Statistical analysis of subjective preferences for video enhancement

Russell L. Woods*, PremNandhini Satgunam, P. Matthew Bronstad, Eli Peli
Schepens Eye Research Institute, Department of Ophthalmology,
Harvard Medical School, Boston, MA, USA

ABSTRACT

Traditional Thurstone scaling (1927) constructs a perceptual scale from pairwise comparisons without providing statistical inferences. We show that subjective preferences for moving video using pairwise comparisons can be analyzed to construct a perceptual scale and provide the statistical significance of preference differences. Two statistical methods (binary logistic regression and linear regression) are described. Data sets from two studies are used to demonstrate the perceptual scale construction from the traditional Thurstone method and from the described statistical methods. Both the studies showed videos on two side-by-side TVs. Four enhancement levels (Off, Low, Medium and High) were applied to the videos using a commercial device. Subjects made pairwise comparisons to indicate their preference of one video over another. The perceptual scales constructed from the three methods were comparable, except when there were cells missing from the preference matrix. Binary logistic regression easily permitted modeling of additional factors, such as side bias. Video quality can be systematically assessed using pairwise comparisons and statistical methods that permits construction of a perceptual scale and provide statistical significance for the compared levels.

Keywords: logistic regression, subjective preference, pairwise comparisons, video enhancement

1. INTRODUCTION

Measuring the perceived image quality of moving video is harder than for static images, due to the fleeting and variable nature of moving video [1]. Subjective image-quality preferences can be measured by observers indicating their preference for one image over another image. Such pairwise comparisons of image quality can be analyzed using Thurstone scaling [2-4].

In 1927, Thurstone first published his law of comparative judgments for application towards qualitative measurement of psychological values such as the perceived seriousness of a crime [5, 6], opinion polls and political voting [7]. Now, Thurstone scaling is widely used in areas such as applied psychology, marketing, food tasting and advertising research [8-11]. According to Thurstone's model, subjective sensation or response for a physical stimulus varies as a random variable that has a normal distribution with mean sensation, S_i, and standard deviation, σ_i. Two distinct stimuli will generate different response distributions. The perceived difference between the stimuli is the distance between the means, S_j and S_k , and can be represented as a normalized distance, Z_{jk}. There are five cases explained in Thurstone's analysis, with Case V, the simplest, being the most widely used [11, 12]. Assumptions to be satisfied for Case V include that σ_i should be equal and uncorrelated among the compared pairs of stimuli. Thurstone's model has further been modified by several others subsequently [13, 14]. To rank and construct the perceptual scale for the compared stimuli (stimuli; e.g. levels of image quality), the preference frequency (or proportion) of one stimulus over another is used. The inverse of the standard cumulative normal distribution function (Φ^{-1}) is then computed for each of the compared stimulus pairs.

The calculation for preference, $P(X_j \succ X_k)$, of one stimulus, X_j, over another, X_k, is:

$$P(X_j \succ X_k) = Z_{j\succ k} = \Phi^{-1}(S_j - S_k) = \Phi^{-1}\left[\frac{f_{j\succ k}}{f_{j\succ k} + f_{k\succ j}}\right], \quad (1)$$

where, $f_{j\succ k}$ and $f_{k\succ j}$ are the frequencies at which the stimuli X_j and X_k were preferred when compared against each other. The preference is expressed as a normalized distance (Z-score), $Z_{j\succ k}$, for each of the compared pairs. As illustrated in

* russell.woods@schepens.harvard.edu; telephone 1 617 912 2589

Human Vision and Electronic Imaging XV, edited by Bernice E. Rogowitz, Thrasyvoulos N. Pappas,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 7527, 75270E · © 2010 SPIE-IS&T · CCC code:
0277-786X/10/$18 · doi: 10.1117/12.843858

Table 1 for the comparison of three stimuli, the marginal sums are averaged (divided by the number of stimuli, n) and the resultant value, Z_i,

$$Z_i = \sum Z_{i \succ k} / n \tag{2}$$

is the arbitrary location on the perceptual scale of S_i. By convention, the least preferred stimulus is scaled to zero with the rest of the stimuli shifted accordingly and ordered in increasing magnitude to indicate the preference of one stimulus over another stimulus. To facilitate comparisons between methods, we followed the common practice of normalizing the perceptual scale to have a zero-to-one range, as seen in Figure 1a.

