Monte Carlo Analysis of Incomplete Paired-Comparison Experiments

Stephen Westland, Yuan Li and Vien Cheung
School of Design, University of Leeds, Leeds, LS2 9JT, UK
E-mail: s.westland@leeds.ac.uk

Abstract. *The use of paired-comparison psychophysical experiments is an important technique that is used widely in imaging studies. It is sometimes difficult to compare every stimulus with every other; the number of paired comparisons for n stimuli becomes prohibitive for large values of n. Thus, experiments are often designed by missing some pairs. However, the effect on the accuracy of the estimations of the scale values is not clear. Similarly, if more resources are available, would it be better to recruit more observers making the same paired comparisons or to have the original observers carry out additional paired comparisons? This work seeks to develop a framework for addressing these practical questions surrounding incomplete paired-comparison experiments design. A Monte-Carlo computational simulation is carried out with an ideal observer model. Results suggest that the proportion of paired comparisons is more critical than the number of observers with small number of stimuli.*

INTRODUCTION

A common problem for the psychophysicist is to derive the best possible set of numerical responses from a set of mental comparisons made by an observer or by a group of observers[1-2]. Not only are these responses to be arranged in their correct subjective order, as determined from the consensus of comparisons by all observers, but also they are to be correctly spaced along a scale of numerical response values (i.e. interval scale data). Thurstone described the technique now known as paired comparisons as a means of accomplishing this objective[3-4]. The technique is widely used in the color-imaging domain[5-7].

The paired-comparison technique may be described as follows for n stimuli and k observers. The n stimuli are considered in pairs. Each of the k observers is required to indicate their opinion as to which of the two stimuli in each pair evokes the greater response (thus, by way of example, if the brightness of the n stimuli is being considered the observers would be expected to indicate which of a pair of stimuli is brighter). In the case where every stimulus in a set is compared with every other stimulus in the set there are simple and well-documented techniques to allow the estimation of scale values for each of the stimuli which are based on case V of Thurstone's law of Comparative Judgment. These usually involve calculating the preference ratio for each paired comparison.

Thurstone's model is not the only method for conversion of experiment proportions to scale data. There are some alternative models with a similar function to Thurstone's model such as the Gaussian model[8], the logistic Bradley-Terry model[9-13], Angular Transformation model[14] and Uniform Distribution model[2].

Hohle[15] compared the Logistic Bradley-Terry model and Thurstone's Case V using maximum likelihood methods of scale estimation. He found that the logistic Bradley-Terry model had a slight edge for experimental data with less complexity in mathematics and fewer assumptions. Jackson and Fleckenstein[16] compared the Thurstone-Mosteller model, the Scheffe method, the Morrissey-Gulliksen model and the Bradley-Terry model and summarized that the Scheffe model could provide a method for estimation and testing order of presentation; the Bradley-Terry model provided the most effective analytical procedure for a complete paired-comparison experiment; if the primary interest of research is to obtain response scales, Thurstone-Mosteller model was preferred because of easy computation; the Morrissey-Gulliksen model was helpful to reduce the size of the experiment. Later, the superiority of the logistic Bradley-Terry model was confirmed again by Handley[17] by comparing the Logistic Bradley-Terry model and Thurstone's Case V. Handley's experiment indicated that the logistic Bradley-Terry model yielded almost the same estimated scale values as the Thurstone's Case V for complete paired-comparison data with advantages of simplicity for analysis, availability for incomplete data and suitability for more statistical analyses (e.g. maximum likelihood estimate for scale parameters with confidence and hypothesis tests for uniformity and preference agreements among groups) than Thurstone's Case V. Handley's suggested that the logistic Bradley-Terry model had overwhelming advantages over Thurstone's Case V in the imaging community and should be widely used instead of Thurstone's Case V.

