Comparative Analysis of Bradley-Terry and Thurstone-Mosteller Paired Comparison Models for Image Quality Assessment

John C. Handley
Xerox Corporation, Digital Imaging Technology Center
Webster, New York

Abstract

In image quality assessment, preference for various image processing algorithms or treatments is often determined using paired comparisons. In this experimental design, pairs of images processed by different algorithms or "treatments" are presented to a judge. The preferred treatment is selected and a tally is kept of the number of times each treatment is preferred to another. It is possible to estimate an interval scale for treatments in a hypothetical psychological space using this method.

There are two dominate paired comparison statistical models: Thurstone-Mosteller Case V (TM) (corresponding to Thurstone's Law of Comparative Judgment, Case V) and Bradley-Terry (BT). Although TM is used almost exclusively in the imaging literature, the BT formulation is more mathematically developed. Owing to its parsimony, it provides tractable maximum-likelihood estimators for scales, simultaneous confidence intervals and hypothesis tests for model fit, uniformity, and differences among populations of judges. In practice, TM and BT yield nearly identical scale estimates for complete data. In some experimental designs, many treatments are compared. Owing to the large number of possible treatment pairs, not every comparison is made, leading to an incomplete matrix of preference counts. Unlike TM, BT model applies directly to incomplete data under mild restrictions

We compare and critique TM and BT models. Statistical analyses, many not available under TM, are demonstrated. An argument is made that BT offers overwhelming advantages to the imaging community and should be used instead of TM.

Introduction

This paper compares two well-known paired comparison models: the Thurstone-Mosteller (TM) model (by which we mean Thurstone's Law of Comparative Judgment, Case V) and the Bradley-Terry (BT) model. (Mosteller's name is included in TM due to his work on the statistical analysis of Thurstone's model). We argue here that BT model should be used in place of TM because presently the former is more developed mathematically than the latter. In particular, easy formulas exist for maximum likelihood estimates (mle) of scale parameters. The asymptotic theory of mle's yields estimators for confidence regions and test statistics based on likelihood ratios for hypothesis testing. TM is privileged within the imaging community ostensibly owing to its origins in psychophysics. Yet it is universally acknowledged that TM and BT yield similar scale estimates. The theory (and software) for generalized linear models can produce mle's yet BT, with its roots in experimental design and consumer choice modeling, offers numerically easier statistical procedures. We present no new research although we do show an alternative analysis to previously published data. Our intent is to provide the imaging community with a general context for paired comparisons, compare and contrast the two models, and demonstrate the advantages of BT.

The Linear Model

TM and BT models are both linear models of paired comparisons. In such models, probabilities of preference can be mapped to scales. Formally (following David, 1988[4]), let V_i and V_j represent "merits" of objects A_i and A_j, respectively. In a psychophysics setting, the V_i might represent sensation magnitudes on a scale. We represent the observed merit of object A_i by random variable X_i owing to observation-to-observation variation. A linear model takes the form

$$P(X_i > X_j) \equiv \pi_{ij} = H(V_i - V_j) \tag{1}$$

where H is a monotonic, increasing function such that $H(-\infty) = 0$, $H(+\infty) = 1$, and $H(-x) = 1 - H(x)$. There are obviously an infinite number of choices for function *H*, the two of concern here are the Thurstone-Mosteller model where *H* is the normal cumulative distribution function with zero mean and the Bradley-Terry model where

$$H(x) = \frac{1}{2}\left[1 + \tanh(x/2)\right] \tag{2}$$

The task is to produce estimates v_i of V_i, $i = 1,\ldots,m$. If the function H has additional parameters, we need to estimate those as well. Assume without loss of generality $\sum_{i=1}^{m} V_i = 0$ and define $\delta_{ij} = V_i - V_j$. Estimation proceeds by tallying α_{ij}, the number of times object A_i is preferred to object A_j after n_{ij} comparisons. A sample estimate of π_{ij} is $p_{ij} = \alpha_{ij} / n_{ij}$. We define $H(d_{ij}) = p_{ij}$ and compute merit or scale estimates v_i by $d_{ij} = v_i - v_j$, $i \neq j$, $i, j = 1,\ldots,m$. It can be shown that a least squares estimate of V_i is

$$v_i = \frac{1}{m}\sum_{i \neq j}^{m} d_{ij} \tag{3}$$

This estimate holds regardless of *H* and is the usual method for Thurstone's Case V model.