Table 1. Components of a Thurstone matrix used to construct Thurstone's perceptual scale for three stimuli (a, b & c) that are compared against each other.

	a	b	c	Z_i
a		$Z_{a \succ b}$	$Z_{a \succ c}$	Z_a
b	$Z_{b \succ a}$		$Z_{b \succ c}$	Z_b
c	$Z_{c \succ a}$	$Z_{c \succ b}$		Z_c

A major limitation of Thurstone scaling is that the statistical significance of differences between the stimuli on the perceptual scale is not determined. Recent papers [4, 12] have provided inferential statistical methods that produce an outcome similar to Thurstone scaling. Lipovetsky and Conklin [12] demonstrated that binary logistic regression can be used to construct a Thurstone-like perceptual scale and provide the statistical significance of differences between preferences directly from the regression outcomes from typical statistical packages. The model is developed without the construction of a Thurstone matrix as described above. Instead, each paired comparison is entered in a matrix (logistic regression table), where for each comparison, i (row), the preferred stimulus (column) is allocated +1 and the non-preferred stimulus is allocated -1:

$$\text{if } X_{ij} \succ X_{ik} \quad \text{then } X_{ij} = +1 \quad \text{and} \quad X_{ik} = -1 \tag{3}$$

All other stimuli are allocated zeros. The identity vector (dependent variable, column), e_i, is randomly assigned to a value of 0 or 1. When $e_i = 0$ ("false"), the signs of the responses for that comparison are reversed, such that:

$$\text{if } X_{ij} \succ X_{ik} \quad \text{then } X_{ij} = -1 \quad \text{and} \quad X_{ik} = +1 \tag{4}$$

The outcome of the analysis is independent of the proportion of e_i=0 comparisons, so long as $0 < p(e_i=0) < 1$, where $p(e_i=0)$ is the proportion of comparisons with e_i=0. Such preference table is illustrated in Table 2. Common statistical software (e.g. SPSS, SAS) can be used to perform the binary logistic regression with the stimuli entered as factors. The model is constructed without an intercept (i.e. forced through the origin). One of the stimuli is kept constant (not entered in the analysis). The computed coefficients are the relative preferences of the stimuli with the excluded stimulus having a relative preference of zero. The statistical significance for stimulus j reported in each analysis is for the difference between the missing stimulus, k, and stimulus j: Z_{jk}. When n stimuli are compared, to obtain the statistical significance for all $n(n\text{-}1)/2$ comparisons, the analysis is performed n-1 times, with a different stimulus kept constant (missing) from each analysis. In the case of three stimuli, there are three comparisons and the analysis is conducted twice.

Table 2. Illustration of the construction of a logistic regression table for the binary logistic regression analysis of three stimuli (a, b & c) that are compared against each other by m subjects for a total of t trials (comparisons). The preferred stimulus is shown in bold in each row.

Trial (i)	Subject	a	b	c	Identity Vector (e)
1	1	**1**	0	-1	1
2	1	0	1	**-1**	0
⋮	⋮	⋮	⋮	⋮	⋮
i	1	**1**	-1	0	1
i+1	2	**-1**	0	1	0
⋮	⋮	⋮	⋮	⋮	⋮
t	m	0	**1**	-1	1

Using sample data from 319 subjects, each comparing five of 15 flavors, Lipovetsky and Conklin [12] showed that, when the relative preferences were normalized to the range of 0 to 1, the relative preferences obtained using binary logistic regression were very similar to those obtained with normalized Thurstone scaling.

Rajae-Joordens and Engel [4] demonstrated that linear regression (rather than binary logistic regression) can be used to construct a Thurstone-like scale for image quality comparisons. The model is constructed from a matrix (linear regression table) as illustrated in Table 3 for paired comparisons of three stimuli. For each paired comparison of stimuli X_j and X_k , X_j is allocated +1 and X_k is allocated -1 and other stimuli, X_i , are allocated 0. These are the factors in the model. From, $p_{j \succ k}$, the proportion that stimuli X_j was preferred over X_k , the inverse of the cumulative frequency $Z_{j \succ k}$ is calculated. This is the dependent variable in the model. The linear regression model for n compared stimuli has the form:

$$\Phi^{-1}(p_{j \succ k}) = Z_{j \succ k} = \beta_1.X_1 + \ldots + \beta_{n-1}.X_{n-1} \quad (5)$$

Which for three compared stimuli, a, b and c, has the form:

$$\Phi^{-1}(p_{j \succ k}) = Z_{j \succ k} = \beta_a.X_a + \beta_b.X_b \quad (6)$$

Like the binary logistic regression model, one of the stimuli is not included in the linear regression. Common statistical software can be used to perform the linear regression. The computed coefficients, β_i , are the relative preferences for stimuli, with the excluded stimulus having a relative preference of zero. A disadvantage of the linear regression approach is that additional steps are required to determine the statistical significance of preference differences [4]. Rajae-Joordens and Engel [4] also demonstrated that the effect of other factors (e.g. gender of observer) on preferences can be included in the model, but again, additional steps are required to determine the statistical significance of such factors.