However, when the complete matrix of comparisons is carried out the work required becomes prohibitive for large numbers n of stimuli[18-19]. Thus, the investigator most likely conducts an incomplete paired-comparison experiment with a high number of stimuli[20]. In addition, depending on the spacing of the stimuli relative to the discriminal unit (or just-noticeable difference) it is possible that some of the preference ratios will be 1 or 0. If all observers agree that one stimulus is preferred over another there is no information available as to the relative difference between the scale values for those two stimuli, only the rank order of those scale values. These two problems can both result in an incomplete table of response-difference values and this requires alternative methods for estimating the scale values[2,21]. We refer to this as the incomplete-matrix problem and it is with methods for

solving this problem that this work is concerned. Dittrich *et al.*[20] have recently also conducted a various-scenario analysis on the missing data for paired-comparison study. Their work was based on a decision-analysis approach rather than on a statistical-modeling approach and used the Bradley-Terry model as the method of obtaining scale values. However, the work by Dittrich *et al.* was not concerned with the questions that the current study was designed to address. We note, however, that use of blocks to separate the stimuli into two of more groups is an alternative method of effectively reducing the number of paired comparisons when the number of samples is large.[12] Durbin, for example, suggested using balanced incomplete block (BIB) designs for incomplete paired comparisons[22]. The block sizes were suggested to be more than two[12]. Within each block the rank orders of objects are obtained[23-29]. To guarantee the stimuli in each block can be comparable, two or more stimuli in one block must be appear in the adjacent block[30-31] and every stimuli should appear equally often in all blocks[32]. A computer-sorting algorithm can also be used for work reduction, which can reduce the average number of comparisons and the number of comparisons from the samples far apart from each other and also produce a sorted list according to the rank order of samples[33-36]. According to Whaley's model[37], an average of no more than $n^{3/2}$ comparisons are needed. According to the procedure of heap sort an average of no more than $n\log_2^n$ comparisons are needed[38]. However, this sorting technique tends to present one sample of a pair twice in a row, which breaks the basic rule of keeping the same sample separated in time[39]. Later, Silverstein and Farrell[40] proposed a binary tree sorting method, which can provide a more accurate estimation of the original values with the disadvantage of the difficulty of dealing with hardcopy samples.

In this study we consider how to solve the problem of estimating the scale values from incomplete matrices of preference ratios. Note, however, that we only address the problem that results from all of the pairs not being considered; we do not explicitly address the problem that occurs when the preference rations are 0 or 1, We investigate the method developed by Morrissey that determines scale values according to a least-squares solution[21, 41]. Although the Morrissey method is not the only method[42-44] that can be used to solve the incomplete-matrix problem it is a method that is widely used. The substantial research questions that this study addresses are: (1) What proportion of the matrix is required in order for the method to be valid and how robust is the method as the matrix becomes more sparse? (2) What is the relationship between the sparseness of the matrix and the number of observers who take part in the paired-comparison experiment? These questions are addressed via a Monte-Carlo computational simulation using an ideal observer model.

EXPERIMENTAL

Ideal-Observer Model

According to Morrissey's method (1955) from the data from all k observers, a preference ratio (the ratio of actual to possible number of times that one stimulus is judged greater or better than the other) is computed for each pair. The preference ratio is interpreted as the area under the normal frequency function; the upper limit of integration is interpreted, both in magnitude and in sense, as the response difference between the two stimuli constituting the pair. Again, 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; 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).

An ideal observer model has been constructed to simulate the response to a paired-comparison experiment. The perceptual response P to a stimulus S is modeled by a normal distribution with mean S and standard deviation σ where σ is inversely related to the discriminatory power of the perceptual system. Figure 1 illustrates the situation for two stimuli S_2 and S_1 whose physical values are 10 and 5 respectively; the corresponding perceptual responses P_2 and P_1 are normally distributed around S_2 and S_1 each with standard deviation σ (in the example shown in Figure 1, $\sigma = 1$).