Assume that each pair is observed a fixed (but possibly unequal) number of times. That is, the sums n_{ij} are fixed and the tallies α_{ij} are binomial random variables:

$$P(\alpha_{ij}) = \binom{n_{ij}}{\alpha_{ij}} \pi_{ij}^{\alpha_{ij}} (1-\pi_{ij})^{n_{ij}-\alpha_{ij}}, \alpha_{ij} = 0,1,\ldots,n_{ij} \tag{4}$$

Owing to independence, the likelihood function is

$$\begin{aligned} L(\boldsymbol{\alpha}) &= \prod_{i<j} P(\alpha_{ij}) \\ &= \prod_{i<j} \binom{n_{ij}}{\alpha_{ij}} H(V_i - V_j)^{\alpha_{ij}} \left[1 - H(V_i - V_j)\right]^{n_{ij}-\alpha_{ij}} \end{aligned} \tag{5}$$

where $\boldsymbol{\alpha} = [\alpha_{ij}]$, the matrix of preference counts.

The Thurstone-Mosteller Model

The most general Thurstonian model on *m* stimuli posits a multivariate distribution on $(X_1,\ldots,X_m)$. In paired comparisons, one observes incomplete rankings where stimuli are presented two at a time. Pair-wise choice probabilities take the form

$$\begin{aligned} P(X_i > X_j) &= \frac{1}{\sqrt{2\pi(\sigma_i^2 + \sigma_j^2 - 2\sigma_{ij})}} \times \\ & \int_{-(\mu_i-\mu_j)/(\sigma_i^2+\sigma_j^2-2\sigma_{ij})^{1/2}}^{\infty} \exp(-y^2/2)dy \end{aligned} \tag{6}$$

For a scaling interpretation, means are considered ordered along a continuum in a psychological space. As discussed elsewhere (e.g., Engledrum[5] or Torgerson,[9]] the full-blown Thurstone model has too many parameters (means, variances, and covariances), so simplifying assumptions are applied. Perhaps the most-used model in paired comparisons in Thurstone's Case V, where X_i's are assumed independent and identically distributed save for location parameters μ_i, $i = 1,\ldots,m$ ($\mu_i = v_i$, $i = 1,\ldots,m$ in the linear model discussion):

$$P(X_i > X_j) = \frac{1}{\sqrt{2\pi}} \int_{-(\mu_i-\mu_j)}^{\infty} \exp(-y^2/2)dy \tag{7}$$

In this case, one usually computes least squares estimates $\hat{\mu}_i$ using Eq. 3. Inferences regarding $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1,\ldots,\hat{\mu}_m)$ are difficult to obtain owing to its unknown (asymptotic) distribution.

A likelihood function based on comparisons matrix $\boldsymbol{\alpha}$ is

$$\begin{aligned} L(\boldsymbol{\alpha};\boldsymbol{\mu}) &= \prod_{i<j} P(\alpha_{ij}) \\ &= \prod_{i<j} \binom{n_{ij}}{\alpha_{ij}} \left[\Phi(\mu_i-\mu_j)\right]^{\alpha_{ij}} \left[1-\Phi(\mu_i-\mu_j)\right]^{n_{ij}-\alpha_{ij}} \end{aligned} \tag{8}$$

The log of this likelihood function can be optimized numerically.