Table 3. Calculations for constructing the linear regression table used in the linear regression model for three stimuli (a, b & c) that are compared against each other.

a	b	c	$p_{j \succ k}$	$Z_{j \succ k}$
1	-1	0	$p_{a \succ b} = f_{a \succ b} / (f_{a \succ b} + f_{b \succ a})$	$Z_{a \succ b}$
1	0	-1	$p_{a \succ c} = f_{a \succ c} / (f_{a \succ c} + f_{c \succ a})$	$Z_{a \succ c}$
0	1	-1	$p_{b \succ c} = f_{b \succ c} / (f_{b \succ c} + f_{c \succ b})$	$Z_{b \succ c}$

Using sample data from 12 subjects that ranked the picture quality of four displays, Rajae-Joordens and Engel [4] showed, that the relative preferences obtained using linear regression were very similar to those obtained with Thurstone scaling. Using simple modeling, they demonstrated that the approach was robust even with relatively small sample sizes.

These three methods for estimating relative preferences all can suffer from the Hauck–Donner effect [15] when there are extreme proportions in one or more cells of the Thurstone matrix (Table 1). This problem occurs when there is complete (or in small data sets, nearly complete) separation between responses for stimuli: for example, all $p_{j \succ k} \approx 1$. In this case, $Z_{j \succ k}$ becomes very large, and the relative preference cannot be reliably estimated. In the case of binary logistic regression, this appears as a failure of the Wald statistic and the report of no significant effect (e.g. p = 1), despite a clear effect and little or no measurement noise. Some authors (e.g. [16]) have suggested removing comparisons with very low (e.g. $p_{j \succ k} < 0.02$) and very high (e.g. $p_{j \succ k} > 0.98$) proportions, but omission of such extreme values produces empty cells. Rajae-Joordens and Engel [4] suggested replacing very low and very high values with $(1/2c_{jk})$ and $(1-(1/2c_{jk}))$, respectively, where c_{jk} is the number of comparisons for that pair. Cohen et al. [17] suggested using the likelihood ration test to compare the model deviances for n stimuli, D_{n-1}, and for n-1 stimuli, D_{n-2}, where the eliminated stimulus is a stimulus that produced extreme proportions. The deviances are reported in SPSS output as the -2 Log Likelihood (-2LL). The difference between the two model deviances yields a likelihood ratio χ^2 test with 1 degree of freedom, the associated p-value then denotes the significance (or not) for the paired comparison. We prefer Cohen et al.'s approach, as it does not involve removing or altering data that is meaningful.

When making pairwise comparisons, usually the two stimuli are temporally or spatially separated. Sequential presentation of the two stimuli is common, but the presentation order may affect responses. In testing sound preferences using pairwise comparisons, Wickelmaier and Choisel [18] reported a strong bias for the second sound, and that these could be accounted for in a Bradley-Terry-Luce model (an alternative to Thurstone scaling). In the comparison of videos, spatially separated stimuli are preferable to sequential presentations as it reduces memory limitations that may

occur with comparisons made between views seen at different times. Side (spatial) bias could result from a subject favoring one side over another (inherent preference) or from the apparatus (e.g. a difference in one of the phosphors of two CRT monitors). The impact of such biases can be minimized by careful study design (e.g. counter-balancing), but a model (data analysis approach) will be a better fit of the data if it can account for such biases.

Using an, as yet, unpublished data set (Study 1), we demonstrate that these three approaches to estimation of relative preferences measured using paired comparisons of moving video image quality produce very similar results. Also, we demonstrate that additional factors, such as response bias (here for side), can be included in the binary logistic regression approach. Using that data set and data from a related published study (Study 2) [19], we demonstrate that it is possible to estimate the relative preferences with a severely constrained (reduced) data set: where paired comparisons always included one stimulus, so response to many potential comparison pairs were not available.

2. METHODS

2.1 Study 1 – High Definition Video Enhancement

Four stimuli were compared by 40 subjects with normal sight. The stimuli were levels of image enhancement (image quality). Each subject made 64 pairwise comparisons, in random order, of videos presented on two side-by-side displays for a total of 2,560 responses. For each subject, all possible pairwise comparisons among the four stimuli (4 x 4 = 16) were made, including comparisons with the same enhancement level on each side. Each pairwise comparison (e.g. High enhancement on Left versus Low enhancement on Right) was made four times, counterbalanced for side of presentation.

High Definition (HD) videos were presented on two 42" HDTVs (VIZIO VO42L FHDTV10A, Irvine, CA) placed side by side at an angle of 148° between them and viewed from 7 feet. The two HDTVs (1080p) had closely matched luminance properties. HD (1920 x 1080 pixel) videos consisting of movie trailers and documentaries were downloaded from Apple websites (www.apple.com). Videos were edited using QuickTime 7 Pro (Apple Inc., Cupertino, CA) into 30-s segments. Following a HDMI splitter (HSP 12, ConnectGear Inc.,Fremont, CA), the videos were processed independently by two PureAV RazorVision devices (Belkin International, Inc., Los Angeles, CA). The RazorVision device is commercially marketed to enhance HD videos at three (Low, Medium and High) modest levels of enhancement and also provides a fourth level (Off) that presents the original unenhanced video. The two RazorVision devices were computer controlled and each applied one of four enhancement levels (Off, Low, Medium and High) per trial.

