons so that some paired comparisons are never made. Secondly, each observer would undertake a different 50% of the comparisons increasing the likelihood that all pairs are considered at least once. Both of these methods of sampling were considered in this work.

Results

Table 2 shows the mean correlation coefficients with standard error for the L* values and the predicted scale values for the Morrisey-Gulliksen's method for various numbers of observers *k* and for two levels of completeness. The values in the upper table are generally lower than those in the lower table when the matrix is full and higher than the lower table when the matrix is 90 per cent completed; this suggests that it is better for all observers to undertake different paired comparisons in the case of an incomplete experiment.

Table 3 shows results obtained when not all paired comparisons are considered. In the upper table, for each of the 50 trials a different (randomly selected) set of paired comparisons was evaluated according to the chosen percent completion rate required but within a trial the same set of paired comparisons was evaluated by each observer. In the lower table, a different set of paired comparisons was evaluated by each observer for each of the 50 trails.

Number of observers *k*	Complete matrix (standard error)	90% (standard error)
5	0.9554 (0.0029)	0.9497 (0.0030)
15	0.9649 (0.0012)	0.9606 (0.0018)
25	0.9684 (-)	0.9655 (0.0009)

Number of observers *k*	Complete matrix (standard error)	90% (standard error)
5	0.9547 (0.0027)	0.9520 (0.0030)
15	0.9623 (0.0012)	0.9646 (0.0011)
25	0.9684 (-)	0.9656 (0.0005)

Table 2: the upper table shows the mean correlation coefficients (50 trials) for complete matrix of pair-wise comparisons for 10 stimuli (n = 10) and various numbers of observers, when each observer estimates the same proportion of paired comparisons for each trial. The lower table shows the mean correlation coefficients (50 trials) for complete matrix of pair-wise comparisons for 10 stimuli (n = 10) and various numbers of observers, when each observer estimates a different proportion of paired comparisons for each trial. The standard error for each mean correlation coefficient is in the bracket.

Per cent comparisons	*k* = 10 (standard error)	*k* = 20 (standard error)
100	0.9569 (0.0021)	0.9637 (0.0007)
90	0.9531 (0.0024)	0.961 (0.0014)
70	0.9442 (0.0031)	0.9548 (0.0030)
50	0.9207 (0.0057)	0.9341 (0.0039)
30	0.8201 (0.0144)	0.8405 (0.0113)

Per cent comparisons	*k* = 10 (standard error)	*k* = 20 (standard error)
100	0.9578 (0.0022)	0.9636 (0.0008)
90	0.9598 (0.0021)	0.9636 (0.0009)
70	0.9554 (0.0029)	0.9605 (0.0018)
50	0.9486 (0.0043)	0.9557 (0.0025)
30	0.9095 (0.0076)	0.9489 (0.0030)

Table 3: Mean correlation coefficients (50 trials) for incomplete matrix of pair-wise comparisons for 10 and 20 observers, 10 stimuli, and five levels of completeness. The upper table is for the situation of each observer estimates the same set of paired comparisons. The lower table is for the situation of each observer estimates a different set of paired comparisons. The standard error for each mean correlation coefficient is in the bracket.

The upper table in Table 3 shows results for 5 different completion rates. It is evident that in the cases of both 10 and 20 observers the performance drops as the preference matrix become more sparsely populated.

Figure 4 illustrates the relationship between the correlation coefficient, the number of observers and the degree of completeness of the experiment. The two lower lines (green for 10 observers and purple for 20 observers) in Figure 4 correspond to the same situation as the upper table of Table 3. The two upper lines (blue for 10 observers and red for 20 observers) are corresponding to the same situation as the lower table of Table 3. It is evident from both situations that the effect of the per cent completion of the paired comparisons has a far greater effect than the number of observers on the performance. This is the same observation that was made in the previously simulated experimental work [1] as illustrated in Figure 2. The yellow and purple lines are above the blue and red lines in the Figure 4. It indicates it is better that each observer undertakes a different sub-set of paired comparisons from all those available. The Figure 4 also shows the rms value grows more rapidly in the yellow and purple lines than the blue and red lines as the proportion of the full matrix grows and when the proportion is as low as 30 per cent the distance of rms value between the two situations are much higher than the high proportions after 50 per cent. It indicates that to obtain a certain rms value or accuracy result the a smaller number of observers and proportion of matrix can be applied when all observers estimate a different set of paired comparisons than all observers estimate the same set of paired comparisons.