The Bradley-Terry Model

One can rewrite Eq. 2 as

$$\log(\pi_{ij}/(1-\pi_{ij})) = V_i - V_j, \tag{9}$$

That is, the scale or merit differences obey a logistic model (instead of a probit model in the Thurstonian case). This model can be simplified to *m*-1 parameters by

$$\pi_{ij} = \frac{\pi_i}{\pi_i + \pi_j}, i \neq j, \tag{10}$$

where

$$\pi_i > 0 \text{ and } \sum_{i=1}^{m} \pi_i = 1$$

so that Eq. 9 takes the form $\log \pi_i - \log \pi_j = V_i - V_j$. This is the Bradley-Terry model of paired comparisons. One can write the model in a form similar to Eq. 7:

$$P(X_i > X_j) = \frac{1}{4} \int_{-(\log\pi_i - \log\pi_j)}^{\infty} \operatorname{sech}^2(y/2)dy, \tag{11}$$

and $V_i = \log \pi_i$ provide scale parameters. Owing to Eq. 10, the likelihood function, Eq. 5, has a simple form in terms of $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_m)$ and can be solved iteratively:

$$p_i = \frac{a_i}{\sum_{i \neq j} n_{ij}(p_i + p_j)^{-1}} \tag{12}$$

where $a_i = \sum_{i<j} \alpha_{ij}$,

the total number of comparisons preferring A_i. A sufficient condition for a maximum likelihood is that each partition of the objects into two nonempty subsets such that some object in the second set has been preferred to at least once to some object in the first set.[6] David (1988) points out that if this condition is violated, it means one of two things: 1) there exists subsets S and T of objects such that no object in S is compared to object in T; or, 2) there exists subsets S and T such that every comparison of objects between them favors objects in S.[4] These conditions can often be detected by inspecting the comparisons matrix $\boldsymbol{\alpha}$.

BT (essentially Eq. 10) can be developed into a general distance model on ranked data. Mallows[7] invoked the so-called Babington-Smith transitivity model (which allows only paired comparisons that produce a complete ranking on *m* objects) on BT to produce the Mallows θ model. This is discussed in Marden.[8]

The remainder of this section follows Bradley.[2] In addition to MLE for scale parameters, BT also provides a means to test whether the data are statistically different from uniform. To test the hypothesis

$$H_0: \ \pi_1 = \cdots = \pi_m = 1/m \tag{13}$$

against the alternative

$$H_a: \ \pi_i \neq \pi_j \text{ for some } i, j, \ i \neq j, \ i, j = 1, \ldots, m \tag{14}$$

use the test statistic

$$\begin{aligned} T_U &= 2N \log 2 - 2B_1 \\ B_1 &= \sum_{i<j} n_{ij} \log(p_i + p_j) - \sum_i a_i \log p_i \end{aligned} \tag{15}$$

which is distributed approximately chi-squared with *t*-1 degrees of freedom (df) for large n_{ij} under H_0.

Sometimes we wish to test whether there are differences among groups of responses. In the example below, we test whether there is a difference between experts and nonexperts. Let each of *g* groups have its own set of *m* parameters indexed the following way: $\pi_i^u, \ i = 1, \ldots, m, \ u = 1, \ldots, g$. To test

$$H_0: \pi_i^u = \pi_i, \ i = 1, \ldots, m, \ u = 1, \ldots, g \tag{16}$$

versus the alternative

$$H_a: \pi_i^u \neq \pi_i \text{ for some } i \text{ and } u, \tag{17}$$

use the the test statistic

$$T_G = 2\left(B_1 - \sum_{u=1}^{g} B_{1u} \right) \tag{18}$$

where B_1 is computed as above using data pooled over groups and B_{1u} is computed for each group. Under H_0 for large n_{iju} this test statistic has an approximate chi-squared distribution with (*g*-1)(*t*-1) degrees of freedom.

Bradley also provides a confidence region for the vector parameter

$$\pi = (\pi_1, \ldots, \pi_m).$$

Approximate (1 - α) 100% confidence intervals for the location parameters of interest are

$$\left(\log p_i - z_{\alpha/2}\sqrt{\hat{\sigma}_{ii}/N}/p_i, \log p_i + z_{\alpha/2}\sqrt{\hat{\sigma}_{ii}/N}/p_i \right) \tag{19}$$

$$i = 1,\ldots,m, \text{ where } N = \sum_{i<j} n_{ij}$$

is the total number of comparisons, the p_i are the mle's, $\hat{\sigma}_{ii}$ is the *i*th diagonal element of the (*m* + 1) by (*m* + 1) matrix