Thus, the ideal-observer model operates by generating perceptual responses for pairs of stimuli drawn from normal distributions N (S_2,) and N (S_1,). The output of the model $R_{1,2}$ is 1 if the perceptual response to P_2 is greater than P_1 and 0 if P_1 is greater than P_2 (if P_1=P_2 then we assume chance performance).

The ideal-observer model described allows us to simulate a paired-comparison experiment for n stimuli and k stochastically similar observers. In order to carry out the Monte-Carlo simulation it is necessary to define the value of the internal noise in the perceptual system . The appropriate selection of σ must be influenced by the stimuli values S_1 ... S_n. If σ is too large then adjacent stimuli will not be discriminable by the ideal observer. Similarly, if σ is too small then the preference ratio for the comparison of two adjacent stimuli will be 0 or 1. Stimuli were selected (see section on Monte Carlo simulation later) such that, when the stimuli are arranged in rank order, on average adjacent stimuli differ by 1 unit. For this work, we defined σ such that adjacent stimuli (differing by one unit) were at discrimination threshold. We assume that, for adjacent stimuli, the difference P_2- P_1 is normally distributed N(1, $\sqrt{(2\sigma^2)}$) and wish to find the value of σ for which there is a 75% chance that a value drawn from this distribution would be greater than zero (this corresponds to the ideal observer making 75% correct decisions which we define as threshold performance). Use of tables or simple computational methods reveal that = 1.048. The use of this value of σ implies that the most similar stimuli in the work will be at discrimination threshold; of course, the difference between other stimuli (which will form the majority of the paired comparisons) will be much greater than threshold.

Stimuli are randomly selected from a range that depends upon the number of stimuli n so that on average neighboring stimuli (when the stimuli are arranged in rank order) would have a difference of 1 stimulus units. Specifically, n stimuli are randomly selected from the range $[-n/2 \ldots n/2]$.

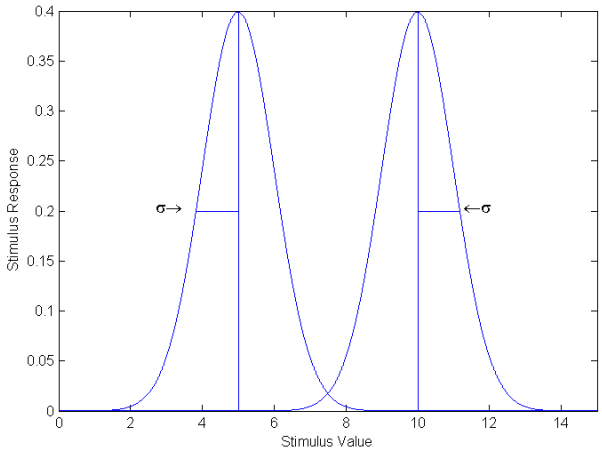

Figure 1. Schematic diagram of the ideal-observer response function. In this case, two stimuli (S_1 = 5 and S_2 = 10) are presented to the observer. The perceptual responses to the stimuli S_1 and S_2 are drawn from normal distributions $N(S_1,\sigma)$ and $N(S_2, \sigma)$ respectively where σ is the internal noise in the perceptual system (and in this case σ = 1). The probability that S_2 will elicit a stronger response than S_1 is determined by both the distance S_2-S_1 between the stimuli and the sensitivity of the system (governed by σ).

Morrissey Method

The ideal-observer model allows us to construct a matrix of preference ratios and according to Morrisey's method the application of Thurstone's law allows us to construct a matrix of response differences. For p paired comparisons we construct matrices $\mathbf{A}$ and $\mathbf{d}$ such that

$$\mathbf{Av} = \mathbf{d} \tag{1}$$

where $\mathbf{d}$ is a $(p + 1) \times 1$ matrix of response differences, $\mathbf{v}$ is a $p \times 1$ matrix of scale values and $\mathbf{A}$ is a $(p + 1) \times n$ operational matrix that defines the pair-wise comparisons that are made. For clarity, in the case where $n = 3$ Equation 2 can be written in full as