2.2 Study 2 – Standard Definition Video Enhancement

In an earlier study [19] 11 subjects with normal sight compared four stimuli. The stimuli were levels of image enhancement (image quality). Each subject made 16 pairwise comparisons, in random order, of videos presented on two side-by-side displays for a total of 176 responses. For each subject, in every pairwise comparison the original video was presented on one side and one of the four levels of image enhancement, including the Off level (original video) was presented on the other side. Each pairwise comparison (e.g. High enhancement on Left versus original video on Right) was made four times, counterbalanced for side of presentation.

Videos taken from DVD were presented on two 27" standard definition TVs placed side by side at an angle of 180° between them and viewed from 3 feet. The two TVs were the same model purchased at the same time, but did not have perfectly matched luminance properties. The original video was always seen on one TV and one of the four image enhancement levels (Off, Low, Medium and High) was presented on the other TV. The videos were switched using an Extron MMX 42 SVA RCA video-switching unit (Extron Electronics, Anaheim, California). The videos were processed by a CMOS chip (DV1000, DigiVision, San Diego, CA) with S-video input and output, that provided the same processing as the PureAV RazorVision device. The chip was computer controlled and applied one of four enhancement levels per trial. The DigiVision DV1000 chip adapts the adaptive enhancement algorithm originally developed to assist people with visual impairments [20].

This study had a much smaller data set than study 1, and was limited by having measured responses to only four of the possible 16 pairwise comparisons. As the Thurstone matrix (Table 1) was incomplete, it was not clear that it could be analyzed using Thurstone scaling. Therefore, in the published study, simple non-parametric analyses (Wilcoxon Signed Rank test) were employed.

2.3 Data Analysis

The Thurstone matrices for the two studies are shown in Table 4. The Thurstone perceptual scales were computed by normalizing the relative preferences, Z_i, to the standard zero-to-one scale.

Table 4. Thurstone matrices for the two studies. Since in Study 2 the pairwise comparisons always included the Off condition, some comparisons are marked as not available (na).

Study 1	*Off*	*Low*	*Medium*	*High*	Z_i
Off		-0.29	-0.28	0.08	-0.12
Low	0.29		0.02	-0.06	0.06
Medium	0.28	-0.02		0.18	0.11
High	-0.08	0.06	-0.18		-0.05

Study 2	*Off*	*Low*	*Medium*	*High*	Z_i
Off		-0.52	0.00	0.23	-0.07
Low	0.52		na	na	0.13
Medium	0.00	na		na	0.00
High	-0.23	na	na		-0.06

For the binary logistic regression approach [12], logistic regression tables were constructed with the four possible enhancement levels (stimuli) represented in four columns. This is illustrated for Study 1 in Table 5. For study 1, two additional columns that represented the side (left or right) and the monitor (A or B) that was preferred were included to allow testing for possible biases (as discussed below). When the preferred stimulus was on the right monitor, $Side_i$= 1, and when on the left, $Side_i$= -1, except for e_i = 0 ("false") when the assigned value of $Side_i$ was reversed. Similarly, when the preferred stimulus was on monitor A, $Display_i$= 1, and when on monitor B, $Display_i$= -1, except for e_i = 0 ("false") when the assigned value of $Display_i$ was reversed. The rows included the subject number and trial number for each subject. Binary logistic regression analyses were performed using SPSS 16.0.2 for Mac (SPSS, Chicago, IL). For study 1, the logistic regression was conducted three times, each with a different stimulus (enhancement level) excluded, so that the statistical significance of all between-stimulus differences could be found. Since in study 2 all comparisons included the original video (Off), a single logistic regression, with the Off stimulus excluded, was sufficient. Thurstone-like perceptual scales were found by normalizing the regression coefficients to a zero-to-one scale.

Table 5. Illustration of the logistic regression table for all 64 trials for all 40 subjects, constructed for Study 1 and used for the binary logistic regression analyses. Stimuli (enhancement levels) not presented during the trial are indicated by 0. Presented stimuli were allocated 1 or -1 depending on the reported preference and the identity vector (*e*), as described above. The preferred stimulus is shown in bold in each row.

Subject number	Subject trial	*Off*	*Low*	*Medium*	*High*	*Side*	*Display*	*e*
1	1	0	1	**-1**	0	1	-1	0
1	2	0	-1	**1**	0	-1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	64	0	0	**1**	-1	-1	-1	1
2	1	0	1	0	**-1**	1	-1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
40	63	-1	**1**	0	0	1	1	1
40	64	1	**-1**	0	0	1	-1	0

For the linear regression approach [4], linear regression tables were constructed with the four possible enhancement levels represented in four columns (Table 6). Linear regression analyses were conducted using the GLM component of SPSS 16.0.2 for Mac (SPSS, Chicago, IL). Thurstone-like perceptual scales were found by normalizing the regression coefficients to a zero-to-one scale.