$$\hat{\Sigma} = \begin{bmatrix} \hat{\Lambda} & \mathbf{1} \\ \mathbf{1}' & 0 \end{bmatrix}^{-1} \tag{20}$$

where $\hat{\Lambda} = [\hat{\lambda}_{ij}]$,

$$\begin{aligned} \hat{\lambda}_{ii} &= \frac{1}{p_i} \sum_{j \neq i} p_j n_{ij} / [N(p_i + p_j)^2], \ i = 1, \ldots, m \\ \hat{\lambda}_{ij} &= -n_{ij} / [N(p_i + p_j)^2], \ i \neq j, \ i, j = 1, \ldots, m. \end{aligned} \tag{21}$$

With the aid of a matrix inversion routine, these statistics are easily coded into C.

Analysis Example

We analyze a data set using BT model to demonstrate its advantages over TM. The experiment is discussed in detail in [1]. Four gamut-mapping algorithms were evaluated in two ways. In the first part, subjects chose the better rendition from a pair of prints. In the second, subjects chose the better reproduction of reference prints. Tables 1 and 2 contain the comparison data.

Table 1. Comparisons matrix for "preference" experiment.

	1	2	3	4
1	-	26	28	22
2	64	-	46	34
3	62	44	-	64
4	68	56	64	-

Table 2. Comparisons matrix for "reproduction" experiment.

	1	2	3	4
1	-	46	29	48
2	44	-	34	43
3	61	56	-	50
4	42	47	40	-

Each of eighteen judges viewed five images and each print was an image/algorithm combination. Judges were partitioned into two classes based on experience: experts (11) and non-experts (7). From the data, we wish to establish for each task, whether preferences exist, and if so, a estimate a preference scale. Further, we wish to access whether differences exist between experts and non-experts.

Preference Data

Using procedures summarized above we perform a hypothesis test to determine whether the data are statistically significant from uniform: T_U = 74.01 with 3 df. The 95% chi-square cutoff is 7.82, so we conclude the data are nonuniform. The estimated scale:

$$(\log(p_i),\ i = 1,\ldots,4) \text{ is } (-2.22, -1.39, -1.53, -0.86).$$

The data can be grouped into comparisons made by expert and nonexperts. For expert data, the estimated scale is: (-2.25, -1.43, -1.58, -0.80) and the test statistics for uniformity it T_U = 50.1 with 3 df, which is significant at 95%. For nonexpert data, the estimated scale is: (-2.17,-1.34, -1.46, -0.94) and T_U = 24.7 with 3 df, also significant at 95%. The scales for experts and nonexperts appear to be similar. We can do a hypothesis test to compare these two populations for preference data. The test statistic for uniformity of these two groups is T_G = 0.75 with 3 df, which is not significant at 95% and therefore we conclude there is no statistical difference in the preferences of these two populations. Estimated scales and confidence intervals are shown in Figures 1 through 3.

In summary, algorithm 4 is preferred by experts and preferred weakly by nonexperts for the experiment in which subjects were asked which rendition they preferred.

Reproduction Data

The estimated scale for the entire reproduction data set is (-1.54, -1.57, -1.05, -1.48) . The test statistic T_U = 15.7 with 3 df, N = 540, which significant at 95%. We therefore conclude that the data is statistically different from a pure random sample from a uniform distribution and that the data show a preference structure. For the expert responses among the reproduction data, the estimated scale is: (-1.52, -1.64, -0.92, -1.67). The test for uniformity: T_U = 19.3 with 3 df, significant at 95%, from which we conclude that the data for experts show a preference structure.

For nonexperts, the estimated scale is: (-1.62, -1.5, -1.27, -1.21) and T_U = 3.76 with 3 df, which is not significant at 95%. We conclude that the data are not statistically different from uniform (there is a 28.8% chance we would have gotten this test statistic value were the data from a uniform distribution).