$$\begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \begin{bmatrix} d_{12} \\ d_{13} \\ d_{23} \\ 0 \end{bmatrix} \tag{2}$$

where v_i are the scale values and d_{ij} are the response differences for $i \in \{1,2,3\}$[2, 19]. The last row in matrices $\mathbf{A}$ and $\mathbf{d}$ imposes the constraint that the sum of all scale values is zero. Equation 2 can be solved using MATLAB's backslash operator (which is equivalent to Gaussian elimination) thus $\mathbf{v} = \mathbf{A}\backslash\mathbf{d}$. The advantage of Morrissey's method is that it can be solved even when every possible paired comparison is not carried out. We can therefore evaluate the effectiveness of the Morrissey method for different degrees of experimental completeness.

Monte-Carlo Simulation

A Monte-Carlo simulation of a paired-comparison experiment was conducted to explore the accuracy of the Morrissey method to estimate scale values according to the following steps:

1) Randomly select n scale values from the uniform distribution $[-n/2, n/2]$.
2) For each of the $n(n\text{-}1)/2$ pair-wise comparisons, present the two stimuli to the ideal observer model (defined by σ) and obtain the observer preference. Repeat for k observers.
3) Construct the preference ratio matrix.
4) Estimate scale values using the Morrissey method.
5) Compare the estimated scale values with the actual scale values.

In order to compare the performance of the methods the scale values (actual and estimated) were normalized to the range 0-1 and the correlation coefficient r^2 calculated for the estimated and actual normalized scale values. The simulation was repeated 1000 times, each time starting with a different random set of scale values and the mean correlation coefficient (averaged over all 1000 trials) was used as a measure of performance. The experiment was repeated for different values of n and k and also using only some of the possible paired comparisons (for a completion rate of 50%, for example, only half of the paired comparisons were used and these were randomly selected for each of the 1000 trials). The number of different conditions was 405 composed of 5 observer numbers (k = 5, 10, 15, 20, 25) × 9 stimulus conditions (n = 10, 20, 30, 40, 50, 60, 70, 80, 90) × 9 matrix conditions (completion rate = 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%).

In this work described so far the observers were all statistically identical. A modification to the main experiment was also carried out in which each observer was assigned a small bias for each of the n stimuli. In this modification instead of the observers response to the ith stimulus S_i being $N(S_i, \sigma)$, the observer's response was $N(S_i, \sigma) + B_i$ where B_i is the observer's bias for sample i. The value of B_i was selected for each observer and for each stimulus from the range [-1.04 1.04].

