Predicting Perceptions: Proceedings of the 3rd International Conference on Appearance, pp. 89-91, Edinburgh, UK, ISBN 978-1-4716-6869-2, April 2012.

Models Estimation for Incomplete Paired-Comparison Experiments Based on Least-Squares Solution

Yuan Li, Stephen Westland, Vien Cheung
School of Design
University of Leeds, Leeds, LS2 9JT, UK
+44 (0) 113 343 3752
{tex6yl, s.westland, t.l.v.cheung} @leeds.ac.uk

ABSTRACT
Incomplete paired comparison is an important scaling technique in vision science since the total 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 for the proportion of paired comparisons included that will still allow reliable estimations of scale values? Monte-Carlo computational simulations were previously carried out using a model of an ideal observer. The results showed that the proportion of paired comparisons that is included is more critical than the number of observers who make those observations. This work aims to test the results from the computational simulation with 25 real observers and 10 stimuli from the gray scale. The psychophysical data suggest that when each observer evaluates the same pairs, accuracy of the derived scale values increases with the proportion of pairs evaluated and the number of observers; the proportion of pairs is, however, more critical and this agrees with the results of the simulation. The psychophysical data also suggest when the each observer estimates a different pairs (albeit with the same proportion of pairs being evaluated) the accuracy of scale values does not always increase monotonically with the number of pairs being evaluated.

Categories and Subject Descriptors
B.2.2 [**Arithmetic and Logic Structures**]: Performance Analysis and Design Aids – *Simulation, Verification.*

General Terms
Performance, Design, Reliability, Experimentation, Verification.

Keywords
Incomplete paired-comparison experiment, model verification, Morrissey's least squares solution.

1. INTRODUCTION
One of the fundamental problems in psychophysics is the assignment of scale values to 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 [1]. To obtain interval scale values, which have equally spaced units between each pair of neighbor scales, the paired-comparison technique is widely used [2-4]. 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 for the attribute being investigated. The raw data are used to construct a table of preference ratios [3].

***Predicting Perceptions: The 3rd International Conference on Appearance**, 17-19 April, 2012, Edinburgh, UK. Conference Proceedings Publication ISBN: 978-1-4716-6869-2, Pages: 89-91*

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 [2,5]. If all possible pairs are compared then a technique known as the Summation method can be used to derive the scale values. However, there are two limitations with this Summation method. Firstly, the method requires the complete matrix of comparisons and this could be prohibitively expensive. Secondly, if all observations agree that one stimulus is preferred over another there is no information available as to the magnitude of the difference, so that some of the preference ratios will be 1 or 0 [3]. Both of these two problems lead to an incomplete matrix of paired comparisons and Morrissey's least- squares solution is popularly used to solve these problems [4].

In 2009, computational simulation experiments were carried to investigate what proportion of the matrix is required in order for the Morrisey-Gulliksen's methods to be valid; how robust the methods are as the matrix becomes sparser; and whether each observer should evaluate the same or different set of comparisons [6]. 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 for incomplete paired-comparison experiments the accuracy of estimated scale values increases if observers evaluate different pairs as shown in Figure 1. For the case that each observer evaluates the same set of comparisons, 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 shown in Figure 2 (n=10) and 40-50% of all the possible paired comparisons are suggested to be considered as the green plots show in Figure 1. For the case that each observer evaluates a different set of comparisons, the accuracy of the estimated scale values is more sensitive to the number of observers who take part in the experiment and relatively invariant to the proportion of possible paired-comparisons that are carried out. This is especially true when n is large as the red plots show in Figure 2 and 20-30% of all the possible paired comparisons are suggested to be considered as the red plots show in Figure 1.

2. EXPERIMENTAL WORK
This work aims to test the model based on computational simulation experiments by the Morrisey-Gulliksen's methods by analyzing 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. A psychophysical experiment was therefore carried out whereby observers were shown pairs of achromatic stimuli of varying

Lightness and were asked to indicate which of each pair was darkest.

2.1 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. For the 10 stimuli there are 45 possible paired comparisons and each observer evaluated all of these. Pairs were displayed against a grey background (L*=43.4) on a CRT monitor and observers were requested to indicate which stimulus in each pair was darker.

2.2 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.

2.3 Experiment Procedure

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. The monitor was controlled using a GUI written in MATLAB. When observers were ready to conduct the experiment, the Start button was pressed to commence the experiment. 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 each condition (e.g. number of observers considered and the proportion of paired comparisons included), the full data set was sub-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 could undertake a different 50% of the comparisons increasing the likelihood that all pairs are considered at least once (in this case, each observer evaluates 50% of the pairs; the selected pairs are randomly determined). Both of these methods of sampling were considered in this work.

3. RESULTS

In Figure 3, the estimated r^2 values are plotted against the proportions for different numbers of *k*. The psychophysical data suggest that the accuracy of the estimates of scale values increases with the number of observers and the proportion of the full matrix evaluated. However, more accurate estimates of the scale values result when observers evaluate different pairs (the red data in Figure 3). It was also found that in the case of observers evaluating the same pairs 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%). Compared with Figure 2 (from the computational simulation) it was found that both red and green plots were corresponding to lower values of r^2 in Figure 3, especially for lower proportions. This indicates that the performance of the model was over-estimated especially for lower proportions for the case of n=10. The experimental data in this study generally supports the findings from the previous study that was based on computational simulation [6]. A preliminary analysis of these results was presented in 2011 [7].