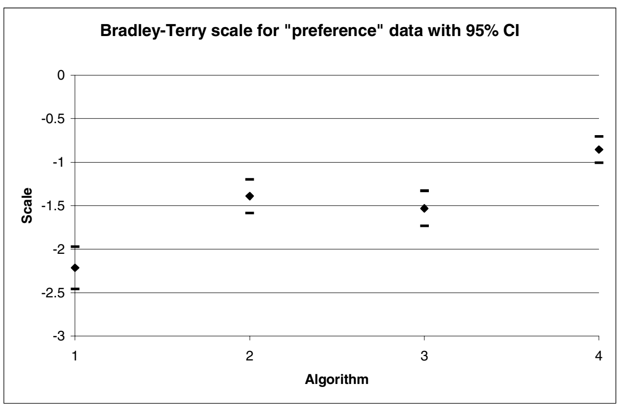

Figure 1. Scale for "preference" data.

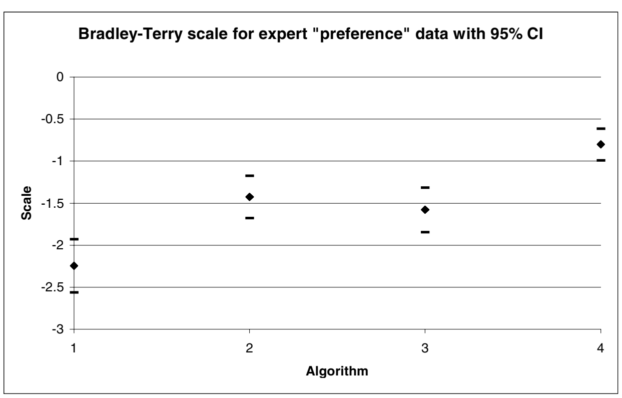

Figure 2. Scale for expert "preference" data.

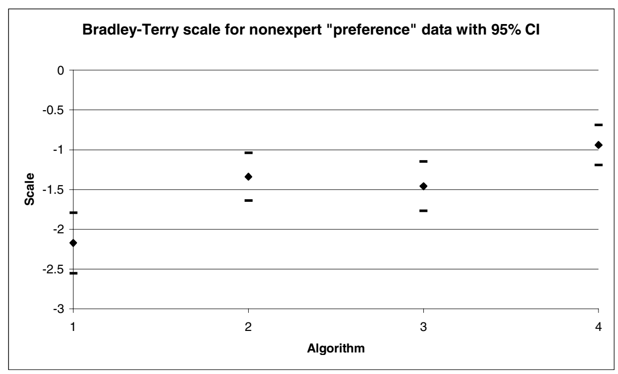

Figure 3. Scale for nonexpert "preference" data

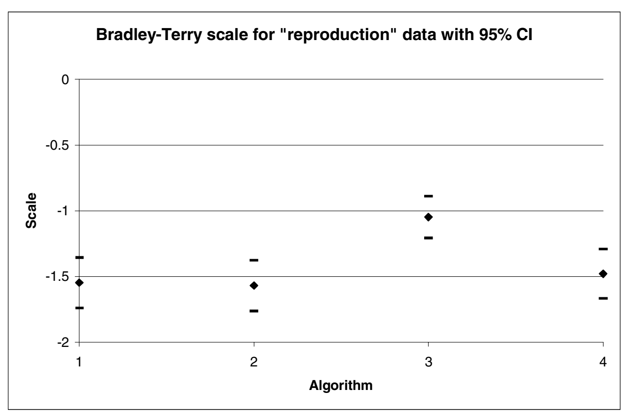

Figure 4. Scales for"reproduction" data.

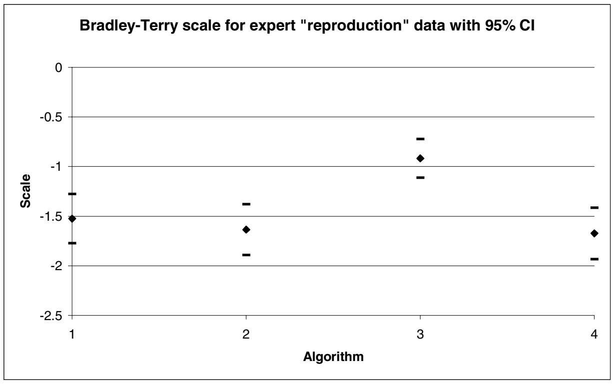

Figure 5. Scale for expert "reproduction" data.