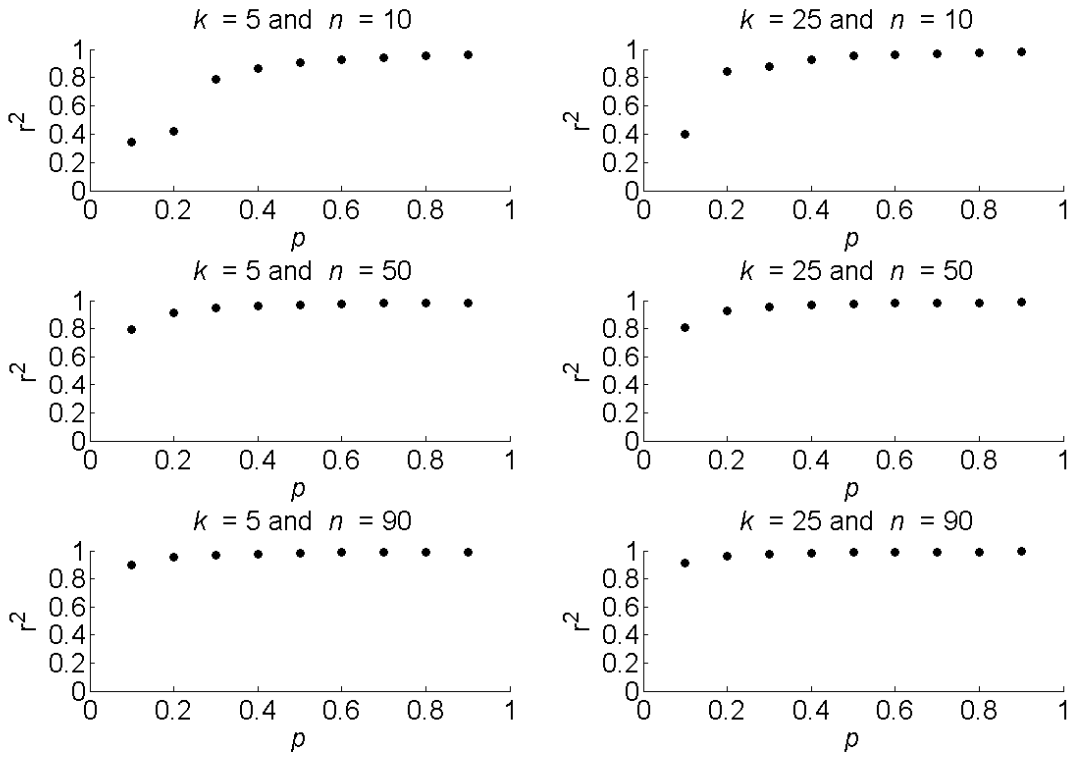

Figure 2. *The performance (mean correlation coefficient) is plotted against the proportion of the paired comparisons for various stimuli (n = 10, 50 and 90) and the number of observers (k = 5 and 25).*

This range was chosen so that the size of the bias was comparable with the noise (defined by σ) in the observer's response. The bias was selected differently for each of the 1000 simulations.

RESULTS

Figure 2 illustrates some of the data obtained from the main experiment (where observers are stochastically identical). In Figure 2, the performance (mean correlation coefficient) is plotted against the proportion of the paired comparisons and the number of observers for various values of n. These plots indicate that the correlation coefficient is relatively invariant to the proportion of paired comparisons considered except when the number of stimuli n is small. It is also apparent that as the number of stimuli increases the proportion of paired comparisons required for a given performance reduces.

In order to further analyze the data we have determined the proportion of paired comparisons required in order to yield a given performance which we have somewhat arbitrarily defined as $r^2 = 0.95$. For each condition (defined by k and n) we generate a plot of r^2 versus proportion and fit the data with a natural log function and use this to determine the proportion of comparisons required for our threshold performance ($r^2 = 0.95$). Figure 3 shows an example for the case of $k = 10$ and $n = 20$. In this plot we omitted the data obtained for very low proportions of paired comparisons. The reason for this is that when the proportion of comparisons was less than 30% the matrix solution became unstable and in some of the 1000 simulations the matrix was so ill-conditioned that no solution was possible; in these situations the performance (mean correlation coefficient) was computed from the remaining simulations where a solution was possible. This reduced the reliability of the data at very low proportions of paired comparisons and therefore, since these data typically resulted in quite small r^2 values anyway, it was decided that the logarithmic fits would only apply to 30% or greater of paired comparisons. The quality of fit in Figure 3 was typical of all 25 plots (the r^2 values for the logarithmic fits ranged from 0.9012 to 0.9794).

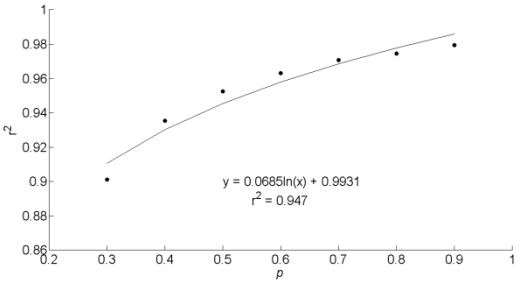

Figure 3. The performance (mean correlation coefficient) is plotted against the proportion of the paired comparisons for $n = 20$ and $k = 10$.