Table 6. The linear regression tables constructed for the two studies used for the linear regression analyses. Stimuli (enhancement levels) not presented during the trial are indicated by 0. Presented stimuli were allocated 1 or -1 depending on the reported preference.

Study 1

Off	*Low*	*Medium*	*High*	$p_{j \succ k}$	$Z_{j \succ k}$
1	0	0	-1	0.53	0.08
1	0	-1	0	0.39	-0.28
1	-1	0	0	0.38	-0.29
0	1	0	-1	0.48	-0.06
0	1	-1	0	0.51	0.02
0	0	1	-1	0.57	0.18

Study 2

Off	*Low*	*Medium*	*High*	$p_{j \succ k}$	$Z_{j \succ k}$
1	0	0	-1	0.59	0.23
1	0	-1	0	0.50	0.00
1	-1	0	0	0.30	-0.52

2.4 Side Bias Analysis

In our two example studies there was a risk of subjects choosing one side whenever they were uncertain (people find it hard to respond randomly in such situations [21]). As an example, in study 1, there was one subject who responded right 15 of the 16 times (94%) that the same level of enhancement was shown on the two monitors. However, that subject responded right 25 of the 48 times (52%) when the enhancement levels were different on the two monitors. Thus, this subject appears to have had a strong right side bias in uncertain situations (here when the two stimuli were the same), but when able to distinguish differences between the stimuli, did not show a bias. Even when a subject appears to have a bias, it is possible that the apparent bias is a consequence of the experimental design. For example, when measuring video image quality using a different video segment for every comparison, if there is some interaction between an aspect of the videos (e.g. inclusion of fast action) and the method of altering the image quality, then it is not possible to distinguish between a side bias and an (unfortunate) interaction between the combination of video segment content with image quality manipulation and the side of presentation, unless there is prior knowledge of the interaction and that knowledge is used in the study design or its analysis.

For differences between display devices it is possible to minimize these by obtaining "identical" (e.g. from the same production batch) devices and by adjusting the devices to have equal properties of interest (e.g. luminance, color). Even when this is done, it may not be adequate. Therefore, in study 1, the locations of the two HDTVs were swapped after running the experiment on the first half of the subjects. Thus, it was possible to construct, in the same manner as the *Side* variable, another variable that reported the preferred monitor.

The outcomes of the side bias analysis conducted within the binary logistic regression were compared to the score test of proportions, where the observed proportions for a preferred side and preferred monitor were compared to the proportion expected if there was no bias, 50%.

3. RESULTS

For the data from Study 1, the perceptual scales found using the three analysis approaches were almost identical (Fig. 1a, b). The differences between the three approaches were that the relative preferences for High were 0.317, 0.315 and 0.316, and for Low were 0.790, 0.788 and 0.791 for Thurstone scale, binary logistic regression and linear regression, respectively. Figure 1b shows the statistical significance of the significant differences, which were obtained directly from the SPSS output from the binary logistic regression analysis. Binary logistic regression found that Low and Medium enhancement levels were significantly different from Off and High ($p \leq 0.03$), and that Low and Medium ($p=0.32$) and Off and High ($p=0.14$) were not significantly different from one another. As the enhancement levels were ordinal, Figure 1c shows the binary logistic regression perceptual scale as a two-dimensional graph instead of a

traditional linear scale. This presentation of the data makes it easier to see that the image-quality preferences were a non-monotonic function of the enhancement level.