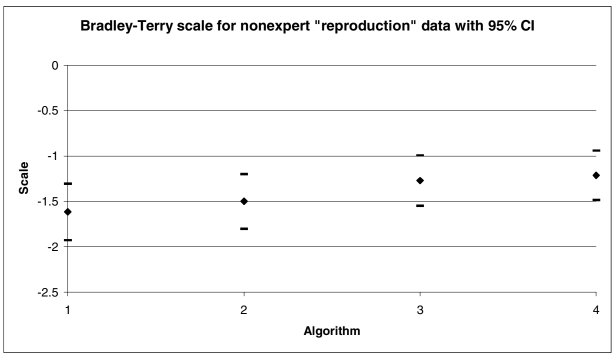

Figure 6.Scale for nonexpert "reproduction" data.

To compare experts and nonexperts, the test for uniformity of these two groups: $T_G = 7.4$ with 3 df, which is not signficant at 95% (but it is significant at 94%; that is, there is a 6% probability that this test statistic value would be obtained under uniformity). Thus algorithm 3 is preferred by experts for the reproduction experiment in which viewers were asked to judge which algorithm produced a closer match to an original. Nonexpert judgments are not statstically different from uniformly random preferences.

Summary

Owing to its simplicity, BT is much more developed analytically than TM (Case V). Many statistical procedures are available and easily implemented. We have demonstrated a few: mle's for scale parameters with confidence intervals (and regions), hypothesis tests for uniformity, and hypothesis tests for preference agreements among groups. In the main, both models can be cast into the framework of generalized linear models and numeric techniques used to perform similar analyses.[3] Should one wish to model interactions between pairs of stimuli and dispersion variations, alternatives to TM are available.[10] In the modern setting, we are no longer restricted to least-squares solutions to TM models. We can explore many general models using modern statistical theory and software. But for the bulk of our work (Thurstone's Case V), BT provides powerful analyses easily implemented in a few tens of lines of C-code.

References

1. R. Balasubramanian et al., Gamut mapping to preserve spatial luminance variations, *Proc. 8th Color Imaging Conference*, pp. 122-128 (2000).
2. R. A. Bradley, Paired comparisons: Some basic procedures and examples, *Handbook of Statistics*, *Vol. 4*, P. R. Krishnaiah and P. K. Sen, eds., Elsevier Science Publishers, pp. 299-326 (1984)
3. D. E. Critchlow and M. A. Fligner, Paired comparison, triple comparison, and ranking experiments as generalized linear models, and their implementation in GLIM, *Psychometrika*, 56(3), pp. 517-533 (1991).
4. H. A. David, *The Method of Paired Comparisons* (2nd ed.), New York, Oxford University Press (1988).
5. P. G. Engledrum, *Psychometric Scaling: A Toolkit for Imaging Systems Development*, Winchester, Imoteck Press (2000).
6. L. R. Ford, Jr., Solution of a ranking problem from binary comparisons, *American Mathematical Monthly*, 64, pp. 28-33, (1957).
7. C. L. Mallows, Non-null ranking models: I, *Biometrika*, 44, pp. 114-130 (1957).
8. J. I. Marden, *Analyzing and Modeling Rank Data*, New York, Chapman & Hall (1995)
9. W. S. Torgerson, *Theory and Methods of Scaling*, New York, Wiley (1958).
10. M. Zhou and C. Cui, New mathematical model for the law of comparative judgment, *Proc. IS&T's 16th International Congress on Digital Printing Technologies*,Vancouver, BC, Canada, pp. 383-387 (2000).

Biography

John C. Handley is a Member of the Research and Technical Staff at Xerox Corporation. He received a B.S. and M.S. in mathematics from The Ohio State University and Ph.D. in Imaging Science at Rochester Institute of Technology. He has published papers in document image processing, OCR, applied statistics, and random sets. He is currently co-editing a special section on statistical issues in psychometric assessment of image quality for the Journal of Electronic Imaging.