Figure 4 plots the threshold values for the proportion of paired comparisons for different values of n and k. This figure further emphasizes that the number of stimuli has more impact than the number of observers and that as the number of stimuli increases a lesser proportion of paired comparisons is required. For small scale experiments ($n < 20$) it is necessary to carry out more than half of the possible paired comparisons. However, for larger scale experiments as few as 20% or 30% of paired comparisons are required to achieve good performance.

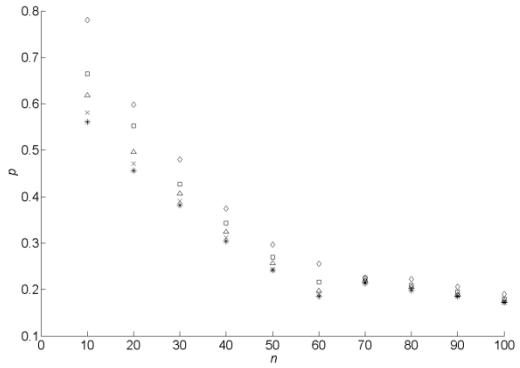

Figure 4. The threshold values for the proportion of paired comparisons for different values of n and for $k = 5$ (diamond), $k = 10$ (square), $k = 15$ (triangle), $k = 20$ (cross) and $k = 25$ (star).

However, Fig. 4 results from our simulations that involve k statistically identical observers. In any real-life experiment the observers are unlikely to be statistically identical and may exhibit personal bias for various stimuli. Therefore the complete Monte Carlo simulation was repeated but including an additional factor to represent observer bias. Figure 5 shows the final outcome of the simulation with bias. In fact, the inclusion of observer bias made relatively little difference to the final results.

Table 1 is provided as a summary of the results and as a resource for other researchers who wish to undertake incomplete paired-comparison experiments to estimate scale values. It is based on the data from the model without observer bias and indicates the threshold percent of comparisons that are required for different numbers of stimuli and observers.

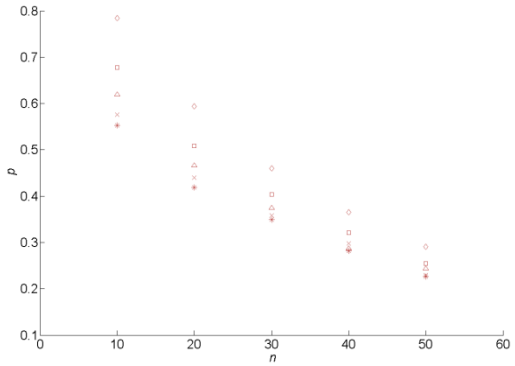

Figure 5. The threshold values with bias for the proportion of paired comparisons for different values of n and for $k = 5$ (diamond), $k = 10$ (square), $k = 15$ (triangle), $k = 20$ (cross) and $k = 25$ (star).

Table 1 Threshold values for the per cent of paired comparisons needed to achieve a criterion performance in incomplete paired comparison experiments for different numbers of stimuli (across the columns) and different numbers of observers (down the rows).

k	10 (%)	20 (%)	30 (%)	40 (%)	50 (%)	60 (%)	70 (%)	80 (%)	90 (%)	100 (%)
5	78	60	48	37	30	26	23	22	21	19
10	67	55	43	34	27	22	22	21	20	18
15	62	50	41	32	26	20	22	20	19	18
20	58	47	39	31	24	19	22	20	19	17
25	56	46	38	30	24	19	21	20	18	17

CONCLUSIONS

The design of paired-comparison experiments is an important psychophysical technique that can be applied to a wide range of problems. For large numbers 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 work through a computational simulation that incorporates an ideal-observer model (characterized by a standard deviation σ) that allows the estimation of scale values from a simulated experiment when the ideal-observer is presented with paired comparisons of stimuli of known scale values. The findings suggest 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. This has important implications for the design of psychophysical experiments and it would seem that reasonable results are obtained when 40-50% of the paired comparisons are made. Further work is underway to further explore this issue. This will include simulations of the experiment for different values of the observer variable σ and the condition of every observer evaluating a different set of stimuli.