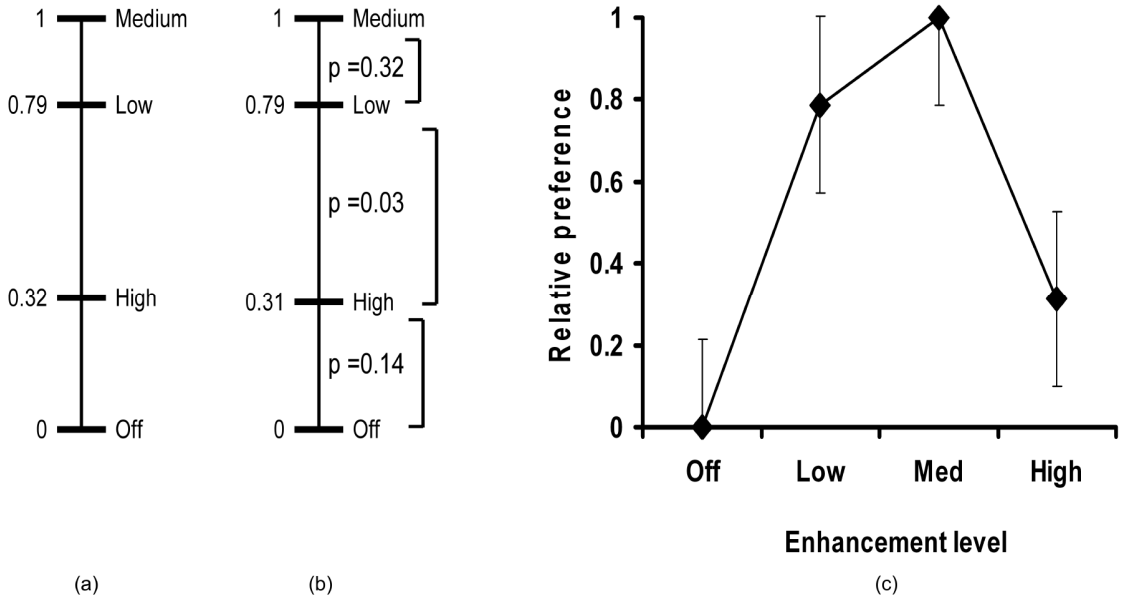

Figure 1: The perceptual scales for Study 1 (with all 40 subjects and all available pairwise comparisons) found using the three data analysis methods (a) The (normalized) Thurstone [5-7] scale and the linear regression [4] scale were almost the same. (b) The binary logistic regression [12] scale was almost the same as the outcomes of the other two analyses (see text for details). The statistical significance (or lack thereof) of differences between stimuli found using binary logistic regression is shown. (c) As our stimuli were ordinal, the perceptual scale is graphed, which illustrates the non-monotonic preference function for this image enhancement. Error bars are 95% confidence intervals of the relative preferences (1.96 X scaled standard error of the coefficient).

When the variables *Side* and *Display* were included in the binary logistic regression, there was no change in the relative preference scale. Monitor A was chosen in 1283 of 2560 trials (50.1%) which was not significantly different from the expected 50% (score test, $z = 0.12$, $p = 0.91$), which is consistent with the lack of an effect of the display on responses found with binary logistic regression (Wald = 0.014, $p = 0.91$). Subjects chose the video on the right in 1343 of 2560 trials (52.5%), which was significantly different from the expected 50% (score test, $z = 2.49$, $p = 0.013$), which is consistent with the effect of side on responses found with binary logistic regression (Wald = 6.63, $p = 0.012$). When only considering the 1920 trials in which the two stimuli were different, subjects chose the right monitor in 925 trials (48%; i.e. slight left bias), which was not statistically significant (score test, $z = 1.60$, $p = 0.11$; Wald = 2.59, $p = 0.11$). Conversely, there was a strong bias for the right side (65.3%; $z = 8.14$, $p < 0.001$), when the two stimuli were the same, presumably when the subjects were uncertain. This seems to discount other external explanations of side bias such as lighting differences (the only other light in the test room was the computer monitor to the left of the left display that was turned away from the subject's view). Removing the same-stimulus trials has no impact on the relative preferences for the enhancement levels, as those trials do not contribute to the simulus factors in the logistic regression.

In their analysis of the data from Study 2, Fullerton and Peli [19] reported the proportion of times that Low, Medium and High were preferred over the original video (Fig. 2a). By comparing the proportions, Low was preferred significantly over Off (score test, $z = 2.89$, $p = 0.004$), Medium ($z = 1.91$, $p = 0.03$) and High ($z = 2.74$, $p = 0.003$). The Thurstone scale found using Table 4b produced relative preferences (Fig. 2b) that were different from the preference proportions (raw data; Fig 2a). Binary logistic regression (Fig. 2c) and linear regression (Fig. 2d) produced a relative preference scale that was almost identical to the preference proportions (Fig 2a), when those were scaled from zero-to-one. Similar to the original report [19], the binary logistic regression found statistical significance for the differences between Low and Off ($p = 0.009$) and High ($p = 0.006$), and the difference between Low and Medium ($p = 0.052$) was almost statistically significant (at the classic $\alpha \leq 0.05$ level).

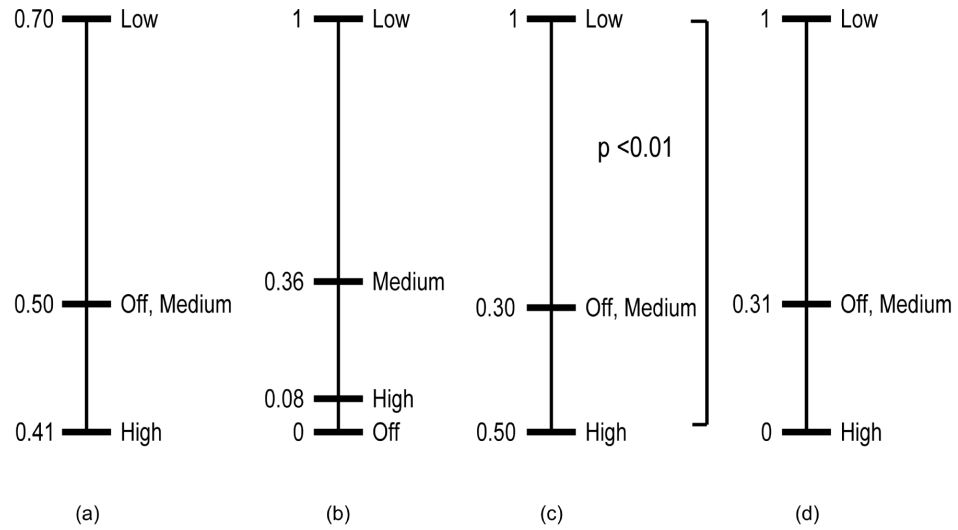

Figure 2: For Study 2, (a) the raw proportions that Low, Medium and High image enhancement were preferred over Off (original video), and the perceptual scales obtained with (b) Thurstone scaling, (c) binary logistic regression, and (d) linear regression.