4. FUTURE WORK AND DISCUSSION

Only 10 stimuli were considered in the experiment and it would be useful to explore how these results scale with very large experiments. This study indicated that performance was slightly over-estimated by the models. This work and previous simulations were based on Morrissey's least-squares solution. However, to solve incomplete paired-comparison experiments, alternative methods are available. It might be interesting to model these methods.

5. REFERENCES

1. Marks LE, Gescheider GA. Phychophysical scaling. In: Pashler HE, Atkinson RD, editors. The Stevens' Handbook of Experimental Psychology. 2nd ed.
Volume 4. Chichester: Wiley; 2002. p 91-95.
2. Thurstone LL. A law of comparative judgement. Psychological Review 1927;34(4):273-286.
3. Torgerson WS. Theory and methods of scaling. New York: Wiley; 1958. 460 p.
4. Morrissey JH. New method for the assignment of psychometric scale values from incomplete paired comparisons. Journal of the Optical Society of America 1955;45(5):373-378.
5. Thurstone LL. Psychophysical analysis. American Journal of Psychology 1927;38(3):368-389.
6. Cheung V, Westland S, Li Y. Experimental Design in Incomplete Paired-Comparison Experiments. 2009; Albuquerque, New Mexico. p 107-110.
7. Li Y, Westland S, Cheung V. Evaluation of Incomplete Paired-Comparison Experiments. 2011; San Jose, CA.

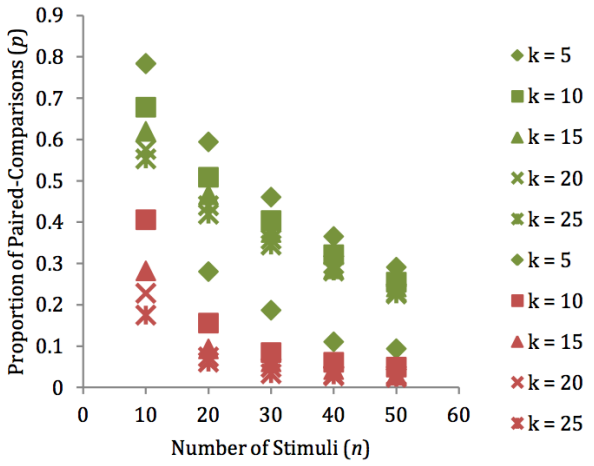

Figure 1. The proportion required to achieve given performance of mean $r^2 = 0.95$ for each case defined by n and k for both experimental designs that each observer evaluates the same set of pairs (in red) and each observer evaluates a different set of pairs (in blue). Data obtained from computational simulation [6].

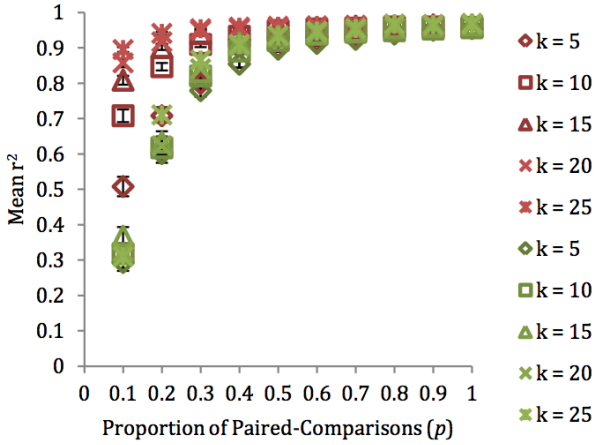

Figure 3. Mean correlation coefficient with standard error is plotted against various degrees of completion of the paired-comparison experiment for various numbers of observers based on psychophysical experimental results. Plots in red are for the case that each observer evaluates a different set of paired comparisons and the plots in green are for the case that each observer evaluates the same set of paired comparisons.

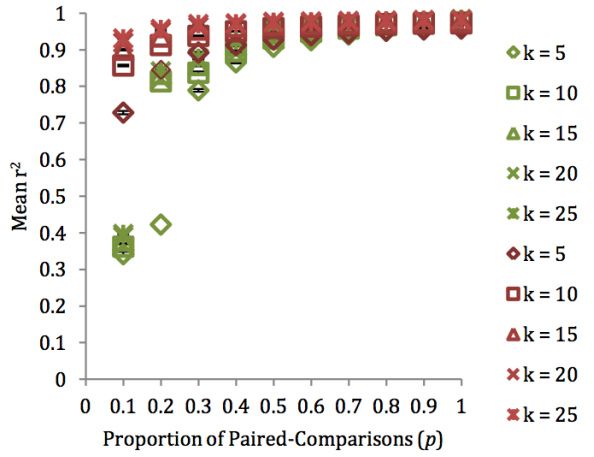

Figure 2. Mean correlation coefficient with standard error is plotted against various degrees of completion of the paired-comparison experiment for various numbers of observers for n stimuli. Plots in red are for the case that each observer evaluates a different set of pairs and the plots in green are for the case that each observer evaluates the same set of pairs. Data obtained from computational simulation [6].