Note Morrissey's least-square solution is used as the analysis method in this study where stimuli were

randomly selected from the complete set of possible pairs. However, there are other designs for incomplete paired-comparisons experiments. McCormick and Bachus[45] conducted a personnel-rating experiment to evaluate the reliability of partial pairings experimental design. In their experiment the cycle type of incomplete design was adopted where pairs were not randomly selected but chosen in a given pattern according to different 'rhythms'[19]. The results showed that as the number of pairs was reduced, the correlations between the results from the full matrix and partial matrix declined consistently. When the total number of personnel was 50, 35% of partial pairs could yield reliable results with correlation of around 0.95. When the total number of personnel was 30, 41% of partial pairs were needed to achieve the correlation of around 0.95. These findings are consistent with our key results in Table 1. However, the previously published results were only tested using the Personnel Comparison System, which is applied particularly in employee rating with consideration of more than one attribute of objects. Furthermore, our work gives more general and robust results that also take into account the number of observers.

For our results to be useful it is important to understand the assumptions that we made in the model. In the first experiment, without bias, the observers were stochastically identical. This means that there is no material difference between two observers participating each once and one observer participating twice (inter- and intra-observer variability were both controlled by our single parameter σ). This assumption may be reasonable when all observers would essentially make the same judgment subject to noise. An example of this might be if observers were asked to evaluate the lightness of uniform stimuli. However, it is easy to consider examples where the assumption would certainly not be reasonable. One such example would be if observers were asked to rate the beautifulness of a number of different faces. In such an example, we would expect some observers to vary quite wildly from one another in terms of their judgments. To address this limitation, the second experiment that we reported assigned a bias for each observer for each stimulus. The bias was selected randomly and to be of a similar magnitude to the noise term (σ) but for each observer was fixed for each stimulus. The implication of this is that now some observers may consistently rate one sample as stronger than another despite their underlying properties (in our model) suggesting otherwise. The introduction of this observer-bias term made relatively little difference to the results. However, it is possible that this was because the bias used was quite small. Further work is certainly required to ascertain the effect of larger observer bias and to therefore increase the applicability of our findings.