To examine the effect of not having all available comparisons, the data from Study 1 were reanalyzed using only the comparisons between Off and the other three enhancement levels. Again, the outcomes of Thurstone scaling (Fig. 3b) were not consistent with preference proportions (Fig. 3a), while binary logistic regression (Fig. 3c) and linear regression (Fig. 3d) were almost identical to the preference proportions when those were scaled from zero-to-one. When using this subset of the data for Study 1, the relative preference scales were not the same as when all available data was used (compare Figs. 3c and 3d with Figs. 1b and 1a, respectively).

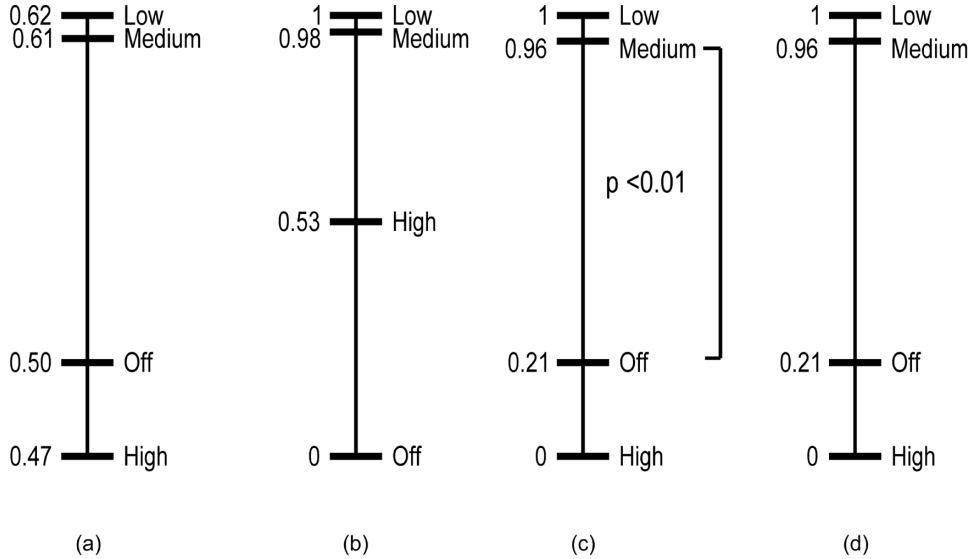

Figure 3: For Study 1, when using only comparisons that included Off (original video), (a) the raw proportions, and the perceptual scales obtained using (b) Thurstone scaling, (c) binary logistic regression, and (d) linear regression.

4. DISCUSSION

Evaluating the image quality of moving video is difficult [1]. We were able to find differences in perceived image quality, assessed as a preference, produced by a commercial image-enhancement device and to derive statistical inferences about measured differences. In Study 1, subjects watching HD moving videos, enhanced with the RazorVision device on HDTVs, preferred Medium and Low levels over Off and High (but the difference in preference between Medium and Off and that between High and Off was not significant). Those preferences were different, particularly for the High enhancement level, from those reported in Study 2 [19] when standard definitions TVs were viewed from a smaller distance. From the analysis of a subset of the data from Study 1, it appears that some of that difference in the preferences for the image enhancement levels may have been due to the use of only three of the six possible comparisons. At the time of that earlier study [19], availability of hardware constrained the study design.

Pairwise comparison experimental designs are used in image quality evaluation studies [2-4]. Classic Thurstone scaling constructs a perceptual scale from the pairwise comparison data, with a limitation of not deriving the statistical significance among the compared pairs. In this paper, we describe two statistical methods (binary logistic regression and linear regression) to analyze pairwise comparison data and produce perceptual scales that are almost identical to the Thurstone scale, with the advantage of determining the statistical significance between stimuli. The advantage of the binary logistic regression over the linear regression approach is that the statistical significance can be found with typical statistical software without the need for additional calculations.

Using the binary logistic regression approach, we demonstrated that it is possible to examine the impact of other factors on preferences within a single analysis. For Study 1, there was no effect of the apparatus (Display) but there was a significant effect of the physical location (Side). That effect of video location was found for the comparisons for which there was no difference in the two stimuli, but it was not found for comparisons between different stimuli. This supports the contention that people often do not equally disperse their responses (guesses) when they are uncertain [21]. In an analysis, the details of which are not reported here, we found that when subjects were separated into two groups based on an interview, it was possible to use binary logistic regression to show that the two groups had different responses to the image enhancement. This is comparable to the demonstration of using a dummy variable to examine the effect of gender using the linear regression approach [4].