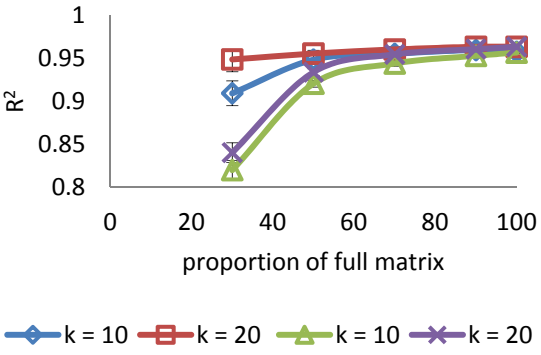

Figure 4: Mean correlation coefficient with standard error for various degrees of completion of the paired-comparison experiment for 10 (blue) and 20 (red) observers in the case that each observer evaluates a different set of paired comparisons for each of 50 trials; and 10 (green) and 20 (purple) observers when each observer evaluates the same set of paired comparisons within each of 50 trials but different sets for each trial.

Conclusions

The design of paired-comparison experiments is important for a number of color-imaging related problems. For large number of stimuli it is not always practical to be able to complete all the possible paired comparisons and scale values are often estimated from a partially complete experiment. The design of such experiments has been explored in this and previous work [1] and two situations were considered. The first one is where each observer estimates the same set of paired comparisons for each of the 50 trials. The second one is where each observer estimates different sets.

The findings in this study, based on an analysis of real experimental data, suggest that the greater the number of observers and the greater the proportion of the full matrix evaluated, the more accurate the estimates of the scale values. However, more accurate estimates of the scale values result when observers evaluate different sets of paired comparisons. It was also found that in the case of observers undertaking the same paired comparisons the number of observers is relatively unimportant compared with the proportion of paired comparisons evaluated. By contrast, in the case of observers undertaking different paired comparisons the number of observers is important when the proportion of paired comparisons evaluated is low (<50%).

The analysis of experimental data conducted in this study supports the findings from the previous study that was based on computational simulation [1]. This work therefore validates the previous computational study and suggests that further investigation using computational simulation could be fruitful.

References

[1] Cheung V, Westland S & Li Y (2009), Experimental Design in Incomplete Paired-Comparison Experiments, *Proceedings of the IS&T/SID's Seventeenth Colour Imaging Conference*, 107-110, Albuquerque, New Mexico.

[2] Marks LE & Gescheider GA (2002), *Phychophysical Scaling*. In Pashler HE & Atkinson RD (Eds.) *The Stevens' Handbook Of Experimental Psychology*. John Wiley, Chichester. 2nd edition.

[3] Thurstone LL (1927a), Psychophysical Analysis, *American Journal of Psychology*, **38**, 368-389.

[4] Thurstone LL (1927b), A law of comparative judgment, *Psychological Review*, **34** (4), 273-286.

[5] Morrissey JH (1955), New method for the assignment of psychometric scale values from incomplete paired comparisons, *Journal of the Optical Society of America*, **45** (5), 373-378.

[6] Gulliksen H (1956), A least-squares solution for paired comparisons with incomplete data, *Psychometrika*, **21**(2), 125-134.

[7] Torgerson WS (1958), *Theory and Methods of Scaling*. John Wiley, New York. 1st edition.

[8] Gescheider GA (1997), *Psychophysics: The Fundamentals*, Lawrence Erlbaum Associates, London.

[9] Engeldrum PG (2000), *Psychometric Scaling: A Toolkit for Imaging Systems Development*, Imcotek Press.

Author Biography

Yuan Li obtained BA Accountancy from Xiamen University (China) and MA Design from Leeds University (UK). She is currently a graduate student undertaking a PhD at Leeds University in psychophysical methods.