REFERENCES

1 P. Dunn-Rankin, G. A. Knezek, S. Wallace and S. Zhang, *Scaling methods*. (Lawrence Erlbaum Associates, 2004).
2 P. G. Engeldrum, *Psychometric Scaling A Toolkit Imaging Systems Development*. (Imcotek Press, Winchester, 2000).
3 L. L. Thurstone, *Psychological Review* **34** (4), 273-286 (1927).
4 L. L. Thurstone, *American Journal of Psychology* **38**, 368-389 (1927).
5 Y. Iwasaki, T. Yamaguchi, T. Watanabe and Y. Hoshino, presented at the NIP & Digital Fabrication Conference, Fort Lauderdale, Florida, USA., 2001 (unpublished).
6 C. Cui, presented at the Color and Imaging Conference Scottsdale, Arizona, USA. , 2000 (unpublished).
7 P. Zolliker and Z. Barańczuk, presented at the Conference on Colour in Graphics, Imaging, and Vision., 2010 (unpublished).
8 H. Whaley and G. K. Lee, *Atmospheric Environment* 16 (4), 871-871 (1982).
9 T. Augustin, *Journal of Mathematical Psychology* 49 (1), 70-79 (2005).
10 D. Causeur and F. Husson, *Journal of Statistical Planning and Inference* 135 (2), 245-259 (2005).
11 U. Grasshoff and R. Schwabe, *Statistical Methods and Applications* 17 (3), 275-289 (2008).
12 R. A. Bradley and M. E. Terry, *Biometrika* 39 (3/4), 324-345 (1952).
13 R. D. Luce, *Individual choice behaviours: a theoretical analysis*. (J. Wiley., New York, 1959).
14 L. V. Jones and R. D. Bock, 1957.
15 R. H. Hohle, *Journal of Mathematical Psychology* 3 (1), 174-183 (1966).
16 J. E. Jackson and M. Fleckenstein, *Biometrics* 13 (1), 51-64 (1957).
17 J. C. Handley, presented at the Pics 2001: Image Processing, Image Quality, Image Capture, Systems Conference, Proceedings, 2001 (unpublished).
18 G. A. Gescheider, *Psychophysics: The Fundamentals*, 3rd ed. (Lawrence Erlbraum Associates, 1997).
19 H. A. David, *The Method of Paired Comparisons*, 2nd ed. (Oxford University Press, New York, 1988).
20 R. Dittrich, B. Francis, R. Hatzinger and W. Katzenbeisser, *Statistical Modelling* 12 (2), 117-143 (2012).
21 J. H. Morrissey, *Journal of the Optical Society of America* 45 (5), 373-378 (1955).
22 J. Durbin, *British Journal of Statistical Psychology* 4 (2), 85-90 (1951).
23 J. E. Oliver, *Journal of Applied Psychology* 37 (2), 129-130 (1953).
24 M. A. Olson, *Journal of Educational Measurement* 15 (1), 49-52 (1978).
25 W. W. Rambo, *Psychological Reports* 5 (2), 341-344 (1959).
26 W. W. Rambo, *Journal of Applied Psychology* 43 (6), 379-381 (1959).
27 D. O. Shaw and H. G. Osburn, *Journal of Applied Psychology* 54 (6), 526-& (1970).
28 D.M. Shoemake, *Journal of Educational Measurement* 8 (4), 279-282 (1971).
29 S. L. Witryol, *Journal of Applied Psychology* 38 (1), 31-37 (1954).
30 S. S. Stevens and J. Volkmann, *The American Journal of Psychology* 53 (3), 329-353 (1940).

31 N. Burningham and Y. Ng, presented at the Is&Ts Eighth International Congress on Advances in Non-Impact Printing Technologies, 1992 (unpublished).
32 M. G. Kendall, *Biometrics* 11 (1), 43-62 (1955).
33 M. Brisson, M. Dewson and C. Whissell, *Perceptual and Motor Skills* 55 (3), 745-746 (1982).
34 M. H. Chignell and B. W. Patty, *Psychological Bulletin* 101 (2), 304-311 (1987).
35 L. Knowles, B. C. Jarvis and M. W. Starr, *Behavior Research Methods Instruments & Computers* 22 (3), 335-336 (1990).
36 R. F. Stevens, *International Journal of Man-Machine Studies* 23 (5), 563-585 (1985).
37 C. P. Whaley, *Behavior Research Methods & Instrumentation* 11 (2), 147-150 (1979).
38 W. H. Press, B. P. Flannery, S. A. Teukolsky and W. T. Vetterling, *Numerical recipes : the art of scientific computing*, 3rd ed. (Cambridge University Press, Cambridge, 2007).
39 W. S. Torgerson, *Theory and Methods of Scaling*, 1st ed. (Wiley, New York, 1958).
40 D. A. Silverstein and J. E. Farrell, *J. Electron. Imaging* 10 (2), 394-398 (2001).
41 G. J. Borse, *Numerical methods with MATLAB.* (PWS Publishing Company, Boston, 1997).
42 H. Gulliksen, *Pyschometrika* 21, 125 (1956).
43 J. A. Clark, *Educational and Psychological Measurement* 37, 603-611 (1977).
44 B. Golany and M. Kress, *European Journal of Operational Research* 69, 210-220 (1993).
45 E. J. McCormick and W. K. Roberts, *Journal of Applied Psychology* 36 (3), 188-192 (1952).