Analysis of the data from Study 2 and a subset of data from Study 1 showed that a Thurstone scale constructed from an incomplete data matrix (i.e. with missing comparisons) will produce an unreliable outcome. However, binary logistic regression and linear regression produced outcomes that were consistent with the proportions of preferences (in the raw data) and consistent with the outcomes of multiple non-parametric tests. The advantage of the statistics derived from the binary logistic regression is that it involves only a single analysis, and thus does not contain the risk of multiple-comparison (type 1) errors. In this case, the linear regression model was a perfect fit of the data, as there were as many stimuli (factors) as data (i.e. only three rows of data), and thus various statistics could not be generated.

5. CONCLUSION

This study demonstrates that side-by-side pairwise comparisons can be used to evaluate subjective preferences for moving videos. Binary logistic regression produces a perceptual scale very similar to a Thurstone scale, with the advantage that the statistical significance of differences between stimuli on the perceptual scale can be evaluated. These perceptual scales allow easy interpretation and visualization of the results.

ACKNOWLEDGMENTS

Supported by NIH grants EY05957 and EY16093, and a grant from Analog Devices Inc.

REFERENCES

[1] E. Peli and R. L. Woods, "Image enhancement for impaired vision: The challenge of evaluation," *International Journal on Artificial Intelligence Tools* **18** (3), 415-438 (2009).

[2] J. E. Farrell, "Colour Imaging: Vision and Technology," in, L. W. MacDonald and M. R. Luo, Eds., 285-313, John Wiley & Sons Ltd. (1999).

[3] J. C. Handley, "Comparative analysis of Bradley-Terry and Thurstone-Mosteller paired comparison models for image quality assessment," in *PICS 2001: Image Processing, Image Quality, Image Capture Systems Conference* **4**, 108-112, The Society for Imaging Science and Technology (2001).
[4] R. Rajae-Joordens and J. Engel, "Paired comparisons in visual perception studies using small sample sizes," *Displays* **26**:1, 1-7 (2005).
[5] L. L. Thurstone, "A law of comparative judgment," *Psychological Review* **34**, 273-286 (1927).
[6] L. L. Thurstone, "Psychophysical Analysis," *The American Journal of Psychology* **38** (3), 368-389 (1927).
[7] L. L. Thurstone, *The Measurement of Values*, Chicago: The University of Chicago Press, Chicago (1959).
[8] C. H. Coombs, "Thurstone's measurement of social values revisited forty years later," *Journal of Personality and Social Psychology* **6** (1), 85-91 (1967).
[9] Y. K. Kwan, W. C. Ip and P. Kwan, "A crime index with Thurstone's scaling of crime severity," *Journal of Criminal Justice* **28**, 237-244 (2000).
[10] P. E. Green, D. S. Tull and G. S. Albaum, *Research for Marketing Decisions*, Prentice Hall, Englewood Cliffs, N.J. (1988).
[11] M. C. Gacula and J. Singh, *Statistical Methods in Food and Consumer Research*, Academic Press, Orlando, FL (1984).
[12] S. Lipovetsky and M. W. Conklin, "Thurstone scaling via binary response regression," *Statistical Methodology* **1**:93, 93-104 (2004).
[13] F. Mosteller, "Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations," *Psychometrika* **16** (1), 3-9 (1951).
[14] W. A. Glenn and H. A. David, "Ties in paired-comparison experiments using a modified Thurstone-Mosteller model," *Biometrics* **16** (1), 86-109 (1960).
[15] W. W. Hauck and A. Donner, "Wald's test as applied to hypotheses in logit analysis," *Journal of the American Statistical Association* **72**, 851-853 (1977).
[16] J. P. Guilford, *Psychometric Methods*, McGraw-Hill, New York, (1954).
[17] J. Cohen, P. Cohen, S. G. West and L. S. Aiken, *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates Inc., New Jersey (2003).
[18] F. Wickelmaier and S. Choisel, "Modeling within-pair order effects in paired-comparison judgments," in *Fechner Day 2006. Proceedings of the 22nd Annual Meeting of the International Society for Psychophysics*, 89-94, The International Society for Psychophysics (2006).
[19] M. Fullerton and E. Peli, "Digital enhancement of television signals for people with visual impairments: Evaluation of a consumer product," *Journal of the Society for Information Display* **16** (3), 493-500 (2008).
[20] E. Peli and T. Peli, "Image enhancement for the visually impaired," *Optical Engineering* **23** (1), 47-51 (1984).
[21] M. A. García-Pérez, "Denoising forced-choice detection data [Epub ahead of print]," *British Journal of Mathematical and Statistical Psychology* (PMID: 19422731) (2009).